Analiza tekstova pesama sa Evrovizije

Vesna Prica
Fakultet Tehničkih Nauka
Novi Sad, Srbija
prica.vesna98@gmail.com

Aleksa Matić
Fakultet Tehničkih Nauka
Novi Sad, Srbija
aleksa.matic408@gmail.com

Abstrakt* — *Ovaj rad se bavi istraživanjem uticaja tekstova pesama na njihov plasman na takmičenju Evrovizije. Ukoliko se pokaže da postoji jaka korelacija između reči pesama i ostvarene pozicije, to bi moglo da dovede da u budućnosti pesme budu slične, jer bi svi učesnici težili da budu maksimalno efikasni. Sa druge strane, ukoliko se pokaže da veza nije toliko jaka, to može ohrabriti učesnike da budu dosledni svojim idejama. Problemu je pristupljeno korišćenjem algoritama mašinskog učenja za predikciju plasmana pojedinačnih numera. Ulazni podaci su preuzeti iz skupa podataka „Song lyrics and entry metadata from Eurovision contests held in 1956-2021“. Na njima su primenjeni različiti oblici pretprocesiranja kako bi se postigla tačnija klasifikacija, a zatim su upotrebljeni algoritmi Naïve Bayes, Logistic Regression i Random Forest. Rezultati su potvrdili da postoji veoma slaba korelacija između tekstova i uspeha pesama. F-score modela se kreće između 0.5 i 0.6. Zaključeno je da nema magične formule, skupa reči koji drastično popravljaja šanse takmičara. Pesme se sastoje iz velikog skupa drugih faktora koji potencijalno imaju veći udeo u uspehu numere.

Ključne reči — mašinsko učenje, klasifikacija, Evrovizija, analiza teksta

I. Uvod

Evrovizija predstavlja muzičko takmičenje evropskih država, sa dugom istorijom održavanja. Na samom takmičenju svaki predstavnik nastupa sa pesmom kojom predstavlja svoju državu. Nakon svih nastupa otvaraju se glasanja i proglašava se pobenik nakon dodele poena od strane svih zemalja. Od 1980. godine okvirno učestvuje oko 30 država, što čini skup pesama mnogobrojnim i pogodnim za analizu.

Ideja ovog projekta je analiza tekstova i samih reči pesama sa Evrovizije, kako bi se objasnio njihov uticaj na plasman i popularnost. Neki od zaključaka koje težimo izvući su potencijalni šabloni koji konzistentno utiču na bolji ili lošiji plasmana na takmičenju, kao i uticaj pevanja na maternjem ili engleskom jeziku.

Da bi došli do željenih rezultata, ulazni skup podataka smo pretprocesirali, kako bi bio pogodniji ulaz u algoritme mašinskog učenja koje smo upotrebili, konkretno *Naïve Bayes*, *Logistic Regression* i *Random Forest*. Rezultati su imali prilično slab *F-Score*. Iz ovoga smo zaključili da tekstovi pesama nemaju veliki uticaj na njihov plasman na takmičenju.

Ostatak rada je struktuiran u sledeca poglavlja: poglavlje 2 - osvrt na radove sa sličnom tematikom, poglavlje 3 – opis skupa podataka i pretprocesiranje skupa podataka, poglavlje 4 – opis upotrebljenih algoritama mašinskog učenja i poređenje njihovih rezultata, poglavlje 5 - sažetak i zaključak.

II. Pregled postojeće relevantne literature

Prilikom odabira teme ovog rada naišli smo na velik broj radova koji pokrivaju sentimentalnu analizu tekstova pesama kao i analizu žanra. Kako smo planirali da analiziramo pesme sa takmičenja Evrovizije naišli smo na rad koji obrađuje politički i demografski uticaj država učesnica na bodovanje i njihov konačni plasman. Primetili smo da nismo našli ni jedan rad koji pokriva uticaj samih tekstova pesama na njihov plasman na takmičenju.

Zbog manjka radova vezanih za Evroviziju I analize pesama sa navedenog takmičenja, ugledali smo se najviše na radove koji se bave predviđanjem uspešnosti pesme, da li je uspešna(*hit*) ili ne, u nastavku slede neki od radova.

Istraživanje [1] bavi se pitanjem šta zapravo čini uspešnu pesmu, fokusirajući se na problem klasifikacije plesnih hitova. Izgrađena je baza *dance hit* pesama od 1985. do 2013. godine, uključujući osnovne muzičke karakteristike, kao i naprednije karakteristike koje obuhvataju vremenski aspekt. Brojni različiti klasifikatori(*C4.5 Tree, Naive Bayes, Logistic Regression, Support Vector Machines*) se koriste za pravljenje i testiranje modela predviđanja plesnih hitova. Model ima dobre performanse kada predvidi da li je pesma “top 10” *dance hit* u odnosu na nižu poziciju na listi.

U radu [2] cilj je da se nađe odgovor na pitanje “Postoji li magična formula za predviđanje hit pesme?”. Razmatraju tehničke parametre pesme da bi se predvideo uspeh. Prikupljali su podatke sa više platformi i kombinovanjem su napravili jedinstven skup podataka, dodata je *Boolean* promenljiva koja označava da li je pesma uspešna ili ne. Iz skupa podataka su izbačene pesme koje nisu na engleskom jeziku, kao i one sa nekonzistentnim vrednostima.. Primenjena su četiri algoritma za mašinsko učenje(*Naïve Bayes, Logistic Regression, Decision Tree, Random Forest*) . Za evaluaciju su uzeti u obzir tačnost i preciznost. Kombinovanjem tehničkih osobina pesama i analize sentimenta tekstova se pokazalo da je i dalje teško predvideti uspešnost pesme i da ona u mnogome zavisi od drugih eksternih faktora(ostvarena preciznost 52%). Utvrdili su da postoje elementi izvan tehničkih podataka koji mogu uticati na predviđanje da li je pesma hit ili ne. Ovaj rad zauzima stav da predviđanje muzike još uvek nije aktivnost analize podataka.

U radu [3] se takođe bave identifikacijom verovatne hit pesme. Izdvajaju akustične i informacije o tekstu iz svake pesme, razdvajaju uspešne i neuspešne pesme koristeći standardne klasifikatore, naročito *Support Vector Machines* i *boosting* klasifikatore. Videli su da su karakteristike zasnovane na stihovima nešto korisnije od akustičnih karakteristika u ispravnom identifikovanju uspešne pesme. Došli su do zaključka da spajanje ove dve karakteristike ne proizvodi značajna poboljšanja, takođe i da odsustvo određenih semantičkih informacija ukazuje da je veća verovatnoća da će pesma biti uspešna.

III. Metod

Ovo poglavlje sadrži detalje o koracima koji su izvršeni za prikupljanje podataka, njihovu obradu, kao i detalje o algoritmima mašinskog učenja koji su odabrani za predikciju.

A. Prikupljanje podataka

Glavni izvor podataka za ovaj projekat je „*Song lyrics and entry metadata from Eurovision contests held in 1956-2021*“ [5]. Ključne informacije se tiču izvornih tekstova, njihovih prevoda na engleski jezik, kao i njihovi plasmani i osvojeni bodovi, tabela 1.

#	Država	Pesma	Jezik	Plasman	Godina	Domaćin	Tekst
1	Switzerland	Refrain	French	2	1956	Switzerland	Chorus of love, or or or or Chorus, color of t...
2	Belgium	Straatdeuntje	Dutch	1	1957	West Germany	Along the streets, a tune is dancing And for a...
…							
1332	Ukraine	Shum	Ukrainian	5	2021	The Netherlands	Oh Spring song, Spring song Where have you spe...
1333	United Kingdom	Embers	English	22	2021	The Netherlands	Sometimes I know My fire burns low But as long...

Tabela 1. Prikaz podataka pre obrade

B. Pretprocesiranje podataka

Da bi ovi podaci bili korisni, potrebno je izvršiti pretprocesiranje. Preduzeti su sledeći koraci:

- U orignalnom skupu podataka, podaci u svakoj pesmi su čuvani u po kolonama, umesto po vrstama. Matrica je transponovana kako bi imala smisla za dalju obradu.
- Pesme koje su već bile na engleskom su imale praznu kolonu *Translated Lyrics*, da bi se ova neregularnost eleminisla kolona je popunjena originalnim tekstom.
- Kako bi svaka pesma imala prevod, izbačene su pesme na izmišljenom jeziku.
- Brojčani tipovi koji su originalno bili string su konvertovani u int.
- Kolone koje nisu bile od interesa su bile izbačene(npr. Redni broj nastupa na takmičenju, grad održavanja Evrovizije).
- Prevedeni tekstovi su dodatno procesirani
  - Ukonjeni su novi redovi(“\n”)
  - Stop reči, znakovi interpunkcije su takođe eleminisani
  - Odrađena je lematizacija reči da se svedu da korenski oblik reči
  - Sve reči su svedene na mala slova
- Pojedine kolone su preimenovane kako bi bolje reflektovale sadržaj

Na kraju pretprocesiranja, skup podataka izgleda kao u tabeli 2.

#	Država	Pesma	Jezik	Plasman	Godina	Domaćin	Tekst
1	Switzerland	Refrain	French	1	1956	Switzerland	chorus love or chorus color t...
2	Belgium	Straatdeuntje	Dutch	1	1957	West Germany	along the streets tune dancing for...
...							
1332	Ukraine	Shum	Ukrainian	1	2021	The Netherlands	spring song spring song where have you spe...
1333	United Kingdom	Embers	English	0	2021	The Netherlands	Sometimes i know my fire burns low ...

Tabela 2. Podaci nakon pretprocesiranja

- Neke pesme su upotrebljavale više jezika. Da bi se proces pojednostavio, sve pesme su ograničene na jedan primaran jezik.
- Pesme iz prvih nekoliko godina takmičenja su imale vrednost „-“ u koloni *Placement.* Ove numere su izbačene iz skupa podataka.

C. Ulazni parametri

Parametri koji su odabrani kao ulaz u algoritme su:

- Država izvođača(*Country*): Ovaj atribut je primarno zadržan da bi bilo mogućnosti za poređenje sa kolonom jezika na kojem je pesma izvedena(*Language*).

- Jezik na kojem je pesma izvedena(*Language*): Pokazatelj efikasnosti pevanja na maternjem, odnosno stranom, najčešće engleskom jeziku.
- Godina odžavanja takmičenja(*Year*): Pretposavka je da se trendovi popularnih reči menjaju kroz vreme, pa je iz tog razloga ovaj parametar zadržan.
- Država domaćin(*Host*): Pretpostavka je da različite zemlje gaje različita raspoloženja prema upotrebljenim rečima.
- Reči pesme(*Words*): Kolona izvedena iz *Translated Lyrics* primenom pretprocesiranja opisanog u prethodnom poglavlju.

D. Procesiranje tekstova pesama

Dve varijante procesiranja kolone *Words* su upotrebljene.

U prvoj, da bi se smanjio broj reči, kolona je redukovana na X najčešće ponovljenih reči za svaku pesmu uparen sa *Bag of Words vectorizer*-om. Ovakvo procesiranje je prilično jednostavno i smanjuje količinu podataka, ali se njime gubi težina reči koje su ponovljene neproporcionalan broj puta u pesmi(npr. prva reč se ponavlja 14 puta, dok se druga ponavlja samo 3 puta).

Druga varijanta podrazumeva upotrebu *TF-IDF vectorizer*-a kako bi se preciznije odredila važnost svake reči u pesmi.

Vectorizer-i su potrebni kako bi omogućili algoritmima mašinskog učenja da rade nad podacima predstavljeni u ne brojevnom obliku, kao što su, u ovom slučaju, reči pesme. Praktično kreiraju matricu pojavljivanja reči u svakoj pesmi.

Upotreba *TF-IDF* se bolje pokazala, po cenu duže obrade.

E. Određivanje izlaznog parametra

Izlazni parametar predstavlja plasman pesme na takmičenju(*Placement*).

U prvoj iteraciji smo pokušali da predvidimo tačno koje mesto će pesma postići. Rezultati su bili veoma loši, a naš zaključak da ovaj problem nije realistično rešiv pomoću datih ulaznih parametra i veličine skupa podataka.

Problem smo pretvorili u znatno jednostavniji, gde smo prosto zamenili kolonu *Placement* sa *bool* vrednostima, *true* za pesme koje su bile u top 10, i *false* za pesme sa lošijim plasmanom. Za ovaj broj smo se odlučili jer je znatno fleksibilnije pogoditi top 10 u odnosu na npr. pobedničku pesmu, ali da se raspored ne svede na 50/50 kao što bi bio slučaj da smo tražili top ~20. Primena algoritama mašinskog učenja na ovako formulisan problem je dala znatno bolje rezultate. [1]

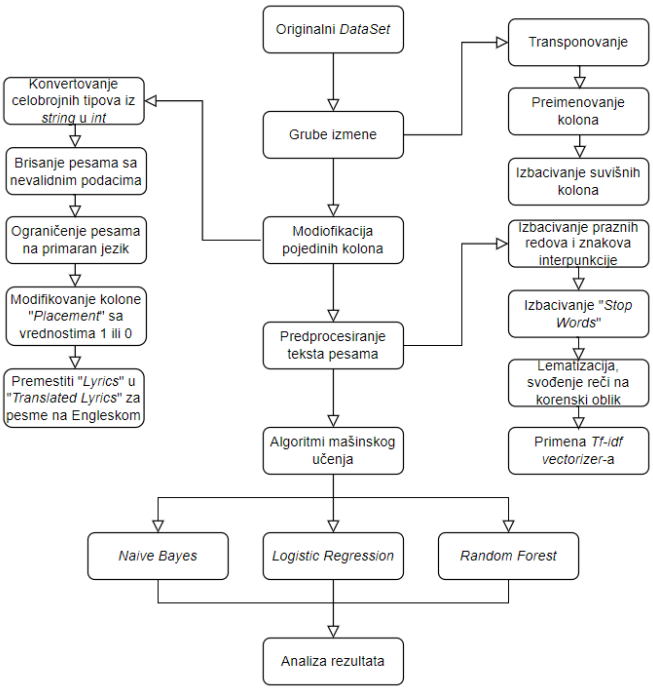
F. Algoritmi mašinskog učenjm

Da bi se razvio model predviđanja uspešnosti teksta pesme na Evroviziji, korišćen je niz algoritama mašinskog učenja. Tri algoritma su odabrana na osnovu njihove prethodne primene u sličnom kontekstu [2, 4] - *Logistic Regression*, *Random Forests* i *Naïve Bayes*.

Logistic Regression: Za rešavanje problema klasifikacije, logistic regression je najosnovnija i najpopularnija među algoritmima mašinskog učenja[4]. Model se zasniva na predviđanju verovatnoće uspeha korišćenjem logističke funkcije.

Decision Tree & Random Forests: Razlog da se uključi stablo odlučivanja(engl. *decision tree*) bilo je da se ponudi predviđanje sa slučajem gde nisu ispunjene pretpostavke logističke regresije. Sposoban da radi sa nelinernim podacima razlikuje se od većine algoritama za mašinsko učenje. Pošto algoritam koristi jednu funkciju po čvoru da podeli podatke, izgradnja stabla oblučivanja je brz proces, sa većom preciznošću rezultata. Random Forests se pominje kao proširenje *Decision Tree*-a, predstavlja algoritam mašinskog učenja koji se može koristiti za klasifikacije kao I rešavanje problema regresije. Funkcioniše kao ansambl koji se sastoji od mnogih stabala odlučivanja [7] koja rade uz konsenzus da se dobije tačniji ishod od pojedinačnog stabla.

Naïve Bayes: To je nagledani algoritam mašinskog učenja na osnovu Bajesove teoreme koja se koristi za rešavanje problema klasifikacije prateći probabilistički pristup. Zove se "Naïve" jer se pretpostavlja da su korišćene karakteristike nezavisne jedne od drugih[6]. Ovaj algoritam je izabran jer zahteva malu količinu podataka o obuci za trening koji će proivesti efikasne rezultate.



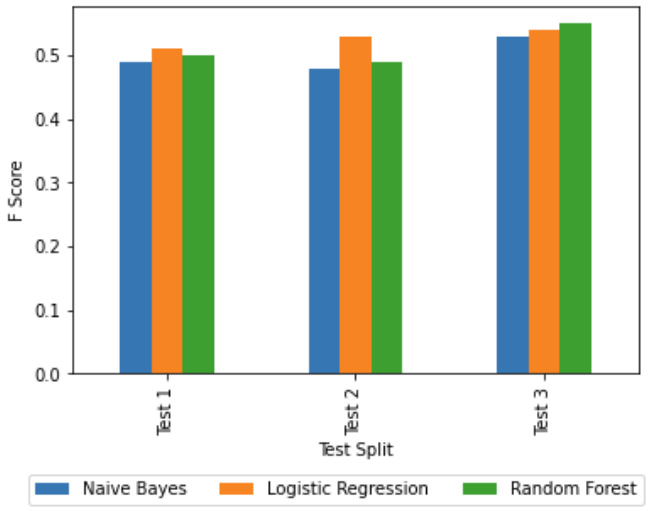
Slika 3: Grafik procedure rada

IV. REZULTATI I DISKUSIJA

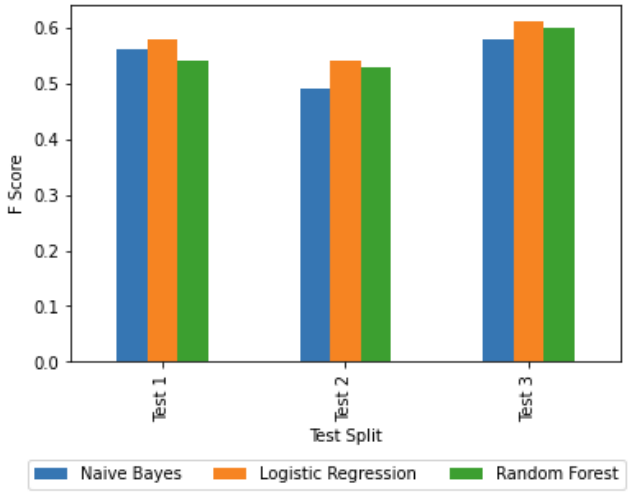
Po ugledu na ostale radove, odnos skupa za učenje i skupa za test je 80%/20%.

Za rangiranje rezultata korišćen je *F-score*, kao solidna metrika koja uzima u obzir i preciznost(*Precision*) kao i povrat(*Recall*). Rezultati upotrebljenih algoritama na 3

različita *seed*-a podele *test*/*train* su prikazani ispod:



Slika 4: Poređenje F-Score-a upotrebljenih algoritama na 3 različita test seed-a(Bag of Words Vectorizer)



Slika 5: Poređenje F-Score-a upotrebljenih algoritama na 3 različita test seed-a(TF-IDF Vectorizer)

Najbolje se pokazao *Logistic Regression*, ali razlika nije velika. Pretpostavka je da bi se greska mogla smanjiti uzimanjem u obzir dodatnih faktora kao što su sentiment ili tempo. Takođe, zbog ogromnog broja reči upotrebljenog u pesmama, i retko ponavljanje nekih od njih, postoji mogućnost da je upotrebljeni skup podataka nedovoljno obiman. Veći skup bi potencionalno dao bolje rezultate.

Rezultati potvrđuju da tekst pesme nije od presudnog značaja za njen plasman. Ne postoji odabran skup reči koje drastično varira uspeh pesama na takmičenju. Svaka pesma se sastoji iz mnogo šireg skupa obeležja koji imaju uticaj na njeu popularnost, kao što su ritam, glasnoća, žanr itd. Rezultati mogu da ohrabre tekstopisce jer se može reći da ne postoji "*meta*" koja se treba slediti za najbolji plasman.

ZAKLJUČAK

Ovim radom smo pokušali da utvrdimo da li tekstovi pesama na takmičenju imaju uticaj na plasman. Ugledali smo se na radove sa sličnim problemom i algoritme korišćene za njihovo rešavanje, konkretno *Logistic Regression*, *Random Forests* i *Naïve Bayes*. Na osnovu dobijenih rezultata, došli smo do zaključka da ne postoji magičan skup reči koji značanjno povećava šanse pesme da se što bolje plasira. Očekivali smo jaču korelaciju između skupa reči i plasmana, međutim možda bi se pored samog teksta trebalo uzeti u obzir i drugi faktori koji čine samu pesmu, poput ritma, akustičnosti, glasnoće itd. koji bi doveli do drugačijih rezultata.

LITERATURA

[1] D. Herremans, D. Martens, and K. Sörensen, “Dance hit song prediction”, Journal of New Music Research, 43(3), 291–302,2014.

[2] A.H.RazaJul, K.Nanath, “Predicting a Hit Song with Machine Learning: Is there an apriori secret formula?” 2020. [Online]. Link: “https://www.researchgate.net/publication/344216655_Predicting_a_Hit_Song_with_Machine_Learning_Is_there_an_apriori_secret_formula:

[3] R. Dhanaraj and B. Logan, “Automatic prediction of hit songs,” in Proceedings of International Society for Music Information Retrieval, pp. 11–15, 2005.

[4] Jaehyun Kim, “Music Popularity Prediction Through Data Analysis of Music’s Characteristics” 2021. [Online]. Link: http://article.ijsts.net/pdf/10.11648.j.ijsts.20210905.16.pdf

[5] Skup podataka - „Song lyrics and entry metadata from Eurovision contests held in 1956-2021“. Link: https://www.kaggle.com/datasets/minitree/eurovision-song-lyrics.

[6] Z. Lateef, "Comprehensive Guide To Logistic Regression In R | Edureka", Edureka, 2019. [Online]. Link: https://www.edureka.co/blog/logistic-regression-in-r/.

[7] T. Yiu, "Understanding Random Forest", Medium, 2019. [Online]. Link: https://towardsdatascience.com/understanding-random-forest-58381e0602d2.