Зборник радова Факултета техничких наука, Нови Сад

UDK: 4.9
DOI: https://doi.org/10.24867/30BE11Lekic

ПРЕДИКЦИЈА ПОПУЛАРНОСТИ YOUTUBE TRENDING ВИДЕО СНИМКА

PREDICTING THE POPULARITY OF A YOUTUBE TRENDING VIDEO

Душан Лекић, *Факултет техничких наука, Нови Сад*

Област – ЕЛЕКТРОТЕХНИКА И РАЧУНАРСТВО

Кратак садржај – *Са порастом популарности друштвених мрежа и платформи за дељење видео садржаја, предвиђање популарности видео записа постало је важно за ауторе садржаја, људе који се баве маркетингом као и оглашиваче. У овом раду истражујемо различите факторе који утичу на популарност YoyTубе trending видео снимка, као што су број лајкова, број дислајкова, број и сентимент коментара, наслов и слика самог видео снимка, као и карактеристике аутора, односно његовог YouTube канала.*

Кључне речи: *предикција популарности, Youtube, неуронске мреже, NLP, Computer Vision*

Abstract – *With the rise in popularity of social networks and video sharing platforms, predicting video popularity has become important for content creators, marketers and advertisers alike. In this paper, we investigate various factors that influence the popularity of a YouTube trending video, such as the number of likes, the number of dislikes, the number and sentiment of comments, the title of the video itself, the image of the video, as well as the characteristics of the author or his YouTube channel.*

Keywords: *popularity prediction, Youtube, neural networks, NLP, Computer Vision*

1. УВОД

Од свог настанка 2005. године, *Youtube* је постао синоним за онлине видео садржај. Иако су се појавиле друге платформе попут *Vimeo*, *TikTok*-а и *Instagram Reels*-а, *Youtube* остаје најпопуларнија са 2 милијарде активних корисника месечно, захваљујући огромној количини садржаја, функцијама претраге, персонализованим препорукама и могућности претплате на канале. *Youtube*-ри могу монетизовати садржај од 2007. године, а додатно зарађивати путем спонзорстава, маркетинга и продаје производа, што је многима постало основни извор прихода. Популаран видео има велики број прегледа и интеракција, док трендинг видео нагло добија много прегледа у кратком периоду. У овом раду представљамо студију о предвиђању популарности *Youtube trending* видеа коришћењем техника машинског учења, анализирајући карактеристике видеа и коментара. Коришћени су подаци о трендинг садржају из САД, укључујући сентимент анализу коментара.

НАПОМЕНА:
Овај рад проистекао је из мастер рада чији ментор је био др Александар Ковачевић, ред. проф.

Након инжењеринга својстава и експлоративне анализе, предикција је извршена класификационим моделима и комбинованом архитектуром неуронских мрежа где су анализирани нумерички, текстуални и визуелни подаци.

2. ПРЕДИКЦИЈА ПОПУЛАРНОСТИ

Предикција популарности онлајн садржаја укључује више корака и техника машинског учења. Прво, прикупљају се подаци о видеима, укључујући наслове, описе, тагове, број прегледа, лајкова, коментара и дељења. Ови подаци се затим обрађују и трансформишу у одговарајући формат за машинско учење. Један од важних корака је инжењеринг својстава, где се генеришу нове карактеристике као што су сентимент анализа коментара и анализа кључних речи у насловима и описима.

Затим се модели машинског учења, као што су регресија, класификација или неуронске мреже, тренирају на овим подацима како би научили да предвиђају популарност садржаја. Ови модели могу користити технике као што су *gradient boosting trees*, *random forests* и *support vector machines* за анализу нумеричких података и рекурентне и конволуцијске неуронске мреже за анализу текстуалних и визуелних података. На крају, модели се евалуирају и оптимизују како би пружили што тачније предикције, што омогућава креаторима садржаја и маркетиншким стручњацима да боље разумеју који садржај ће вероватно привући највише пажње и ангажовања публике.

3. КОРИШЋЕНЕ ТЕХНИКЕ И МОДЕЛИ

3.1 SVM

Support Vector Machine (SVM) је напредни алгоритам машинског учења за класификацију и регресију који проналази оптималну хиперравнину која раздваја различите класе података са највећом могућом маргином. SVM је ефикасан у решавању линеарно одвојивих проблема и може се проширити на нелинеарне коришћењем кернел метода. Кернел трик омогућава трансформацију података у веће димензије, олакшавајући њихово раздвајање. SVM-ови су примењени у областима као што су здравство, обрада природног језика и препознавање слике. Алгоритам је дизајниран за бинарну класификацију, али се може прилагодити за проблеме више класа. Једна од главних предности SVM-а је избегавање прекомерне усклађености избором оптималне хиперравнине.

3.2 Стабла одлучивања

Стабла одлучивања су интуитиван алгоритам машинског учења за класификацију и регресију који деле податке на подскупове, креирајући структуру сличну стаблу са чворовима који представљају одлуке на основу одређених карактеристика. Сваки чвор представља атрибут података, док гране представљају исходе тих одлука, водећи до коначних листова који представљају класе или континуиране вредности. Главна предност стабала је једноставност и лакоћа интерпретације, као и могућност рада са нумеричким и категоријалним подацима без много претходне обраде. Међутим, склона су прекомерној усклађености, нарочито када су дубока и комплексна. Овај проблем се често решава обрезивањем грана. Иако су моћна, стабла одлучивања највећу снагу показују у комбинацији са алгоритмима као што су Random Forest и XGBoost.

3.3 Random Forest

Random Forest је алгоритам машинског учења који комбинује више стабала одлучивања како би побољшао тачност и стабилност предикција. Ова техника креира више стабала одлучивања користећи различите подскупове података и узорке карактеристика, а коначна предикција се добија гласањем или просеком предикција свих стабала. *Random Forest* је отпоран на прекомерно усклађивање, јер различита стабла смањују варијанс и повећавају општу тачност модела. Овај алгоритам обрађује велике скупове података са великим бројем карактеристика без значајне претходне обраде. Такође пружа мере важности карактеристика, корисне за разумевање утицаја различитих карактеристика на предикцију.

3.4 XGBoost

XGBoost (*Extreme Gradient Boosting*) је напредни алгоритам машинског учења који користи градијентно појачавање за стварање високоефикасних модела за класификацију и регресију. Овај алгоритам ради у фазама, при чему свака нова фаза покушава да исправи грешке претходних фаза. XGBoost користи ансамбл стабала одлучивања, где сваки нови модел минимизира резидуалне грешке претходних модела. Овај приступ резултира моделима високе прецизности и перформанси. XGBoost је познат по брзини и ефикасности, захваљујући паралелној обради и оптимизацијама у управљању меморијом. Алгоритам пружа регуларизацију, која спречава прекомерну усклађеност и побољшава генерализацију модела.

3.5 Дубоке неуронске мреже

Дубоке неуронске мреже (DNN) су тип неуронских мрежа са више скривених слојева који могу аутоматски учити и представљати високе нивое апстракције у подацима. Оне се састоје од низа слојева повезаних неуронима, где сваки неурон прима улаз, обрађује га и прослеђује следећем слоју. DNN су примењене у компјутерском виду, обради природног језика и препознавању говора. Њихова обука захтева велике скупове података и значајну рачунарску моћ, што је омогућено развојем графичких процесорских јединица (GPU). Иако су веома моћне, дубоке неуронске мреже су склоне проблемима као што су превелика усклађеност и дуго време обуке. За решавање ових проблема често се користе технике као што су регуларизација и dropout.

3.6 Рекурентне неуронске мреже

Рекурентне неуронске мреже (RNN) су дизајниране за рад са секвенцијалним подацима и способне су да задрже информације о претходним стањима кроз унутрашње петље. Ово их чини идеалним за анализу временских серија, обраду природног језика и друге задатке где је контекст битан. RNN-ови памте претходне улазе кроз своју рекурентну архитектуру, омогућавајући коришћење информација из претходних корака за предвиђање будућих. Проблем нестајања или експлодирања градијената током обуке отежава учење дугорочних зависности, што је делимично решено развојем *Long Short-Term Memory* (LSTM) и *Gated Recurrent Unit* (GRU) архитектура. RNN-ови се користе у машинском преводу, препознавању говора и генерисању текста.

3.7 Конволуцијске неуронске мреже

Конволуцијске неуронске мреже (CNN) су оптимизоване за обраду структурираних података попут слика и видео снимака. CNN-ови користе конволуционе слојеве који примењују филтере на улазне податке како би екстраховали релевантне карактеристике као што су ивице, текстуре и обрасци. Састоје се од више конволуционих и пуллинг слојева, који су затим повезани са једним или више потпуно повезаних слојева. Главна предност CNN-ова је њихова способност да аутоматски уче и екстрахују значајке из слика, смањујући потребу за ручним инжењерингом карактеристика.

4. МЕТОДОЛОГИЈА

Методологија решења описаног у овом раду реализована је кроз следеће кораке: прикупљање података, експлоративна анализа података, инжењеринг својстава, претпроцесирање података, сентимент анализа коментара, обучавање и оптимизација класификационих модела, анализа текста рекурентном неуронском мрежом и формирање комплексне неуронске мреже која комбинује све наведене типове улаза за класификацију видео снимка. У наставку овог поглавља детаљније су описане све претходно наведене фазе методологије.

Поред основног скупа података потребно је било и прикупљање, где смо за сваки појединачни видео снимак помоћу Selenium алата преузели 50 најрелевантнијих коментара са видео снимка, слика(*thumbnail*), као и броја претпланика и прегледа које је аутор снимка акумулирао. Коментари нам могу омогућити дубљи залаз у њихов утицај на популарност видео снимка. Ми смо овде анализирали да ли сентимент коментара доприноси прегледима, тј. да ли видео који има добру повратну информацију од гледаоца је инхерентно популарнији због тога.

Информације од канала може се обогатити почетни скуп података и побољшати метрике примењиваних алгоритама машинског учења, што и јесте основна идеја укључивања информација о каналу у овом рад.

Након експлоративне анализе смо закључили да су током година најпопуларније категорија видеа у области музике, забаве, спорта и игрица.Такође смо закључили да је корелација између броја лајкова и броја прегледа доста већа од корелације броја прегледа са бројем коментара и бројем дислајкова. Међутим, ниједна од добијених вредности није занемарљива и има свој допринос у процесу тренирања модела, па ниједна нумеричка колона није избачена.

Инжењерингом својстава користимо постојећи скуп података како би га проширили и добили неко ново знање и значење од њега. Ово се може постићи на мноштво начина. На нашем примеру смо увели податке као што су: временски период који је био потребан да видео постане *trending*, имена категорија спајањем идентификатора са именима како би се добила прегледнија експлоративна анализа као и број негативних, неутралних и позитивних коментара за сваки видео снимак.

Претпроцесирање у овом раду се састоји из више корака који нам омогућују лакши рад са подацима и који дозвољавају да модел има прикладне податке као улазе. Неки од корака су: претварање индетификатора категорије видео снимка у њено име, претварање података кад је видео објављен и кад је ушао у *trending* у податак колико је видео снимку требало да уђе у *trending*, замена *outlier*-а са границама користећи *upper/lower bound* технику, замена недостајућих вредности са вредношћу медијана као и мноштво осталих.

Анализа сентимента прикупљених коментара је урађена уз помоћ две *Python* библиотеке, VADER [1] и *TextBlob* [2]. Резултат семантичке анализе једног конкретног коментара је поларитy и изражен је у опсегу од -1 до 1. Где доња граница -1 представља да је коментар негативан док горња граница 1 представља да је коментар позитиван. Неутралним коментаром сматрамо онај чија апсолутна вредност поларитy не прелази 0.25.

Да би класификација видео снимка била могућа, неопходно је креирање категоричког атрибута који представља дискретне вредности броја прегледа. Ово је постигнуто дискретизацијом континуалне вредности броја прегледа у две групе балансиране по величини. По узору на постојећу литературу коришћени су следећи класификациони модели: *Support Vector* класификатор, *Random Forest* ансамбл класификационих стабала, ансамбл класификационих стабала у *Extreme Gradient Boosting* конфигурацији и вишеслојна мрежа перцептрона.

Како би се предвидела популарност видео снимака, може се применити неуронска мрежа, која је способна да научи сложене обрасце у подацима и да их користи за предвиђање циљне променљиве - популарности видео снимка. Поред нумеричких података, у овом моделу имамо могућност да интегришемо и текстуалне податке као што је наслов видео снимка. Ово ће нам показати колико утицај могу имати ови подаци и колико цлицкбаит култура са насловом и погађање кључних речи описом може да утиче на популарност видео снимка. Мрежа комбинује још један тип података, а то је слика (*thumbnail*) која представља видео визуелно и игра велику улогу у његовој популарности. Желимо да сазнамо да ли квалитет ове слике којој доста аутора дају пажњу и придодају успех видеа има утицај на број прегледа видеа. Мрежа се састоји од три гране која прима одређене улазе, који пролазе кроз неколико слојева, који су специфични по грани, и на крају се конкатенирају како би добили излаз, тј. предикцију модела. Улаз у првој грани је слика, у другој наслов, а у трећој сви остали подаци који су представљени нумериком.

5. ЕКСПЕРИМЕНТИ

Скуп података који се користи је *Youtube Trending Video Datase*t са *Kaggle*-а који се ажурира дневно и може се преузети са [4]. За овај пројекат је задњи пут преузет 28.01.2023 како би имали конзистентан скуп података. Други део скупљања скупа података је прикупљање Youtube коментара које је урађено преко Selenium-а. Неке од значајних колона одабраног скупа су: наслов видео снимка, број лајкова и дислајкова, број коментара, опис, линк ка слици видео снимка и број прегледа. Поред и оригиналних додати су и атрибути канала и анализе сентимента.

Експерименти су се сводили на тестирање тачности модела додавањем и одузимањен одређених атрибута скупа података као што су резултати анализе сентимента и тагови видео снимка. Класификациони модели тестирани су на тест скупу података, издвојеном у почетној фази пројекта (20% од целог скупа података) насумичним поступком. Валидациони скуп није био потребан пошто библиотека са којом се радила унакрсна валидација има уграђено издвајање и евалуирање валидационих података. Извршен је и преглед прошлих експеримената, како би се видела експериментација са подацима и како су они утицали на резултате.

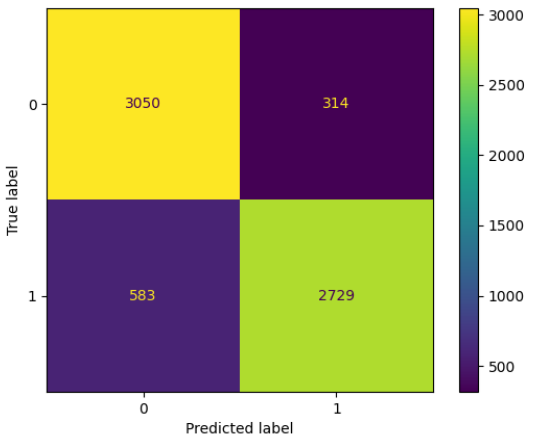
Класификациони модели евалуирани су употребом метрике тачности, AUC, MAE и F-score. Метрике су изабране у складу са проблемом, имајући у обзир да је циљ бинарна класификација, као и да је број позитивних и негативних класа добро избалансиран у датом скупу података. Коришћена је и матрица конфузије, да би се уочило на којим класама модел више греши. Поред матрице конфузије извршена је и додатна анализа грешака како би могли да уочимо зашто наши модели греше и како можемо да побољшамо њихове перформансе. Урађена је и анализа тагови и анализа сентимента утичу на класификационе моделе.

6. РЕЗУЛТАТИ И ДИСКУСИЈА

Табела 1 приказује резултате евалуације свих коришћених модела. Најбоље резултате нумеричке анализе је дао *Random Forest* модел. Видимо како су се, у односу на вишеслојну мрежу перцептрона, мере евалуације са анализом текстуалних елемената смањиле, док се при убацивању визуелних елемента прецизност повећала. Међутим када уместо тренирања наше конволутивне мреже имамо *ResNet50* перформансе се погоршају. Можемо да закључимо да визуелни атрибути имају већи и позитиван утицај у односу на текстуалне са нашим ограниченим ресурсима за тренирање. Такође, видимо да наше конволутивна мрежа има боље перформансе од *ResNet50* која је претходно обучавана. Можемо да претпоставимо да је разлог за то велики корпус података и самим тим слика, где је наша конволутивна мрежа имала довољно ресурса да естрахује смислене атрибуте и да те обрасце примени у разликовању популарних и непопуларних видео снимака.

Табела 1. *Евалуација без тагова са анализом сентимента и оптимизованим хиперпараметрима*

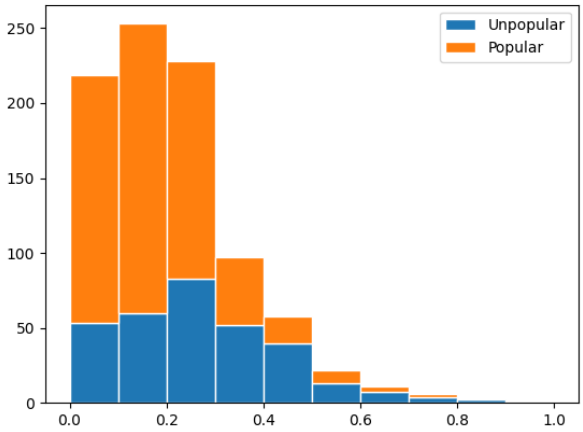
Класификациони модели	Метрике [%]			
	Тачност	AUC	F1	MAE
SVM	86.56	86.53	85.88	13.43
Random Forest	87.60	87.57	87.13	12.40
XGBoost	87.52	87.47	87.01	40.75
MLP	85.77	85.73	84.90	14.23
MLP са LSTM	85.38	85.34	84.46	14.62
MLP са LSTM i CNN	85.82	85.81	85.17	14.21
MLP са LSTM i ResNet50	85.31	85.25	84.03	14.69



Слика 1. *Матрица конфузије* SVM *модела*

Матрице конфузије SVM-а је приказана на слици 1 Матрице модела су у сличним односима као и ова. Видимо да се највише грешака јавља код видео снимака који су популарни а модел их препознаје као непопуларне.

Претпостављамо да се ово може објаснити чињеницом да је теже класификовати популарне снимке који имају мање цифре кад се тиче најважнијих атрибута за класификацију. Ово се може приметити на слици 2 где имамо хистограм лајкова код погрешних предикција, и где видимо да модел греши код малих броја лајкова и то највише код популарних видео снимака.



Слика 2. *Хистограм лајкова код погрешних предикција у односу на популарне и непопуларне видео снимке*

7. ЛИТЕРАТУРА

[1] VADER https://vadersentiment.readthedocs.io/en/latest/

[2] TextBlob. https://textblob.readthedocs.io/en/dev/

[3] YouTube Trending Video Dataset (updated daily): RISHAV SHARMA. Available: https://www.kaggle.com/datasets/rsrishav/YouTube-trending-video-dataset

Кратка биографија:

Душан Лекић рођен је у Краљеву 2000. год. Студент на мастер академским студијама Факултета техничких наука, у оквиру Универзитета у Новом Саду. Након завршене средње школе, уписује 2018. године Факултет техничких наука, смер Рачунарство и аутоматика. У трећој години се усмерава на модул Примењене рачунарске науке и информатика. Након завршених основних студија, уписује 2022. године мастер академске студијена Факултету техничких наука, смер Рачунарство и аутоматика, модул Интелигентни системи.