***Predikcija žanra filma***

Aleksa Santrač Stefan Santrač Nemanja Dimitrijević

Fakultet tehničkih nauka Fakultet tehničkih nauka Fakultet tehničkih nauka

Univerzitet u Novom Sadu Univerzitet u Novom Sadu Univerzitet u Novom Sadu

Trg Dositeja Obradovića 6 Trg Dositeja Obradovića 6 Trg Dositeja Obradovića 6

21000 Novi Sad 21000 Novi Sad 21000 Novi Sad

[aleksantrac@gmail.com](mailto:aleksantrac@gmail.com) [santrac.stefan@gmail.com](mailto:santrac.stefan@gmail.com) n.dimitrijevic1297@gmail.com

*Apstrakt*—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. Informacije se u modernom svetu šire velikom brzinom zbog postojanja interneta i društvenih mreža, ali mnoge od njih su neproverene ili netačne. Ovaj rad nastoji da sa što većom tačnošću korisnicima da informacije o žanru filma i samim tim bi se smanjila količina netačnih informacija na internetu. 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 evaluacije. Prilikom predikcije na osnovu Micro-average evaluacije najbolje se pokazala Classifier chains metoda, dok se na osnovu Macro-average evaluacije najbolje pokazao Multilabel k Nearest Neighbours metoda.

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

# Uvod

Tri najvažnija elementa na kojima počinje kinematografija su fotografije, iluzija pokreta i projekcija. Spojivši ove elemente Luj i Ogist Limijer su 1895. god. snimili prvi film. Sa razvojem filmske produkcije u Holivudu, pored postojećih, javljaju se i novi žanrovi: *western* film, slapstick komedija, gengsterski film i mjuzikl. Druga polovina 20. veka donela je obilje filmskih ostvarenja. Od velikih umetničkih dela visokih estetskih dometa, do značajnih komercijalnih uspeha izraženih u više desetina miliona dolara. [[1](#ref1)]

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.

Detaljniji opis podataka i rešenja izložen je u ostatku rada. Naredno poglavlje se bavi pregledom radova koji se bave sličnom tematikom. Poglavlje III sadrži prikaz algoritama korišćenih za predikciju i metoda korišćenih za evaluaciju rezultata predikcije kao i tehnike korišćene za konverziju reči u numeričke vrednosti. Poglavlje IV opisuje inicijalni i finalni skup podataka i prikazuje korake koji su urađeni kako bi se od inicijalnog skupa došlo do konačnog skupa. Nakon toga u poglavlju V opisan je postupak dobijanja rezultata, njihovo prikazivanje i analiziranje. U poglavlju VI dat je kratak zaključak o radu i navedeni su koraci koji se mogu preduzeti kako bi se rad unapredio i dao još bolje rezultate.

# Pregled postojeće relevantne literature

Prilikom istraživanja radi boljeg upoznavanja sa tematikom ovog rada, pronađeno je nekoliko radova u kojima je obrađen sličan problem. Neki od njih su navedeni u nastavku teksta.

U radu [[2](#ref2)] izvršena je Multilabel klasifikacija žanra filma na osnovu kratkog reklamnog prikaza filma, što predstavlja i sam cilj rada. Autori teksta su razvili arhitekturu neuronske mreže – Convolution-Trough-Time for Multi-label Movie genre Classification (CTT-MMC). Dizajnirana je da iskoristi karakteristike kadrova kratkog reklamnog sadržaja tokom vremena.

Skup podataka korišćen za ovaj rad je LMTD (Labelled Movie Trailer Dataset) koji obuhvata 10000 kratkih reklamnih prikaza filmova koji spadaju u 22 kategorije žanra filma. Podaci su dodeljeni prema IMDB meta-podacima. Skup podataka je podeljen u tri celine: trening skup (70% podataka), validacioni skup (10% podataka) i testni skup (20% podataka).

Za evaluaciju rezultata su korišćeni: Micro-average, Macro-average i Weighted-average. Autori su koristili različite algoritme, gde su koristili multilabel klasifikaciju i uspeli su da postignu maksimalnu tačnost između 0.646 i 0.742, gde je za dobijeni rezultat 0.742 korišćena Micro-average evaluacija. Metodologija navedenog i našeg rada će se razlikovati, ali u našem radu ćemo primeniti iste metode evaluacije, jer su namenjene za Multilabel klasifikaciju.

U radu [[3](#ref3)] autori su koristili iterativni relacioni klasifikator suseda sa više oznaka koji koristi karakteristike društvenog konteksta (SCRN).

Umreženi podaci sa društvenih mreža, internet stranica i različitih bibliografskih baza podataka, mogu da sadrže entitete koji pripadaju više klasa. Cilj je da se korišćenjem Multilabel klasifikacije, instancama u mreži podataka dodeli više labela.

Korišćena su 3 različita skupa podataka. Prvi skup podataka je DBLP Dataset koji obezbeđuje milione referenci na naučne radove iz oblasti računarstva. U ovoj mreži svaki autor predstavlja čvor, gde su 2 autora povezana ukoliko su sarađivali na bar 2 projekta. Za ovaj skup podataka autori su odabrali 15 reprezentativnih konferencija u 6 istraživačkih oblasti na osnovu kojih će se vršiti klasifikacija. Drugi skup podataka je IMDB Dataset. Internet filmska baza podataka (IMDb) je onlajn baza podataka informacija u vezi sa filmovima, televizijskim programima i video igricama, uključujući informacije o rediteljima, glumcima i zapletima. Treći skup podataka je YouTube Dataset, koji je izvučen sa YouTube veb-sajta, koji predstavlja popularan veb-sajt za deljenje video sadržaja. Za prvi skup podataka, trening skup obuhvata od 5% do 30% ukupnog DBLP skupa podataka. Za drugi skup podataka, trening skup obuhvata od 1% do 20% ukupnog IMDB skupa podataka. Dok za treći skup podataka, trening skup obuhvata od 1% do 9% ukupnog YouTube skupa podataka.

Kao mera evaluacije su korišćene metode: Micro-F1, Macro-F1 i Hamming Loss. Na osnovu rezultata ovog naučnog rada zaključujemo da je IMDB Dataset bio pogodan za Multilabel klasifikaciju i dostignut je procenat tačnosti između 45.62% i 71.98% kod Micro-F1 evaluacije i 18.46% do 43.31% kod Macro-F1 evaluacije. Metodologija ovog naučnog rada i našeg će se razlikovati, ali na osnovu ostvarenih rezultata, zaključili smo da je IMDB Dataset validan skup podataka i primenjen je u našem radu.

U radu [[4](#ref4)] kao skup podataka korišćeni su funkcionalni podaci o genu Kvasca u stvarnom svetu. Svaki gen je povezan sa skupom funkcionalnih klasa čiji maksimalna veličina potencijalno može biti veća od 190. Skup podataka je podeljen na trening skup, koji obuhvata 1500 gena, dok testni skup obuhvata 917 gena.

Postojalo je više načina za Multilabel klasifikaciju, ali Multilabel lazy learning pristup nije bio dostupan. Cilj ovog rada je prezentovanje Multilabel k Nearest Neighbor (ML-KNN) algoritma, kao novog načina za Multilabel lazy learning pristup.

U ovom radu korišćen je Multilabel k Nearest Neighbor (ML-KNN) algoritam za Multilabel klasifikaciju. Kao mera evaluacije su korišćene metode: Hamming Loss, One-error, Coverage, Ranking Loss, Average Precision.

Pokazano je da broj suseda ne utiče na vrednost evaluacije prilikom korišćenja metode Hamming Loss, ali za ostale metode evaluacije ML-KNN algoritam je najbolje rezultate ostvario kada je broj suseda bio k=7. U našem radu je korišćena ML-KNN metodologija koja je prezentovana u ovom radu.

Zadatak rada [[5](#ref5)] je komparacija metoda za Multilabel learning, rangiranje metoda po preformansama kao i efikasnosti (vremenu izvršavanja). Cilj rada je pronalaženje najbolje metode za Multilabel learning korišćenjem 16 različitih metoda evaluacije.

U ovom radu autori su koristili 12 metoda: ML-KNN, ML-C4.5, PCT, Binary Relevance, Classifier Chains, Calibrated Label Ranking, Quick Weighted Algorithm For MLL, Hierarchy Of Multilabel classifiers, Random k-labelsets, Ensembles of classifier chains, RF-MLC4.5 (Random Forest), RF-PCT (Random Forest). Mere evaluacije korišćene u ovom radu su: Hamming loss, Accuracy, Precision, Recall, F1 score, Subset accuracy, Macro precision, Macro recall, Macro F1, Micro precision, Micro recall, Micro F1.

Korišćeno je 11 različitih skupova. Ovi skupovi potiču iz tri različita domena: Biologije, Multimedije i Tekst kategorizacije. Skupovi podataka variraju od 391 do 60000 trening primera, od 202 do 27856 testnih primera i od 6 do 983 labela.

Analiza rezultata metoda za Multilabel klasifikaciju je pokazala da su metode sa najboljim rezultatima Random forests of predictive clustering trees (RF-PCT) i Hierarchy of multilabel classifiers (HOMER), zatim Binary relevance (BR) i Classifier chains (CC). U našem radu smo koristili metodologije Binary Relevance i Classifier Chains koje su se dobro pokazale kao metodologije za Multilabel klasifikaciju.

# Metodologija

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 u tekstualnom obliku izvršena je njihova konverzija u numeričke vrednosti.

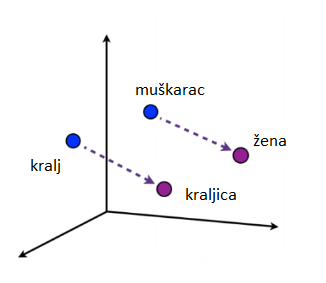
Term frequency-inverse document frequency (TF-IDF) je tehnika pronalaženja informacija koja meri frekvenciju termina (TF) i njegovu inverznu frekvenciju dokumenta (IDF), što je prikazano u formuli [3](#_Slika_3._Term). Svaka reč ili termin koji se pojavljuje u tekstu ima svoj TF i IDF rezultat. TF predstavlja broj pojavljivanja tražene reči u svim dokumentima, što pokazuje koliko je određeni termin bitan, formula [1](#_Formula_1._Term). 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, prikazan je u formuli [2](#_Formula_2._Inverse). 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.

###### Formula 1. Term frequency (TF)

###### Formula 2. *Inverse document frequency* (IDF)

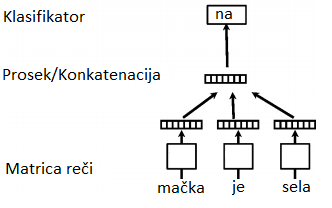
###### Formula 3. *Term frequency-inverse document frequency* (TF-IDF)

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, slika [1](#_Slika_1._Asocijacija).



###### Slika 1. Asocijacija između reči “kralj”-“ kraljica” i “muškarac”-“žena”

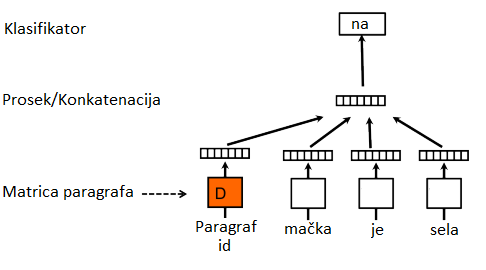
Word2Vec reprezentacija je kreirana korišćenjem 2 algoritma: Continuous Bag-of-Words model (CBOW) i Skip-Gram model. Continuous Bag-of-Words predviđa trenutnu reč na osnovu konteksta, odnosno na osnovu okolnih reči. Korišćenjem Continuous Bag-of-Words algoritma se koriste reči „mačka”, „je”, „sela” za predviđanje reči „na”, primer na slici [2](#_Slika_2._Skica).



###### Slika 2. SkicaCBOWalgoritma

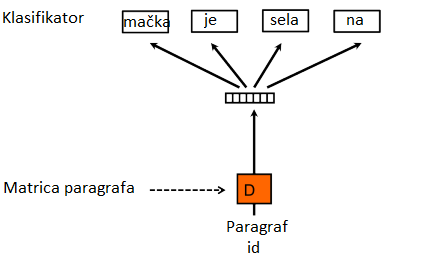
Skip-gram algoritam radi suprotno od CBOW algoritma: umesto da se svaki put predviđa jedna reč, koristi se jedna reč da bi se predvidele sve okolne reči, odnosno kontekst.

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, što je prikazano na slici [3](#_Slika_7._Razlika). Ovaj metod se naziva Distributed Memory version of Paragraph Vector (PV-DM).



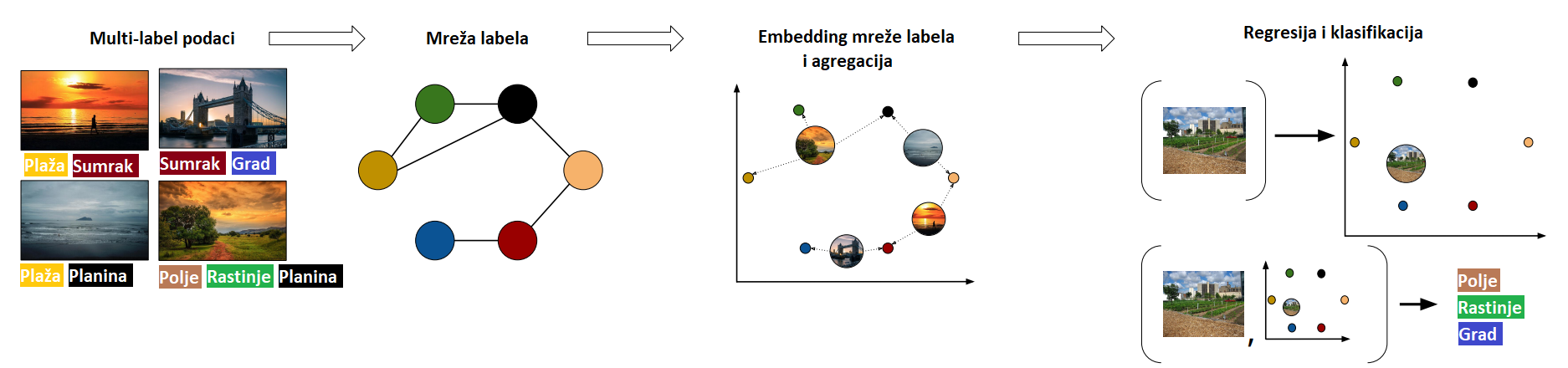
###### Slika 3. *Distributed Memory version of Paragraph Vector* (PV-DM)

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), prikazan na slici [4](#_Slika_8._Distributed).



###### Slika 4. *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.

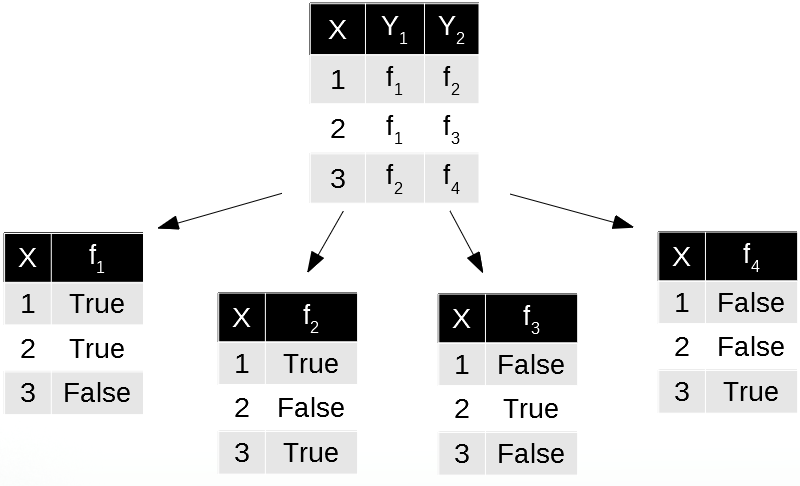
**Multilabel embedding tehnike [[6](#ref6)] pojavile su se kao odgovor na potrebu da se nosi sa velikim prostorom za labele, ali sa razvojem računara postale su metod za poboljšanje kvaliteta klasifikacije. U ovom radu korišćen je LabelNetwork Embeddings koji je prikazan na slici [5](#_Slika_9._Label).

###### Slika 5. Dijagram *Label Network Embeddings* za *Multilabel* klasifikacionu šemu

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) [[7](#ref7)]. 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 [[8](#ref8)], čiji je primer prikazan na slici [6](#_Slika_10._Binary), 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.



###### Slika 6. *Binary relevance*

Porodica metoda poznata kao Classifier chains [[9](#ref9)] 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, što je prikazano na slici [7](#_Slika_11._Classifier). 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.



###### Slika 7. *Classifier chains*

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. Primer računanja Micro-average preciznosti prikazan je u formuli [4](#_Slika_12._Micro-average), gde TP predstavlja broj tačnih predviđanja, a FP broj netačnih predviđanja za svaku klasu.

###### Formula 4. *Micro-average* preciznost

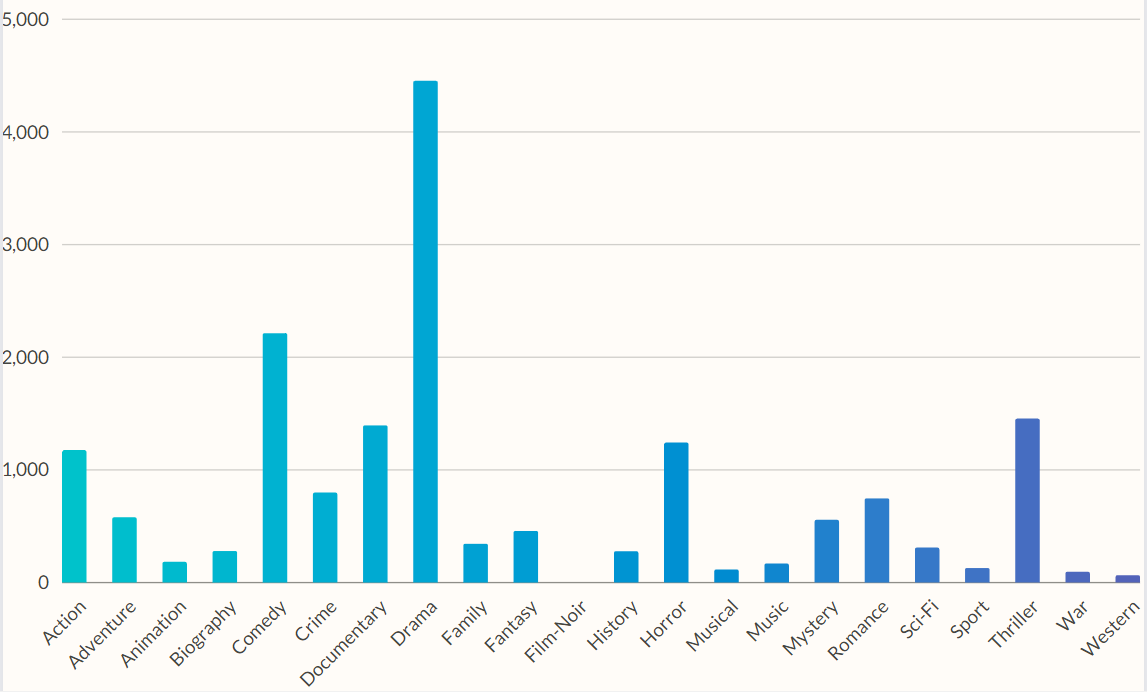
Macro-average precinznost predstavlja prosečnu preciznost sistema nad različitim skupovima. Primer računanja Macro-average preciznosti prikazan je u formuli [5](#_Slika_13._Macro-average), gde je .

###### Formula 5. *Macro-average* preciznost

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, koji su navedeni i objašnjeni u prethodnom poglavlju. 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 slici [8](#_Slika_15._Distribucija_1) je prikazan disbalans žanrova u finalnom skupu podataka i na osnovu toga je zaključeno da je Micro-average bolji metod evaluacije u ovom slučaju.



###### Slika 8. Distribucija žanrova filmova

# Opis skupa podataka

Inicijalni skupovi podataka su preuzeti sa zvaničnog sajta “imdb.com.” [[10](#ref10)] U ovom radu se koriste tri skupa podataka, svaki skup podataka se nalazi u tab-separated values (tsv) formatu.

Prvi skup podataka title.basic.tsv.gz sadrži:

* tconst – jedinstveni identifikator sadržaja
* titleType – tip/format sadržaja (npr. film, serija, video...)
* primaryTitle – naslov sadržaja koji su tvorci koristili u promociji filma
* originalTitle – prvobitni naslov sadržaja
* isAdult – da li je sadržaj namenjen za decu
* startYear – godina kada je sadržaj premijerno prikazan
* endYear – godina kada je završeno prikazivanje serije, kod svih ostalih tipova sadržaja vrednost je '\N'
* runtimeMinutes – trajanje sadržaja u minutama
* genres – skup do najviše 3 žanra filma

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.

Drugi skup podataka podataka title.principals.tsv.gz sadrži:

* tconst – jedinstveni identifikator filma
* ordering – broj jedinstvenih redova za zadati titleId
* nconst – jedinstveni identifikator osobe koja radi na filmu (glumac, režiser, scenarista…)
* category – kategorija posla osobe na filmu
* job – naziv specifičnog posla ukoliko postoji, u suprotnom vrednost je '\N'
* characters – ime uloge u filmu ukoliko postoji, u suprotnom vrednost je '\N'

Atribut ordering nije imao svrhu u skupu podataka u ovom radu, i zbog toga je isključen. Polje job blisko je vezano za polje category, i tačno određuje kojim poslom se bavila osoba koja radi na filmu, s obzirom da polje category dovoljno precizno opisuje kojim poslom se osoba bavila u izradi filma i da je u većini slučajeva vrednost polja job bila '\N', odlučeno je da se ovo polje izuzme iz konačnog skupa podataka. Atribut characters predstavlja ulogu koju glumac glumi u određenom filmu i ne utiče na žanr filma i iz tog razloga je izbačen.

Treći skup podataka podataka name.basics.tsv.gz sadrži:

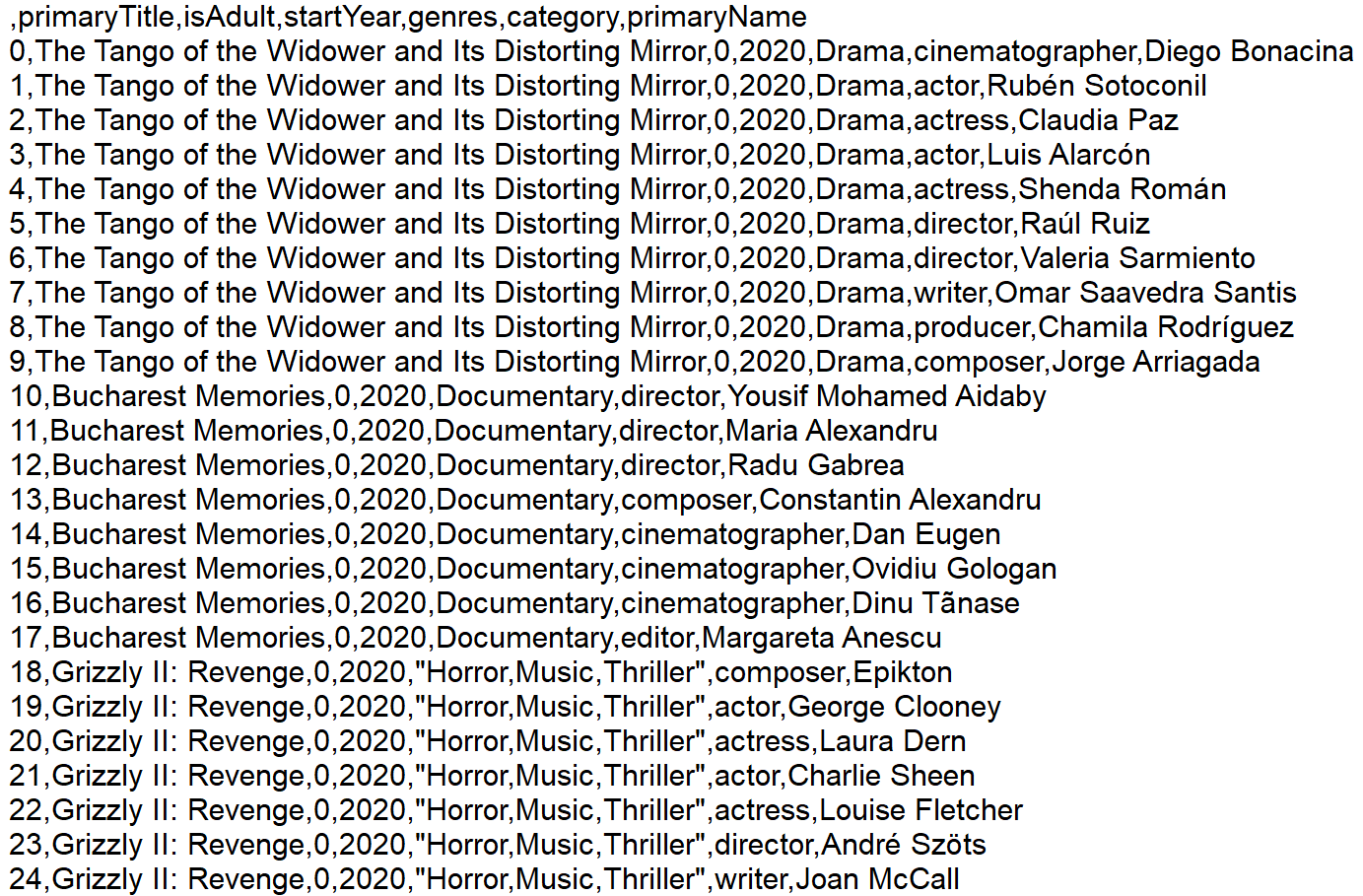
* nconst – jedinstveni identifikator osobe koja radi na filmu (glumac, režiser, scenarista...)
* primaryName – ime osobe koja radi na filmu
* birthYear – godina rođenja
* deathYear – godina smrti ukoliko postoji, u suprotnom vrednost je '\N'
* primaryProfession – tri najčešća zanimanja kojima se ova osoba bavila
* knownForTitles – najpoznatija kinematografska dela u čijoj je izradi ova osoba učestvovala

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, prikazan na slici [9](#_Slika_14._Deo). 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. S obzirom da finalni skup podataka sadrži nekoliko miliona torki, što zahteva veliki broj sati procesiranja i veliku količinu radne memorije, odlučeno je da se skup ograniči na samo filmove koji su premijerno prikazani posle 2020. godine i od njih je napravljen podskup od 10000 torki, kako bi se postigla željena efikasnost.



###### Slika 9. Deo finalnog skupa podataka

# 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 (po ugledu na rad [[2](#ref2)]), 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](#_Tabela_1._Rezultati) 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 |

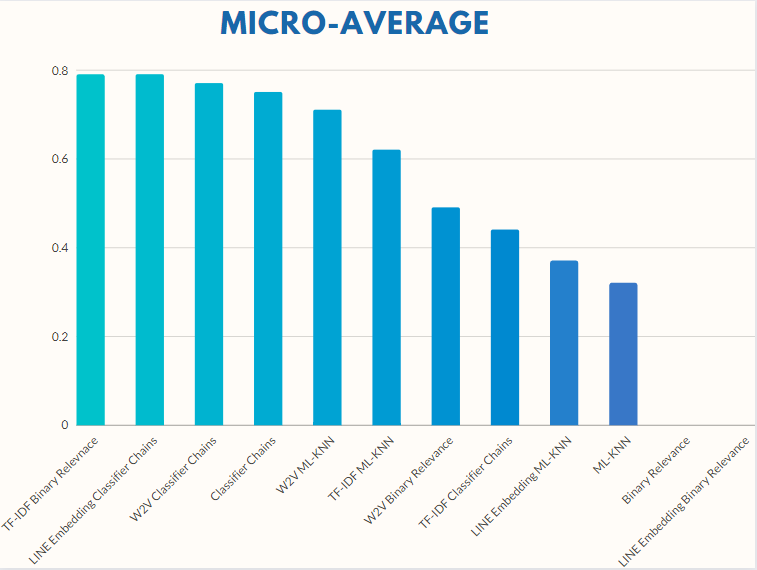
###### Tabela 1. Rezultati *Multilabel k Nearest Neighbours* metode

Optimizovane vrednosti, preuzete sa sajta [[11](#ref11)], 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.

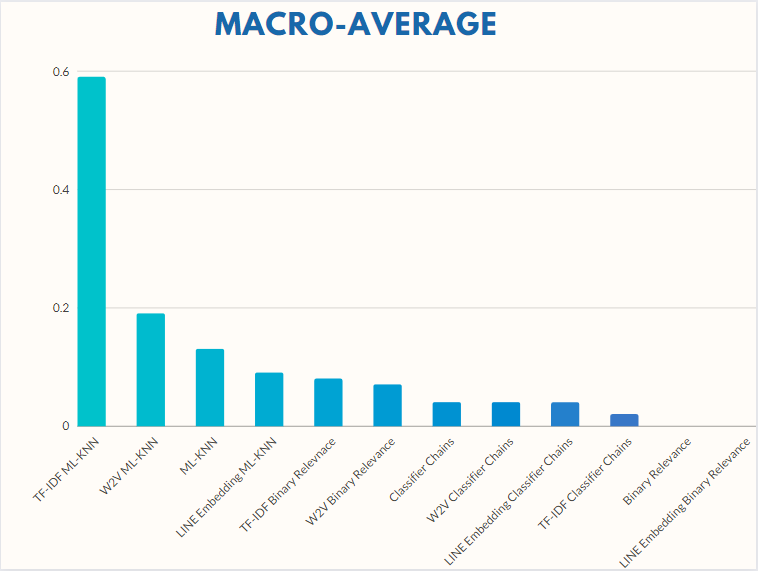
Rezultati dobijeni korišćenjem metodologija opisanih u poglavlju [III](#_Metodologija) prikazani su u tabeli [2](#_Tabela_2._Rezultati).

|  |  |  |
| --- | --- | --- |
| **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



###### Slika 10. Rezultati predikcije evaluirani *Micro-average* metodom



###### Slika 11. Rezultati predikcije evaluirani *Macro-average* metodom

Na osnovu slike [10](#_Slika_16._Rezultati) 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, što se može zaključiti iz poglavlja [III](#_Metodologija).

Sa slike [11](#_Slika_17._Rezultati) 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, što se slaže sa već publikovanim radom [[4](#ref4)]. 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.

# 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 koriščenjem 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 iskoršćeno u budućim radovima koji se bave rešavanjem sličnog problema.

Rad bi se mogao unaprediti proširivanjem skupa podataka, podacima kao što su:

* Vreme trajanja filma
* Država u kojoj se snima film
* Jezik na kom je film snimljen
* Budžet

# Literatura

1. Gimnazija “Vuk Karadžić”, Loznica, “Sve što ste hteli da znate o filmu“.
2. Jônatas Wehrmann, Rodrigo C. Barros (2017) Movie genre classification: A multi-label approach based on convolutions through time [Online]Available: https://www.sciencedirect.com/science/article/abs/pii/S1568494617305112 [Accessed: 20-Apr-2022]
3. Xi Wang, Gita Sukthankar (2013) Multi-Label Relational Neighbor Classification using Social Context Features [Online]Available: <https://dl.acm.org/doi/pdf/10.1145/2487575.2487610> [Accessed: 20-Apr-2022]
4. Min-Ling Zhang and Zhi-Hua Zhou (2005) A k-Nearest Neighbor Based Algorithm for Multi-label Classification [Online]Available: <https://www.researchgate.net/profile/Min-Ling-Zhang-2/publication/4196695_A_k-nearest_neighbor_based_algorithm_for_multi-label_classification/links/565d98f408ae1ef92982f866/A-k-nearest-neighbor-based-algorithm-for-multi-label-classification.pdf> [Accessed: 20-Apr-2022]
5. Gjorgji Madjarov, Dragi Kocev, Dejan Gjorgjevikj, Saso Dzeroski (2012) An extensive experimental comparison of methods for multi-label learning [Online]Available: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.710.3585&rep=rep1&type=pdf> [Accessed: 20-Apr-2022]
6. Piotr Szymański, Tomasz Kajdanowicz (2017) Multi-label embedding-based classification [Online]Available: <http://scikit.ml/multilabelembeddings.html> [Accessed: 22-Apr-2022]
7. D.W. Aha, Special AI review issue on lazy learning, Artif. Intell. Review 11
8. Zhang, Min-Ling, et al. "Binary relevance for multi-label learning: an overview." Frontiers of Computer Science 12.2 (2018): 191-202.
9. Read, Jesse, et al. "Classifier chains for multi-label classification." Machine learning 85.3 (2011): 333-359.
10. [Online]Available: <https://www.imdb.com/interfaces/> [Accessed: 23-Apr-2022]
11. [Online]Available: <http://scikit.ml/multilabelembeddings.html> [Accessed: 23-Apr-2022]