

Analiza uticaja filmskih titlova i žanrova na tržišni uspeh i priznanja filmova

Tim: Laslo Uri E2 163/2023, Denis Dautović E2 166/2023, Nikola Janković E2 130/2025

1. Definicija cilja projekta

Cilj projekta je razvoj sistema za **predikciju tržišnog uspeha filmova** na osnovu analize sadržaja filmskih titlova i karakteristika žanrova. Sistem koristi tehnike obrade prirodnog jezika (NLP) i mašinskog učenja za binarnu klasifikaciju filmova, kao i za predikciju verovatnoće nominacija za filmske nagrade.

Konkretni zadaci:

1. Klasifikacija filmova po tržišnom uspehu (uspešan/neuspešan) na osnovu ROI
 2. Predikcija verovatnoće nominacije za Oscar na osnovu žanra i karakteristika titlova
 3. Identifikacija koje karakteristike titlova (sentiment, leksička raznovrsnost, žanr) najviše utiču na uspeh filma
-

2. Motivacija problema

Filmska industrija generiše preko 100 milijardi dolara godišnje, ali investicije u produkciju nose visok rizik. Sistem za predikciju uspeha filmova ima praktičnu vrednost za:

- **Producente:** Informisanije odluke o investicijama u projekte
- **Streaming platforme:** Unapređenje preporuka i predviđanje popularnosti sadržaja
- **Istraživače:** Kvantitativna analiza odnosa narativnog sadržaja i prijema publike

Kombinacija analize tekstualnog sadržaja (titlova) sa žanrovskim i finansijskim podacima predstavlja pristup koji nije dovoljno istražen u postojećoj literaturi.

3. Relevantna literatura

3.1 Detecting Emotional Scenes Using Semantic Analysis on Subtitles [1]

Cilj rada: Razvoj sistema za automatsku detekciju emotivnih scena iz filmova koristeći NLP tehnike primenjene na titlove.

Metodologija: Implementirana su dva klasifikatora: (a) naivni brojački klasifikator zasnovan na frekvenciji reči i semantičkim konstruktima; (b) MaxEnt klasifikator sa analizom unigrama i emotivnih reči.

Skup podataka: 8 filmova različitih žanrova (Gladiator, Love Actually, Remember the Titans, X-men trilogija, Transformers, Troy) sa manualno anotiranih ~8.000 rečenica, grupisanih u segmente od po 20 rečenica.

Evaluacija: Testiranje na svakom filmu pojedinačno sa ručno anotiranim gold standard labelama. Mera: tačnost klasifikacije (accuracy).

Rezultati: Naivni klasifikator postigao prosečnu tačnost 62% (maks. 80%), dok je MaxEnt klasifikator dostigao 78% na filmu Remember the Titans.

Relevantnost: Studija demonstrira efikasnost NLP tehnika u analizi filmskih titlova. U našem projektu proširujemo ovaj pristup kombinovanjem sa žanrovskim i finansijskim karakteristikama, koristeći TF-IDF umesto jednostavnog brojanja reči.

3.2 Screenplay Quality Assessment: Can We Predict Who Gets Nominated? [2]

Cilj rada: Automatska evaluacija kvaliteta filmskih scenarija sa fokusom na predviđanje nominacija za Oscar.

Metodologija: TF-IDF transformacija teksta, SVM klasifikacija, kombinacija lingvističkih, emocionalnih i strukturnih karakteristika. Poređenje sa deep learning modelima (BERT).

Skup podataka: ScriptBase korpus (897 scenarija) i Movie Screenplay Corpus (868 scenarija). Binarne labele: nominovan/nije nominovan.

Evaluacija: Podela 80/10/10 (trening/validacija/test). Mere: F1-score, precision, recall.

Rezultati: Feature-based pristup (TF-IDF + SVM) nadmašio deep learning metode (BERT) sa F1-score od 62.35% (ScriptBase) i 64.79% (MSC).

Relevantnost: Direktno potvrđuje mogućnost predikcije Oscar nominacija na osnovu teksta. Koristićemo slične feature engineering tehnike, ali umesto scenarija analiziramo titlove i dodajemo finansijske podatke. Koristićemo ensemble metode umesto samo SVM.

3.3 Success in Books: A Big Data Approach to Bestsellers [3]

Cilj rada: Analiza obrazaca prodaje knjiga sa liste bestselera New York Times-a i razvoj modela za predikciju ukupne prodaje na osnovu ranih prodajnih podataka.

Metodologija: Statistički model zasnovan na tri mehanizma: (1) fitness knjige (kvalitet), (2) preferential attachment (popularnost privlači više kupaca), (3) aging (opadanje interesa tokom vremena). Kombinacija ovih faktora u jednu formulu za predikciju prodaje.

Skup podataka: 2468 fiction i 2025 nonfiction bestselera sa NYT liste (2008-2016), sa nedeljnim podacima o prodaji iz NPD BookScan.

Evaluacija: Fitovanje modela na prvih 25-50 nedelja prodaje, predikcija ukupne prodaje. Mera: R^2 za kvalitet fita.

Rezultati: Model tačno predviđa ukupnu prodaju knjige na osnovu prvih 6 meseci ($R^2 > 0.9$). Fikcija se prodaje bolje od nefikcije. Početna pozicija na listi predviđa dužinu ostanka.

Relevantnost: Iako se bavi knjigama, principi analize uspeha kulturnih proizvoda su primenjivi na filmove. Koristićemo sličan pristup za analizu odnosa žanra i uspeha, ali fokusirano na predikciju PRE izlaska filma (ne post-hoc analiza prodaje).

4. Skup podataka

Skup podataka ćemo sami sastaviti kombinovanjem više izvora:

Izvori podataka:

- **TMDB API** (themoviedb.org) — naslov, godina, žanrovi, budžet, box office prihod
- **OpenSubtitles API** (opensubtitles.stoplight.io) — tekstovi titlova filmova
- **Academy Awards Database** (awardsdatabase.oscars.org) — istorijski podaci o Oscar nominacijama

Atributi za predikciju:

- Žanrovi filma (action, drama, comedy, horror...)
- NLP karakteristike titlova (TF-IDF, sentiment, leksička raznovrsnost)
- Budžet filma (log-transformisan)

Ciljno obeležje:

- **Tržišni uspeh:** Klasa 1 ako je $ROI \geq 2.0$ (prihod $\geq 2 \times$ budžet), inače klasa 0. Labele računamo automatski iz budžeta i prihoda.
- **Oscar nominacija:** Klasa 1 ako je film nominovan, inače klasa 0. Labele preuzimamo iz Academy Awards Database.

Očekivani obim: ~5.000-10.000 filmova (2000-2024) sa kompletnim podacima (budžet, prihod i titl). Očekujemo nebalansirane klase — većina filmova nije uspešna niti nominovana.

5. Metodologija

Nakon prikupljanja podataka o filmovima, spojićemo podatke iz TMDB API-ja (naslov, žanr, budžet, prihod) sa titlovima iz OpenSubtitles-a koristeći naslov i godinu kao ključ. Filmove bez kompletnih podataka (nedostaje budžet, prihod ili titl) ćemo isključiti iz analize.

Ekstrakcija karakteristika iz titlova: Za svaki film ćemo iz teksta titlova izvući: (1) TF-IDF vektore najčešćih reči i fraza koje se pojavljuju u dijalozima, (2) sentiment score koji meri da li su dijalozi pretežno pozitivni ili negativni, (3) leksičku raznovrsnost koja pokazuje bogatstvo vokabulara u filmu. Ove karakteristike ćemo kombinovati sa žanrom filma (action, drama, comedy...) koji ćemo enkodirati kao binarne attribute.

Modelovanje: Ulaz u klasifikatore će biti: žanrovi filma, NLP karakteristike titlova i odnos budžeta. Izlaz je binarna predikcija: uspešan ($ROI \geq 2$) ili neuspešan film. Upoređićemo Random Forest, XGBoost i SVM jer su se ovi modeli pokazali kao efikasni za tekstualne karakteristike u relevantnoj literaturi [2]. S obzirom da očekujemo više neuspešnih nego uspešnih filmova, primenićemo SMOTE za balansiranje klasa pre treniranja.

Analiza rezultata: Analiziraćemo koje karakteristike titlova najviše doprinose predikciji uspeha — npr. da li filmovi sa pozitivnijim sentimentom imaju veći ROI, ili da li određeni žanrovi (action vs drama) pokazuju različite obrasce uspeha.

6. Metod evaluacije

6.1 Postupak evaluacije

Podatke ćemo podeliti na trening (70%), validacioni (15%) i test skup (15%) sa stratifikovanim uzorkovanjem kako bismo očuvali proporciju klasa. Za robustniju procenu performansi koristićemo 5-fold stratified cross-

validaciju na trening skupu, dok će finalna evaluacija biti izvršena na izdvojenom test skupu koji model nikada nije video.

6.2 Mere evaluacije

S obzirom da očekujemo nebalansirane klase (više neuspešnih filmova nego uspešnih, više nenominiranih nego nominiranih), **accuracy nije adekvatna mera**. Kao primarnu meru koristićemo **F1-score** jer balansira precision i recall, što je ključno kada je minority klasa (uspešni filmovi / nominovani filmovi) od većeg interesa. Dodatno ćemo pratiti **AUC-ROC** za procenu sposobnosti modela da rangira filmove po verovatnoći uspeha.

Za poređenje različitih modela (Random Forest vs XGBoost vs SVM) koristićemo paired t-test nad rezultatima cross-validacije. Analiziraćemo matricu konfuzije kako bismo identifikovali tipove grešaka — npr. da li model češće pogrešno klasifikuje niskobudžetne hitove ili viskobudžetne propuste.

7. Softver (opciono)

- **Programski jezik:** Python 3.10+ sa Jupyter Notebook okruženjem
- **Data wrangling:** Pandas, NumPy
- **ML:** Scikit-learn (Random Forest, SVM), XGBoost, Imbalanced-learn
- **NLP:** SpaCy, NLTK, Gensim (Word2Vec)
- **Vizuelizacija:** Matplotlib, Seaborn, Plotly

8. Plan rada

Kontrolne tačke:

- **Prva kontrolna tačka:** Kompletan dataset, inicijalna EDA, baseline model
- **Druga kontrolna tačka:** Optimizovani modeli, evaluacija, error analysis
- **Finalna predaja:** Kompletan sistem sa dokumentacijom i izveštajem

Potencijalni rizici: Nedostajući podaci za neke filmove, računarska kompleksnost NLP obrade, nebalansirane klase. Mitigacija: filtriranje podataka, batch processing, SMOTE/undersampling.

Reference

[1] Detecting Emotional Scenes Using Semantic Analysis on Subtitles, Stanford NLP Course Project, 2009. <https://nlp.stanford.edu/courses/cs224n/2009/fp/6.pdf>

[2] M.-C. Chiu, T. Feng, X. Ren, S. Narayanan, "Screenplay Quality Assessment: Can We Predict Who Gets Nominated?", arXiv:2005.06123, 2020. <https://arxiv.org/abs/2005.06123>

[3] Y. Yucesoy et al., "Success in books: a big data approach to bestsellers", EPJ Data Science, vol. 7, 2018. <https://link.springer.com/article/10.1140/epjds/s13688-018-0135-y>