Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 004.8
DOI: https://doi.org/10.24867/13BE18Rajanovic

PREDIKCIJA *BILLBOARD* HITOVA NA OSNOVU AUDIO I TEKSTUALNIH OBELEŽJA PESAMA

BILLBOARD HIT SONG PREDICTION BASED ON AUDIO AND TEXT FEATURES

Sandra Rajanović, *Fakultet tehničkih nauka, Novi Sad*

Oblast – ELEKTROTEHNIKA I RAČUNARSTVO

Kratak sadržaj – *U ovom radu prikazano je kreiranje sistema za predikciju hitova na Billboard top-listama, koristeći audio obeležja i tekstove pesama. Rešenje problema binarne klasifikacije implementirano je uz pomoć više modela mašinskog učenja i dve tehnike obrade prirodnog jezika, a realizovano je u Python programskom jeziku.*

Ključne reči: *mašinsko učenje, predikcija hit pesama, obrada prirodnih jezika*

Abstract – *This paper presents creating a system for Billboard hit song prediction using their audio features and lyrics. The binary classification problem solution was implemented using multiple machine learning models and two natural language processing methods, all created in Python programming language.*

Keywords: *machine learning, hit song prediction, natural language preocessing*

1. UVOD

Hit Song Science se bavi pretpostavkom da se veliki broj popularnih pesama odlikuje istim obeležjima na osnovu kojih se može proceniti da li neka pesma ima šanse da postane hit ili ne [1]. Sa razvojem tehnologije, *state-of-the-art* modeli postaju sve bolji, te ova oblast pronalazi primenu i u komercijalne svrhe. Sa digitalizacijom muzičke industrije i pojavom *streaming* platformi dolazi do promena u načinu konzumiranja muzike, što dovodi do ogromnih količina muzičkog sadržaja i velike konkurencije, gde je jedan od načina izdvajanja uspešna hit pesma.

Kreiranjem modela koji bi mogao da predvidi kakve pesme imaju veću šansu da postanu popularne i koje karakteristike na to imaju uticaja, drugim izvođačima bi bilo omogućeno da uključe te karakteristike u svoja dela i povećaju šanse za svoj uspeh. Popularne pesme donose veliku finansijsku dobit, pa bi dovoljno precizan model bio od ogromnog značaja, jer bi omogućio izvođačima i muzičkim kućama da na potencijalne hitove usmere svoje finansijske resurse, i tako ostvare najveću moguću zaradu. Danas se standardom u proceni popularnosti smatraju top-liste gde se pesme rangiraju na osnovu slušanosti, prodaje i puštanja na radiju, a najpoznatije su *Billboard* top-liste. Ovde je najznačajnija *Billboard Hot* 100 lista koja predstavlja dosta realnu sliku trenutno aktuelne muzike. Stoga, pesme sa ove liste mogu da predstavljaju reprezentativan uzorak hit pesama za analizu.

Tema ovog rada je predikcija *Billboard* hitova na osnovu obeležja pesama koja čine audio i tekstualna obeležja. Ulazni skup podataka čine pesme nastale u poslednjih deset godina, gde su popularne pesme uzete sa *Billboard Hot* 100 listi, a ostatak pesama je izabran nasumično. Dodatni manji podskup podataka je kasnije kreiran, koji sadrži samo hitove sa top 40 pozicija top-listi. Rad se sastoji iz dva različita pristupa treniranju modela, gde se prvi ogleda u korišćenju *topic* modela i četiri različita modela mašinskog učenja, dok drugi koristi BERT model [2] i neuronske mreže. Oba rešenja su evaluirana na isti način koristeći tri mere – tačnost (eng. *accuracy*), preciznost (eng. *precision*) i odziv (eng. *recall*).

2. ARHITEKTURA REŠENJA

Proces modelovanja softverskog rešenja sastoji se iz tri celine. Prva obuhvata prikupljanje podataka, njihovo pretprocesiranje i podelu na trening i test skup. Druga celina se bavi dvema različitim implementacijama rešenja i obuhvata optimizaciju parametara, a zatim i treniranje različitih modela. U trećoj celini testiraju se najbolji modeli. Slika 1 prikazuje šemu celoukupnog procesa.

Za prikupljanje podataka korišćena su tri izvora i formirane su dve verzije skupa podataka i njihovih podskupova. Prvoj verziji su pripojena dodatna obeležja dobijena pomoću *topic* modela, a drugoj obeležje teksta pesme. Pretprocesiranje se sastoji se od skaliranja obeležja na opseg između 0 i 1.

U drugom delu, prvi način implementacije obuhvata optimizaciju parametara pomoću unakrsne validacije za četiri modela mašinskog učenja, a nakon toga i selekciju obeležja. Drugi način implementacije podrazumeva korišćenje BERT modela za procesiranje tekstualnih obeležja i treniranje modela neuronskih mreža.

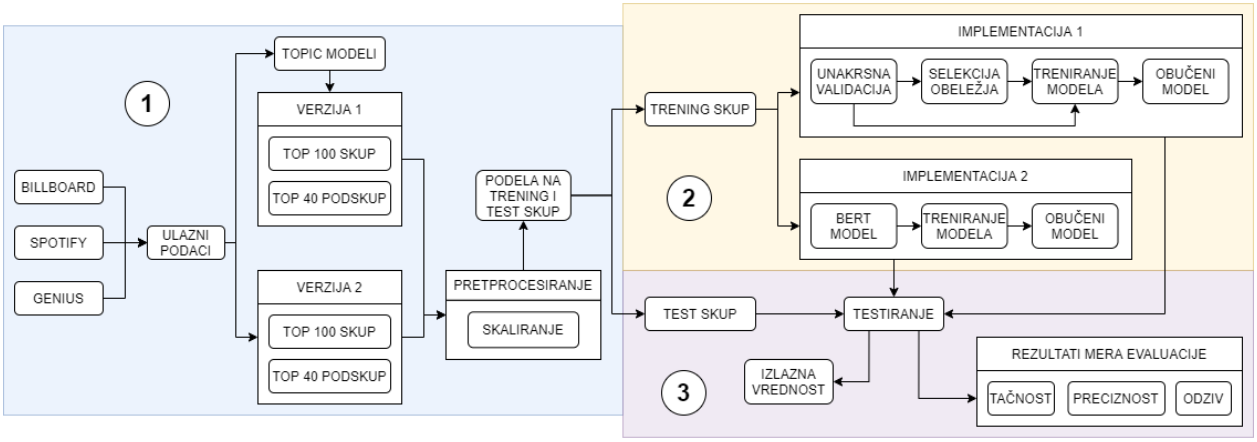
Poslednji deo sastoji se od evaluacije performansi obučenih modela nad testnim podacima.

3. PRIKUPLJANJE I OBRADA PODATAKA

Za potrebe ovog rada prikupljeni su podaci sa *Billboard Hot* 100 top-liste **Error! Reference source not found.**, *Spotify* platforme [4] i *Genius* baze podataka [5], čijim spajanjem je kreiran skup podataka sa 22 obeležja. Kreirane su dve verzije skupa – glavni skup koji sadrži 9126 instanci, i njegov podskup od 3390 numera.

Ciljna labela ima vrednosti – 1 ako je hit, i 0 nije. Oba skupa sadrže približno jednak broj pesama iz obe kategorije radi balasiranosti u skupu podataka.

NAPOMENA:
Ovaj rad proistekao je iz master rada čiji mentor je bila dr Jelena Slivka, vanr. prof.



Slika 1. *Šema arhitekture rešenja*

3.1. Opis skupa podataka

Prikupljeni podaci su iz vremenskog perioda od 2010. do 2019. godine. Podaci o hit pesmama i njihovim izvođačima su dobavljeni sa *Billboard Hot* 100 liste kao u radovima [6] i [7]. Ove liste se ažuriraju jednom nedeljno, pa je na raspolaganju bilo njih 530, tj. 53000 pesama. Nakon uklanjanja duplikata i pesama za koja nisu bila dostupna audio obeležja, dobijen je skup od 4558 pesama. Po ugledu na pristup u radu [8], ne-hitovi su definisani kao sve ostale pesme onih izvođača koji imaju bar jednu pesmu na top-listama. Pomoću *Spotify* API-a dobavljene su sve njihove ostale pesme izdate u posmatranom vremenskom periodu. Ovim procesom prikupljeno je 173227 pesama, od kojih je izdvojen manji podskup od 4568 pesama, kreiran vodeći računa o balansiranosti podataka među klasama. Konačan skup svih podataka brojao je 9126 instanci.

Pomoću *Spotify* API-a dobavljena su audio obeležja svih pesama konačnog skupa, a zatim su koristeći *Genius* API prikupljeni tekstovi ovih pesama. Određena količina pesama u skupu nije na engleskom jeziku, pa je za njih sa interneta ručno preuzet engleski prevod tekstova.

U cilju analiziranja da li je lakše predvideti hitove sa vrha top-lista, kreiran je podskup originalnog skupa za koji se izdvajaju samo pesme koje su se našle u top 40 pozicija. Na ovaj način je dobijeno 1695 hit pesama, a njima je pridruženo isto toliko nasumično izabranih ne-hitova iz prvog skupa, te je konačan broj instanci iznosio 3390.

3.2. Obeležja i labela

Skup podataka ima dve verzije, gde je prva opisana sa 22, a druga sa 19 obeležja, a oba imaju istu izlaznu labelu. Obe verzije skupa imaju obeležja koja mogu biti podeljena u dve grupe: 13 obeležja dobavljenih pomoću *Spotify* API-a i 5 dodatnih boolean vrednosti. Pored toga, obe verzije sadrže i dodatna obeležja koja reprezentuju tekstove pesama.

Audio obeležja pesama, slično kao u radovima [6], [7] i [8], obuhvataju:

- trajanje pesme
- tonalitet
- mod
- promena taktova
- akustičnost
- plesnost
- energija
- instrumentalnost
- živost
- glasnost
- govornost
- valentnost
- tempo

Naredna grupa obeležja obuhvata 5 *boolean* obeležja dodatnih na osnovu uočenih istorijskih trendova kao potencijalni faktori u popularnosti pesama. U nedostatku lakšeg načina za dodelu ispravnih vrednosti ovih obeležja, ceo skup je ručno anotiran. Ovde spadaju sledeći indikatori:

- pesma je originalno na jeziku koji nije engleski
- pesmu izvodi popularan izvođač
- obrada pesme
- božićna pesma
- remiks pesme

Poslednja četiri obeležja u prvoj verziji skupa imaju vrednosti između 0 i 1, a dobijena su kao rezultati *topic* modela, po ugledu na rešenje u radu [9]. Ove vrednosti predstavljaju verovatnoće da numera pripada toj temi na osnovu sadržaja teksta pesama. Druga verzija skupa umesto ovih obeležja koristi samo jedno tekstualno obeležje koje sadrži ceo tekst pesme, i obrađuje se kasnije u toku implementacije.

4. IMPLEMENTACIJA SISTEMA

Rešenje problema implementirano je u programskom jeziku *Python*. Podaci su dobavljeni iz tri izvora koristeći biblioteke *billboard*, *spotipy* i *lyricsgenius*. Pomoću *billboard* biblioteke su dobavljene sve *Billboard Hot* 100 top-liste u vremenskom periodu od 2010. do 2019. godine. Biblioteka *spotipy* predstavlja enkapsulaciju *Spotify* API-a, a korišćena je prvo za pronalaženje svih hit pesama i njihovih izvođača radi dobijanja njihovih *Spotify* ID-eva. ID-evi izvođača se zatim koriste za dobavljanje njihovih ostalih pesama, a potom se pomoću ID-eva pesama preuzimaju njihova audio obeležja. Biblioteka *geniuslyrics* koja takođe predstavlja enkapsulaciju *Genius* API-a, koristi se za prikupljanje tekstova pesama.

Topic modeli implementirani su pomoću biblioteka *nltk*, *gensim* i *scikit-learn*. Biblioteka *nltk* korišćena je za tokenizaciju tekstova i izbacivanje stop reči, a *gensim* korišćen je za određivanje vrednosti *perpexity* mere koja služi za određivanje optimalnog broja tema, pomoću koje je izabran broj 4 kao rešenje. Izabrani *topic* model je

Latentna Dirihleova alokacija (eng. *Latent Dirichlet allocation*, skraćeno LDA) koji se implementira koristeći *scikit-learn* biblioteku. Za njegovo kreiranje korišćene su klase *TfidfVectorizer* i *LatentDirichletAllocation* pomoću kojih se dobijaju verovatnoće pripadanja svakoj od tema za svaki dokument. Kasnija treniranja modela su izvršena sa i bez ovih obeležja radi analize i poređenja rezultata.

Potom se skaliraju vrednosti određenih obeležja, a zatim se skup deli na trening i test skup u razmeri 80:20.

U okviru prvog načina implementacije realizovana je unakrsna validacija, selekcija obeležja i treniranje četiri modela mašinskog učenja. Unakrsna validacija se koristi radi nalaženja najoptimalnijih parametara za treniranje, a izvedena je pomoću klasa *KFold* i *GridSearchCV* iz *scikit-learn* biblioteke. U ovom procesu trening skup se deli na pet delova i definišu se parametri za koje će svaki model biti treniran. Na taj način su izabrani najpovoljniji parametri za treniranje četiri modela – logistička regresija, *random forest*, metoda potpornih vektora (eng. *Support Vector Machine*, skraćeno SVM) i *gradient boosting*. Za svaki od njih, unakrsna validacija je izvršena po četiri puta, odnosno jednom za svaku verziju skupa podataka – originalni skup, originalni skup bez *topic* obeležja, podskup originalnog skupa i podskup bez *topic* obeležja.

Model logističke regresije je kreiran koristeći klasu *LogisticRegression*, a u procesu unakrsne validacije posmatrani su sledeći parametri: *penalty*, *solver*, *class_weight*, *multi_class* i *max_iter*.

Random forest model je implementiran pomoću klase *RandomForestClassifier*, a optimizovani su mu naredni parametri: *n_estimators*, *criterion* i *max_features*.

Za model SVM, kreiran pomoću klase SVC, sledeći parametri su optimizovani: *C*, *gamma* i *kernel*.

Na kraju, za *gradient boosting* model, implementiran koristeći *GradientBoostingClassifier* klasu, sledeći parametri su trenirani u unakrsnoj validaciji: *loss*, *n_estimators*, *criterion* i *max_features*.

Nakon dobijanja najboljih parametara, vrši se treniranje nad trening i predikcija nad test skupom. Pored ovoga je implementirana i selekcija obeležja koja je primenjena radi optimizacije rešenja, a realizovana je koristeći *SelectFromModel* klasu *scikit-learn* biblioteke. Potom se ponovo vrši treniranje modela sa istim parametrima.

Drugi način implementacije rešenja ostvaren je preko BERT modela i modela neuronskih mreža. BERT je tehnika mašinskog učenja baziranu na transformer modelima dubokog učenja. Biblioteka *transformers* dolazi sa pretreniranim BERT modelima koji su bili upotrebljeni za implementaciju rešenja, zajedno klasama *BertTokenizer* i *BertModel*. *BertTokenizer* služi za tokenizaciju tekstova i pripremu ulaznih informacija za model, a njegova izlazna vrednost predstavlja ulaz za klasu *BertModel* koja dati ulaz transformiše u vektorske reprezentacije reči. Izlazu iz ovog modela se redukuje dimenzionalnost i tako se dobija dvodimenzionalna matrica gde je svaki ulazni podatak predstavljen nizom dužine 768. Nakon toga se ovim podacima dodaje i 18 postojećih audio obeležja.

Modeli neuronskih mreža kreirani su uz pomoć *keras* biblioteke, a treniranje je izvršeno i sa svim obeležjima, i samo sa tekstualnim obeležjima iz BERT modela. Struktura modela, broj slojeva i dimenzionalnost određeni su empirijski. U modelima su korišćeni potpuno povezani slojevi predstavljeni objektima klase *Dense*, a za svaku verziju skupa podataka, dodate su različite kombinacije slojeva - uglavnom su u pitanju troslojne strukture sa aktivacionim funkcijama *relu* i *sigmoid*. Ulazni podaci su dimenzija 786 za skup sa svim obeležjima i 768 za skup sa samo tekstualnim. Nakon dodavanja slojeva, vrši se kompajliranje svakog modela sa *binary_crossentropy* funkcijom gubitka i optimizacionom funkcijom *adam*, a treniranje se vrši u 150 epoha sa delovima veličine 10.

5. VERIFIKACIJA

Verifikacija obuhvata procenu kvaliteta performansi modela, a sprovodi se uz pomoć tri mere evaluacije nad testnim skupom podataka. Nakon toga su analizirane greške modela i mogući razlozi za njihovo pojavljivanje.

5.1. Rezultati

Mere evaluacije primenjene u radu su tačnost, preciznost i odziv. Za obučavanje i testiranje modela korišćen je skup od 9126 i njegov podskup od 3390 pesama, oba podeljena u razmeri 80:20.

Tabela 1 predstavlja rezultate modela nad celim skupom podataka. Skoro svi modeli imaju dosta slične rezultate i na trening i na test skupu, dok su u nekim slučajevima performanse čak i bolje na testnom nego na trening skupu. Izuzetak je *random forest* model, kod kojeg su rezultati nad trening skupom skoro savršeni, ali su na testnom značajno slabiji, što može ukazivati na *overfitting* modela. I pored ovoga, ova model je dao najbolje rezultate, zajedno sa neuronskim mrežama i *gradient boosting* modelom. Logistička regresija i SVM imaju dosta niže odzive što stvara veću razliku između vrednosti preciznosti i odziva, te može biti znak da model prediktuje veliki broj lažno negativnih instanci. Neuronske mreže imaju primetno bolju meru odziva od ostalih, što smanjuje broj lažnih negativnih instanci.

Za prva četiri modela treniranje je izvršeno i bez *topic* obeležja što je dalo skoro identične rezultate, pa se može pretpostaviti da tekstovi pesama nemaju mnogo uticaja na popularnost pesama, ili da *topic* modeli nisu dobro rešenje ovog problema. Takođe je izvršena i selekcija obeležja nad oba skupa (sa i bez *topic* obeležja), što je opet dalo identične rezultate. Razlog za ovo može biti to da izbačena obeležja svakako nisu imala veliki uticaj na konačni ishod, ili da su trenutni modeli dostigli svoj maksimum, te dodatne optimizacije nemaju efekta. Nad neuronskim mrežama obučavanje je izvršeno i samo sa tekstualnim obeležjima što je dalo dosta lošije rezultate od prethodnih, te je jasno da predikcija samo na osnovu tekstova pesama nije dovoljno dobra.

Tabela 2 prikazuje performanse nad podskupom podataka. Tačnosti svih modela su poboljšane za 3-5% u odnosu na rezultate sa celim skupom podataka. Mera odziva beleži primetno poboljšanje kod svih modela – porast se kreće u opsegu 7-14%, čime se smanjuje velika razlika između odziva i preciznosti, što je prethodno bio potencijalni izvor problema. Modeli *random forest* i neuronske mreže imaju generalno najbolje rezultate, a odmah nakon njih slede logistička regresija i *gradient boosting*. Model logističke regresije ima najveću tačnost i svaka mera evaluacije ima bolji rezultat nad testnim u odnosu na trening skup, dok model neuronskih mreža ima najbolje vrednosti odziva.

Tabela 1. *Performanse modela nad celim skupom podataka*

	Accuracy		Precision		Recall	
	Trening skup	Test skup	Trening skup	Test skup	Trening skup	Test skup
Logistic Regression	0.69589	0.69769	0.77700	0.78125	0.54854	0.54824
Random Forest	0.99986	0.72562	0.99972	0.75979	1.0	0.65899
SVM	0.69972	0.69989	0.80752	0.80434	0.52358	0.52741
Gradient Boosting	0.72945	0.70974	0.79040	0.76381	0.62369	0.60635
Neural Network	0.82369	0.70920	0.82112	0.71574	0.82720	0.69298

Tabela 2. *Performanse modela nad podskupom podataka*

	Accuracy		Precision		Recall	
	Trening skup	Test skup	Trening skup	Test skup	Trening skup	Test skup
Logistic Regression	0.73266	0.75516	0.78707	0.79522	0.63790	0.68731
Random Forest	0.99963	0.75073	1.0	0.76234	0.99926	0.72861
SVM	0.73045	0.75073	0.83783	0.82692	0.57153	0.63421
Gradient Boosting	0.79240	0.74778	0.83403	0.76415	0.73008	0.71681
Neural Network	0.92072	0.75516	0.90035	0.73441	0.94616	0.79941

5.2. Analiza grešaka modela

Analiza grešaka izvršena je poređenjem stvarnih i prediktovanih izlaznih vrednosti za svaki od modela. Kreirane su matrice zabuna (eng. *confusion matrix*) na osnovu kojih je ustanovljeno da uglavnom ima više lažno negativnih nego lažno pozitivnih, što je i očekivano s obzirom na to da modeli imaju veće preciznosti, a manje odzive. Izuzetak su modeli neuronskih mreža kod kojih su količine ovih instanci balansirane.

Od svih pogrešnih instanci izdvojene su samo one koje su modeli pogrešno predvideli sa pouzdanošću većom od 80%, gde je situacija sa lažnim instancama bila obrnuta, odnosno više je bilo lažno pozitivnih nego negativnih. Ovo je dosta nepovoljan ishod, jer bi s obzirom na svrhu ovog modela cena bila veća za lažno pozitivne rezultate, tj. za ne-hit pesme koje su pogrešno klasifikovane kao hitovi. Analizirane su vrednosti obeležja na kojima modeli najviše greše i uočeni su određeni trendovi gde većina modela greši na sličnim obeležjima. Računate su srednje vrednosti svakog obeležja za pogrešne u odnosu na tačno klasifikovane instance. Za ceo skup podataka, sledeća obeležja imaju drugačije vrednosti za pogrešne u odnosu na tačno prediktovane:

- hitovi: manja plesnost i valentnost, veća instrumentalnost i živost
- ne-hitovi: manja instrumentalnost

U slučaju podskupa podataka, uočeni su sledeći slučajevi:

- hitovi: manja plesnost, valentnost, energija, glasnost i instrumentalnost, veća živost
- ne-hitovi: veći mod, manja instrumentalnost i valentnost

6. ZAKLJUČAK

Cilj ovog rada bio je kreiranje sistema za predikciju hit pesama na *Billboard* top-listama koji bi bio od velikog značaja muzičkim kućama i izvođačima. U radu su kreirani skup podataka od 9126 i njegov podskup od 3390 pesama opisanih audio i tekstualnim obeležjima.

Implementacija je odrađena u *Python* jeziku pomoću *topic* i BERT modela i pet modela mašinskog učenja. Performanse svih obučenih modela su slične – tačnost za ceo skup je u rangu 69-72%, a za podskup oko 74-75%. Dodatne optimizacije modela nisu dale bolje rezultate.

Ideje za dalja proširenja rada obuhvataju odabir dodatnih obeležja koja tačnije opisuju podatke, kao i poboljšanje BERT modela implementacijom ponovnog treniranja pretreniranih modela uključujući i tekstualna obeležja iz skupa u ovom radu.

7. LITERATURA

[1] Y. Ni, R. Santos-Rodriguez, M. Mcvicar and T. De Bie, "Hit song science once again a science", *4th International Workshop on Machine Learning and Music*, 2011.

[2] J. Devlin, M.W. Chang, K. Lee and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding", *arXiv preprint arXiv:1810.04805*, 2018.

[3] https://www.billboard.com/charts/hot-100

[4] https://www.spotify.com/

[5] https://genius.com/

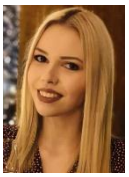
[6] K. Middlebrook and K. Sheik, "Song Hit Prediction: Predicting Billboard Hits Using Spotify Data", *arXiv preprint arXiv:1908.08609*, 2019.

[7] E. Georgieva, M. Suta and N. Burton, "HITPREDICT: PREDICTING HIT SONGS USING SPOTIFY DATA", 2018.

[8] A. Singhi and D.G. Brown, "Can song lyrics predict hits", *Proceedings of the 11th International Symposium on Computer Music Multidisciplinary Research*, pp. 457-471, 2015.

[9] R. Dhanaraj and B. Logan, "Automatic Prediction of Hit Songs", *ISMIR*, pp. 488-491, 2005.

Kratka biografija:



Sandra Rajanović rođena je u Beogradu 1995. god. Master rad na Fakultetu tehničkih nauka iz oblasti Elektrotehnika i računarstvo – Softversko inženjerstvo i informacione tehnologije odbranila je 2020. god.