Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 4.9
DOI: https://doi.org/10.24867/30BE13Santrac

PREDIKCIJA ŽANRA FILMA UPOTREBOM MAŠINSKOG UČENJA

MOVIE GENRE PREDICTION USING MACHINE LEARNING

Stefan Santrač, *Fakultet tehničkih nauka, Novi Sad*

Oblast – PRIMENJENE RAČUNARSKE NAUKE I INFORMATIKA

Kratak sadržaj – *Prilikom izrade novih filmova često dolazi do curenja informacija. Takve informacije su nepotpune i na osnovu njih možemo saznati ko radi na filmu, kada se film premijerno prikazuje, naslov filma, itd. Da bi se formirala puna slika o filmu potrebno je poznavati i kom žanru film pripada. U ovom radu vrši se predikcija žanra filma korišćenjem Multilabel klasifikatora (Multilabel k Nearest Neighbours, Classifier chains, Binary relevance) na osnovu informacija o osobama koje rade na filmu, koji posao te osobe obavljaju, naslova filma, godini premijernog prikazivanja i informacije da li je film namenjen za decu. Rezultati predikcija su evaluirani pomoću Micro-average i Macro-average metoda evaluacij.*

Ključne reči: — *film; predikcija žanra; multilabel klasifikacija; mašinsko učenje*

Abstract – *When creating new films, information often leaks. Such information is incomplete and can reveal details like who is working on the film, when it will premiere, the film's title, etc. To get a complete picture of the film, it's necessary to know the genre it belongs to. In this study, genre prediction is performed using Multilabel classifiers (Multilabel k-Nearest Neighbors, Classifier Chains, Binary Relevance) based on information about the people working on the film, their roles, the film's title, the year of its premiere, and whether the film is intended for children. The prediction results are evaluated using Micro-average and Macro-average evaluation methods.*

Keywords: *film; genre prediction; multilabel classification; machine learning*

1. UVOD

Žanr je koncept koji se koristi u filmskim studijama i teoriji filma za opisivanje sličnosti između grupa filmova zasnovanih na estetskim ili širim društvenim, institucionalnim, kulturnim i psihološkim aspektima. Filmski žanr ima sličnosti u formi i stilu, temi i komunikativnoj funkciji. Filmski žanr se stoga zasniva na skupu konvencija koje utiču kako na produkciju pojedinačnih dela u okviru tog žanra, tako i na očekivanja i iskustva publike. Žanrovi se koriste u industriji, u proizvodnji i marketingu filmova, od strane filmskih analitičara i kritičara u analizi filma, i kao okvir za publiku u odabiru i doživljaju filmova.

Svaki film može da pripada grupi više žanrova, što utiče na njegovu popularnost, kao i na grupu ljudi kojima je namenjen i koji će ga pogledati. Žanr filma u velikoj meri zavisi od glumaca koji se u njemu pojavljuju, režisera, scenarista, godine prikazivanja, itd. Živimo u svetu društvenih mreža i široko rasprostranjenih medija, gde svakodnevno procure informacije o budućoj saradnji poznatih glumaca i režisera, godini u kojoj će film premijerno biti prikazan, kao i mnoge druge informacije vezane za film. Ljubiteljima filmova širom sveta je cilj da saznaju i sam žanr filma, kako bi znali da li im taj film upada u sferu interesovanja i da bi odlučili da li će pratiti njegov razvoj.

Upravo ovim problemom se i bavi ovaj rad. U njemu će biti prikazano više različitih rešenja za predikciju žanra filma na osnovu glumačkog kadra, producenta, režisera, scenariste, kompozitora muzike za film, kinematografa, da li je film namenjen za decu, naslova filma i godine kada je film premijerno prikazan. Sličan problem je rešavan u drugim naučnim radovima, gde je rađena predikcija žanra na osnovu različitih parametara, kao što su filmski posteri, radnja filma ili kratki reklamni prikaz filma (eng. *trailer*). Za *multilabel* klasifikaciju korišćene su metode *Multilabel k Nearest Neighbours*, *Binary Relevance* i *Classifier Chains*.

U ovom radu korišćeno je više metoda predikcije i one će biti poređene pomoću *Micro-average* i *Macro-average* metoda evaluacije. Najbolje rezultate su dale *Classifier Chains* metoda na osnovu Micro-average evaluacije i *Multilabel k Nearest Neighbours* metoda na osnovu *Macro-average* evaluacije. Rezultati predikcija su definisani i rangirani i te informacije se mogu iskoristiti u budućim radovima koji se bave sličnim temama.

2. METODOLOGIJA

U ovom poglavlju je predstavljena implementacija sistema za predikciju žanra filma na osnovu ulaznih podataka kao što su tip/format sadržaja (npr. film, serija, video...), naslov sadržaja koji su tvorci koristili u promociji filma, da li je sadržaj namenjen za decu, godina kada je sadržaj premijerno prikazan, osobe koje rade na filmu i posao koji te osobe obavljaju. Očekivani izlaz sistema predstavlja skup do najviše 3 žanra filma.

Pre same primene metoda za predikciju žanra filma bilo je potrebno izvršiti izmene nad skupom podataka kako bi podaci bili spremni za obuku i testiranje modela. Pošto su se podaci za polja *primaryTitle*, *isAdult*, *startYear*, *genres*, *category*, *primaryName* u skupu podataka nalazila

NAPOMENA:
Ovaj rad proistekao je iz master rada čiji mentor je bio prof. dr Aleksandar Kovačević, red. prof.

u tekstualnom obliku izvršena je njihova konverzija u numeričke vrednosti.

Term frequency-inverse document frequency (TF-IDF) je tehnika pronalaženja informacija koja meri frekvenciju termina (TF) i njegovu inverznu frekvenciju dokumenta (IDF). Svaka reč ili termin koji se pojavljuje u tekstu ima svoj TF i IDF rezultat. TF predstavlja broj pojavljivanja tražene reči u svim dokumentima, što pokazuje koliko je određeni termin bitan. Dok IDF predstavlja logaritam količnika ukupnog broja dokumenata i broja dokumenata u kojima se ovaj termin pojavljuje, on smanjuje važnost termina koji se često pojavljuju u dokumentima. Konkretno u ovom radu TF-IDF je jedna od metoda korišćena za konverziju ulaznih podataka iz tekstualnog formata u numeričke vrednosti pre prosleđivanja u model.

Druga tehnika korišćena za konverziju tekstualnih podataka u vektore numeričkih vrednosti je *Doc2Vec*. On predstavlja generalizaciju *Word2Vec* tehnike. *Word2Vec* može da napravi procene visoke tačnosti o značenju reči na osnovu njihovog pojavljivanja u tekstu. Ove procene daju asocijacije reči sa drugim rečima, na primer, reči poput „kralj“ i „kraljica“ bile bi veoma slične jedna drugoj.

Word2Vec reprezentacija je kreirana korišćenjem 2 algoritma: *Continuous Bag-of-Words model* (CBOW) i *Skip-Gram model*. *Continuous Bag-of-Words* predviđa trenutnu reč na osnovu konteksta, odnosno na osnovu okolnih reči. Korišćenjem *Continuous Bag-of-Words* algoritma se koriste reči „mačka”, „je”, „sela” za predviđanje reči „na”.

Skip-gram algoritam radi suprotno od CBOW algoritma: umesto da se svaki put predviđa jedna reč, koristi se jedna reč da bi se predvidele sve okolne reči, odnosno kontekst.

Cilj *Doc2Vec-a* je da kreira numerički prikaz dokumenta, bez obzira na njegovu dužinu. Dokumenti nisu iste logičke strukture kao reči, algoritam koji rešava ovaj problem je proširenje CBOW modela. Umesto korišćenja samo reči za predviđanje naredne reči, dodaje se još jedan vektor obeležja (Paragraf id), koji je jedinstven za dokument. Ovaj metod se naziva *Distributed Memory version of Paragraph Vector* (PV-DM).

Kao i kod *Word2Vec-a*, može se koristiti drugi algoritam, sličan *Skip-gramu*, koji se naziva *Distributed Bag of Words version of Paragraph Vector* (PV-DBOW).

Dodatno za *Doc2Vec* je prvo potrebno prikupiti sve reči iz labela i proslediti ih *Doc2Vec* modelu kako bi mogao da se napravi rečnik, zatim se taj model obučava i koristi se *infer_vector* metoda kako bi se reči iz skupa podataka transformisale u vektore numeričkih vrednosti.

Multilabel embedding tehnike [1] pojavile su se kao odgovor na potrebu da se nosi sa velikim prostorom za labele, ali sa razvojem računara postale su metod za poboljšanje kvaliteta klasifikacije. U ovom radu korišćen je *LabelNetwork Embeddings*.

S obzirom da film može pripadati grupi više žanrova, u ovom radu će se koristiti *Multilabel k Nearest Neighbours* (ML-KNN) algoritam, *Binary Relevance* algoritam i *Classifier Chains* algoritam za *Multilabel* klasifikaciju. Pomoću njih će se vršiti klasifikacija filmova na više žanrova. Dobijeni rezultati klasifikacije će se evaluirati i potom međusobno porediti.

Multilabel k Nearest Neighbours (ML-KNN) je *lazy learning* algoritam. Kao što mu ime implicira, ML-KNN je izveden iz popularnog *algoritma k Nearest Neighbours* (KNN) [2]. Na početku se identifikuju k najbližih suseda u trening skupu. Zatim, prema statističkim informacijama dobijenim iz skupova labela ovih susednih instanci se određuje skup labela za test instancu. Nakon transformacije podataka iz tekstualnog formata u numeričke vrednosti, korišćenjem neke od gore navedenih metoda za word embedding, prosleđuju se ML-KNN metodi u obliku *dense* matrice i vrši se predikcija.

Binary relevance [3], transformiše problem klasifikacije sa N labela u N *single-label* odvojenih problema binarne klasifikacije koristeći zadati binarni klasifikator. U ovom radu *Binary relevance* metodi prosleđen je SVC kao binarni klasifikator. *Binary relevance* metod zahteva da se ulazni podaci nalaze u obliku *sparse* matrice. Rečenice pretvorene u numeričke vrednosti pomoću gore navedenih *word embedding* metoda će se nalaziti u obliku *dense* matrice, zato je potrebno transformisati je u *sparse* oblik.

Porodica metoda poznata kao *Classifier chains* [4] postala je popularan pristup *Multilabel learning* problemima. Ovaj pristup uključuje povezivanje standardnih binarnih klasifikatora u lančanu strukturu, tako da predviđanja klase labela postaju obeležja za druge klasifikatore. *Classifier chains-u* je zadat SVC kao binarni klasifikator. Kao što je slučaj i kod *Binary relevance* metode i *Classifier chains* metoda zahteva dodatnu konverziju podataka u *sparse* oblik.

Micro-average preciznost je zbir svih pozitivnih rezultata predviđanja i deli se sa zbirom svih pozitivnih i negativnih rezultata predviđanja. U osnovi to je količnik broja tačno identifikovanih predviđanja sa ukupnim brojem predviđanja.

Macro-average precinznost predstavlja prosečnu preciznost sistema nad različitim skupovima.

Macro-average će izračunati metriku nezavisno za svaku klasu, a zatim će uzeti prosečnu vrednost i tako tretirati sve klase podjednako, dok *će Micro-average* agregirati doprinose svih klasa za izračunavanje prosečne metrike.

Cilj ovog rada je predikcija žanra filma na osnovu prosleđenih parametara. Poređeni su rezultati predikcija metoda *Multilabel k Nearest Neighbours*, *Classifier chains* i *Binary relevance* korišćenjem *Micro-average* i *Macro-average* metoda evaluacije. Nakon dobijenog finalnog skupa podataka urađen je *shuffle* podataka kako bi se obezbedilo da modeli ostanu opšti i da se izbegne *overfit*. Nakon toga nad njim je izvršeno čišćenje podataka. Popunjene su nedostajuće vrednosti, uklonjeni su razmaci, dijakritici i zagrade. Pored TF-IDF i *Doc2Vec*, za konverziju reči u numeričke vektore korišćene su i metode Label encoder i *Multilabel binarizer*.

U *Multilabel* klasifikaciji, *Micro-average* je poželjniji ukoliko postoji neuravnoteženost klasa, tj. ima mnogo više primera jedne klase nego drugih klasa. Na osnovu

postojanja disbalansa žanrova filmova u finalnom skupu podataka, zaključeno da je *Micro-average* bolji metod evaluacije u ovom slučaju.

3. OPIS SKUPA PODATAKA

Inicijalni skupovi podataka su preuzeti sa zvaničnog sajta "imdb.com." [5] U ovom radu se koriste tri skupa podataka, svaki skup podataka se nalazi u *tab-separated values* (tsv) formatu. Podaci su filtrirani na osnovu vrednosti polja titleType, gde su uzeti u obzir samo podaci čija je vrednost tog polja *'movie'*. Kako je atribut originalTitle napisan u izvornom jeziku i manje je zastupljen od primaryTitle-a odlučeno je da se ovaj atribut zanemari. Prethodno je navedeno da skup podataka sadrži samo filmove, pa atribut endYear postaje suvišan. Kako je cilj ovog rada predviđanje žanra još neobjavljenog filma atribut runtimeMinutes neće biti poznat, pa će iz tog razloga biti uklonjen iz finalnog skupa podataka.

Kako u finalnom skupu podataka već imamo osobe koje učestvuju u izradi filma, polja birthYear i deathYear ne utiču na predikciju žanra filma, pa su uklonjena. Atribut primaryProfession postaje suvišan iz razloga što se u atributu category već navodi zanimanje kojim se osoba bavila u izradi filma. KnownForTitles je polje koje navodi najpoznatija kinematografska dela u čijoj je izradi određena osoba učestvovala, a za predikciju žanra filma veću važnost ima broj ostvarenih uloga u filmovima određenog žanra što je već prikazano u skupu podataka, iz tog razloga atribut knownForTitles će biti uklonjen.

Nakon filtriranja, tri skupa podataka se spajaju u finalni skup podataka. Spajanje se vrši na osnovu atributa tconst, odnosno nconst, koji će nakon spajanja biti izbačeni jer predstavljaju jedinstvene identifikatore. Finalni skup podataka se sastoji od atributa:

- titleType – tip/format sadržaja (npr. film, serija, video...)
- primaryTitle – naslov sadržaja koji su tvorci koristili u promociji filma
- isAdult – da li je sadržaj namenjen za decu
- startYear – godina kada je sadržaj premijerno prikazan
- genres – skup do najviše 3 žanra filma
- category – kategorija posla osobe na filmu
- primaryName – ime osobe koja radi na filmu

Ciljno obeležje predstavlja atribut genres, gde se njegova vrednost predviđa na osnovu vrednosti ostalih podataka u skupu. Atribut genres predstavlja skup do najviše 3 žanra filma i može uzimati vrednosti iz skupa: *action*, *adventure*, *animation*, *biography*, *comedy*, *crime*, *documentary*, *drama*, *family*, *fantasy*, *film-noir*, *history*, *horror*, *musical*, *music*, *mystery*, *romance*, *sci-fi*, *sport*, *thriller*, *war*, *western*.

4. REZULTATI I DISKUSIJA

Prilikom kreiranja *Doc2Vec* modela optimizovani su sledeći parametri: 'vector_size' = 64, 'window' = 2, ' min_count' = 1, ' workers' = 8, 'epochs' = 20. Finalni skup podataka nakon transformacija je podeljen na trening i test skup u odnosu 80/20, pri čemu je iz test skupa izbačena vrednost ciljne labele.

Kao što je već navedeno u metodologiji, zbog disbalansa žanrova u finalnom skupu podataka zaključeno je da je Micro-average bolji metod evaluacije u ovom slučaju, a na osnovu rezultata iz tabele 1 optimizovana vrednost broja suseda u *Multilabel k Nearest Neighbours* metodi iznosi k = 7.

ML-KNN	Micro-average	Macro-average
k = 3	0.32	0.11
k = 4	0.33	0.13
k = 5	0.38	0.08
k = 6	0.37	0.06
k = 7	**0.42**	0.06
k = 8	0.39	0.10

Tabela 1. *Rezultati Mulitlabek k Nearest Neighbours metode*

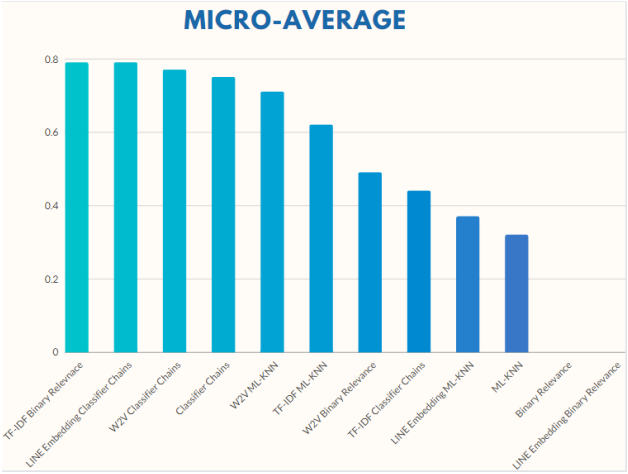
Optimizovane vrednosti za Label Network Embeddings su: 'weighted' = True, 'include_self_edges' = False, 'batch_size' = 1000, 'order' = 3, korišćen je LINE *embedding* metod ('LINE'), 'dimension' = 5, 'aggregation_function' = 'add', 'normalize_weights' = True, kao *regresor* je korišćen *RandomForestRegressor* sa parametrom 'n_estimators' = 10.

Rezultati dobijeni korišćenjem metodologija opisanih u poglavlju II prikazani su u tabeli 2.

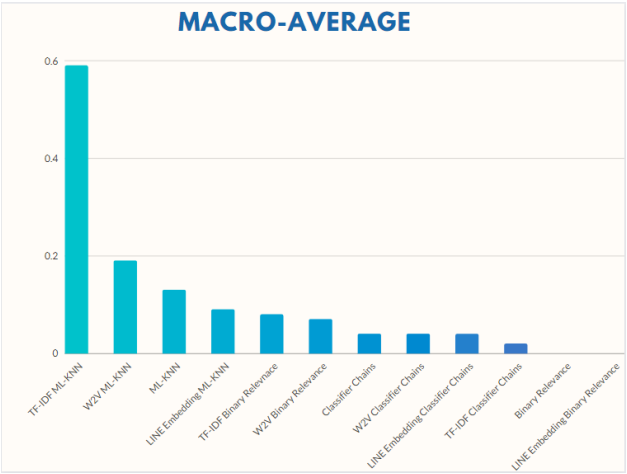
Classifiers	Micro-average	Macro-average
ML-KNN	0.32	0.13
Binary Relevance	0.00	0.00
Classifier chains	0.75	0.04
TF-IDF ML-KNN	0.62	**0.59**
TF-IDF Binary Relevance	**0.79**	0.08
TF-IDF Classifier chains	0.44	0.02
W2V ML-KNN	0.71	0.19
W2V Binary Relevance	0.49	0.07
W2V Classifier chains	0.77	0.04
LINE embedding ML-KNN	0.37	0.09
LINE embedding Binary Relevance	0.00	0.00
LINE embedding Classifier chains	0.79	0.04

Tabela 2. *Rezultati predikcije*

Na osnovu slike 1 se uočava da se *Classifier chains* metoda u proseku najbolje pokazala za *Micro-average* evaluaciju. Jedino odstupanje od očekivanog rezultata ostvarila je metoda *TF-IDF Binary Relevance* sa tačnošću od 0.79 i pokazala se kao najbolja metoda za TF-IDF *embedding*. Dok je korišćenjem drugih *word embedding* tehnika ostvarila najlošije rezultate, kao što je i očekivano jer je ovaj metod najjednostavniji ali ujedno i najbrži.



Slika 1. *Rezultati predikcije evaluirani Micro-average metodom*



Slika 2. *Rezultati predikcije evaluirani Macro-average metodom*

Sa slike 2 vidljivo je da se ML-KNN metoda najbolje pokazala za *Macro-average* evaluaciju i TF-IDF ML-KNN je ostvarila ubedljivo najbolju tačnost od 0.59. Metode predikcije daju znatno slabije rezultate evaluacijom pomoću *Macro-average* metode u odnosu na evaluaciju *Micro-average* metodom, što je i očekivano ponašanje zbog disbalansa žanrova u finalnom skupu podataka.

Kao što je već navedeno, u ovom radu najbolji rezultati za ML-KNN metodu su ostvareni kada je broj suseda k = 7. U skupu podataka su postojala ograničenja u vidu nedostajućih vrednosti za žanr i godinu premijernog prikazivanja, pa su torke sa tim nedostajućim vrednostima izbačene iz finalnog skupa podataka.

4. ZAKLJUČAK

Problem čijim se rešavanjem bavi ovaj rad je predviđanje žanrova filma na osnovu dostupnih podataka o filmu kao što su: osobe koje rade na filmu, godine premijernog prikazivanja filma, da li je film preporučljiv za decu, naslov filma. Motivacija za rešavanje ovog problema je sve veće interesovanje publike za buduće filmove i njihova želja da saznaju što više informacija o filmu kako bi mogli da znaju da li ti filmovi ulaze u njihovu sferu interesovanja. Često se desi da informacije o filmovima koji su u fazi planiranja ili izrade dospeju u javnost i od takvih informacija nastaju novinski članci o pojedinostima filma. Rešavanje ovog problema bi takođe olakšalo novinarima spekulacije o žanru filma.

U prethodnim sekcijama opisan je postupak vršenja predikcije žanra filma korišćenjem tehnika mašinskog učenja. Prvi korak ka rešavanju ovog problema bio je spajanje tri skupa podataka u finalni skup podataka. Finalni skup podataka je bilo potrebno ograničiti i očistiti od nepotrebnih i nevalidnih podataka. Zatim je izvršena transformacija tekstualnih podataka u numeričke vrednosti. Nakon toga su kreirani predikcioni modeli (*Multilabel k Nearest Neighbours*, *Classifier chains* i *Binary relevance*) i izvršena je njihova optimizacija. Poslednji korak u rešavanju ovog problema je bila evaluacija metoda korišćem Micro-average i *Macro-average* evaluacija. Kao najbolja rešenja pokazali su se *Binary Relevance* korišćenjem TF-IDF word embedding-a sa tačnošću od 0.79 i *Classifier chains* korišćenjem LINE embedding-a takođe sa tačnošću od 0.79.

Takođe ovaj rad prilikom rešavanja problema predikcije žanra filma koristi više *Multilabel* klasifikatora i poredi ih. Rezultati poređenja mogu biti analizirani, da bi se izdvojilo najbolje rešenje koje će biti iskorišćeno u budućim radovima koji se bave rešavanjem sličnog problema.

4. LITERATURA

[1] Piotr Szymański, Tomasz Kajdanowicz (2017) Multi-label embedding-based classification Available: http://scikit.ml/multilabelembeddings.html[datum pristupa 10.07.2024.]
[2] D.W. Aha, Special AI review issue on lazy learning, Artif. Intell. Review 11
[3] Zhang, Min-Ling, et al. "Binary relevance for multi-label learning: an overview." Frontiers of Computer Science 12.2 (2018): 191-202.
[4] Read, Jesse, et al. "Classifier chains for multi-label classification." Machine learning 85.3 (2011): 333-359.
[5] [Online]Available: https://www.imdb.com/interfaces/ [datum pristupa 10.07.2024.]

Kratka biografija:

Stefan Santrač rođen je u Novom Sadu 1998. god. Master rad na Fakultetu tehničkih nauka iz oblasti Računarstva i automatike – Primenjene računarske nauke i informatika odbranio je 2024.god.
kontakt: santrac.stefan@gmail.com