Univerzitet u Beogradu

Elektrotehnički fakultet



Analiza socijalnih mreža

Izveštaj

|  |  |
| --- | --- |
| Profesor: | Student: |
| prof. dr. Marko Mišić,  prof. dr. Jelica Protić,  as. ms Predrag Obradović | Andrijana Miković 2023/3042 |

Beograd, avgust 2024.

Sadržaj

[Садржај 0](#_Toc156427544)

[1. Увод 1](#_Toc156427545)

[2. Анализа проблема 4](#_Toc156427546)

[2.1. Улазни фајлови 4](#_Toc156427547)

[3. База података 5](#_Toc156427548)

[3.1. Опште информације 5](#_Toc156427549)

[3.2. Најчешће коришћене речи 5](#_Toc156427550)

[4. Трансформација текста 9](#_Toc156427551)

[4.1 . Лематизација 9](#_Toc156427552)

[4.2 . Отклањање *stopwords* 10](#_Toc156427553)

[4.3 . Преглед трансформисаног текста 10](#_Toc156427554)

[5. Модели машинског учења 11](#_Toc156427555)

[5.1. Векторизација 11](#_Toc156427556)

[5.2. Логистичка регресија 11](#_Toc156427557)

[5.3. КНН – к најближих комшија 11](#_Toc156427558)

[5.4. Стабло одлучивања 12](#_Toc156427559)

[5.5. Мултиномијални наивни Бајесов класификатор 13](#_Toc156427560)

[6. Модели неуралних мрежа 14](#_Toc156427561)

[*6.1.* Токенизација и *GloVe* 14](#_Toc156427562)

[*6.2.* *LSTM* 15](#_Toc156427563)

[*6.3.* *BiLSTM* 15](#_Toc156427564)

[*6.4.* *CNN* 16](#_Toc156427565)

[6.5. Креирање слојева код модела неуралних мрежа 16](#_Toc156427566)

[7. Резултати 17](#_Toc156427567)

[7.1. Метрика 17](#_Toc156427568)

[7.1.1. Тачност 17](#_Toc156427569)

[7.1.2. Прецизност 17](#_Toc156427570)

[7.1.3. Одзив 18](#_Toc156427571)

[7.1.4. Ф1-мера 18](#_Toc156427572)

[7.1.5. Матрица конфузије 18](#_Toc156427573)

[7.1.6. Multi-class Log Loss 18](#_Toc156427574)

[7.2. Резултати добијени алгоритмима машинског учења 19](#_Toc156427575)

[7.3. Резултати добијени неуронским мрежама 22](#_Toc156427576)

[8. Закључак 27](#_Toc156427577)

[Литература 28](#_Toc156427578)

[Списак слика 29](#_Toc156427579)

[Списак табела 30](#_Toc156427580)

1. Uvod

Epidemije, kao fenomen neočekivano brzog širenja zaraznih bolesti unutar određenih populacija ili geografskih regiona tokom ograničenog vremenskog perioda, vekovima su prepoznate kao jedan od najvažnijih izazova za zdravlje čovečanstva. Od najranijih civilizacija pa do savremenog doba, epidemije su uzrokovale masovne gubitke ljudskih života, duboko pogađajući društva i ekonomije širom sveta. Istorijski značajne epidemije, poput kuge u srednjem veku, koja je prepolovila populaciju Evrope, ili pandemija kolere u XIX veku, koja se proširila preko kontinenata usled porasta međunarodnih trgovinskih puteva, ostavile su trajne posledice na čovečanstvo.

Globalizacija u XX veku je olakšala međunarodnu razmenu ljudi i dobara. Kao i mogućnost saradnje naučnika iz raznih oblasti, pa i medicine. Ovakva globalizacija je takođe dovela do toga da se zarazne bolesti brže šire. Što je moglo da se primeti i pandemijom koronavirusa SARS-CoV-2 (Severe Acute Respiratory 3 Syndrome Coronavirus 2), slika 1.1, koja je zadesila svet početkom 2020. godine. Ova pandemija je na dramatičan način demonstrirala koliko je savremeno društvo ranjivo na pojavu novih i neočekivanih epidemija, uprkos napretku u medicini i javnom zdravlju. Istovremeno, pandemija je ukazala na ključnu ulogu koju imaju oblasti medicine koje se bave proučavanje epidemiologije i zaraznih bolesti, u borbi protiv pandemija u savremenom svetu.

A close-up of a virus

Description automatically generated

Slika 1.1. SARS-CoV-2

Predmet istraživanja u okviru ovog projektnog zadatka je kvantitativna i kvalitativna analiza naučne produkcije Medicinskog fakulteta Univerziteta u Beogradu, sa posebnim fokusom na oblasti medicinske nauke koje su direktno povezane sa proučavanjem i suzbijanjem epidemija zaraznih bolesti. U analizu će biti uključeni naučni radovi i kolaboracije zaposlenih na Katedri za imunologiju, Katedri za epidemiologiju, Katedri za infektivne bolesti i Katedri za mikrobiologiju. Kroz ovu analizu, biće ispitani obim i kvalitet naučne produkcije, kao i stepen saradnje među zaposlenima na pomenutim katedrama, sa ciljem da se utvrdi kako doprinosi istraživača sa Medicinskog fakulteta Univerziteta u Beogradu doprinose globalnom razumevanju i borbi protiv epidemija..

Cilj ovog rada je da kroz obradu prikupljenih podataka, i analizu podataka iz oblasti medicinskih nauka prikaže stanje naučne produkcije na Fakultetu, mapira mreže saradnje među istraživačima, i identifikuje ključne pojedince i grupe koji doprinose u borbi protiv epidemija. Analiza će obuhvatiti publikacije kao što su knjige, disertacije, naučni radovi u časopisima, kao i radovi predstavljeni na međunarodnim i domaćim naučnim konferencijama, s posebnim akcentom na one publikacije koje su objavljene u naučnim časopisima sa impakt faktorom. Posebna pažnja će biti posvećena razmatranju naučne produkcije sa aspekta njene vidljivosti i uticaja u globalnoj naučnoj zajednici, što uključuje analizu citiranosti radova, mreže kolaboracija.

1. Analiza problema

U ovom poglavlju biće analizirani ulazni fajlovi koji se koriste pri analizi naučne produkcija iz oblasti epidemiologije i zaraznih bolesti na Medicinskom fakultetu u Beogradu. Koji predsavlja problem konstruisanja mreže saradnje na osnovu ulaznih podataka, kao i donošenje zaključka o osobinama i načinu funkcionisanja ljudi, kao pojedinca, i kao predstavnika svoje katedre. Ovi fajlovi su formirani na osnovu upita u bazu *Scopus.*

* 1. Ulazni fajlovi

Podaci prikupljeni za analiz, dostupni su u vidu odgovarajućih fajlova. Primarni skup podataka je predstavljen u okviru fajla *autori.xlsx,* i on sadrži informacije o autorima koji su se vodili kao zaposleni u nastavnim zvanjima na katedrama za imunologiju, epidemiologiju, infektivne bolesti i mikrobiologiju, Medicinskog fakultea Univerziteta u Beogradu, u oktobru 2020. godine, što je prikupljeno sa internet prezentacija ovih katedra. Autori naučnih publikacija koji se javljaju u okviru skupa podataka, a ne nalaze se u ovom primarnom skupu, nisu od interesa za analizu.

Fajl *autori.xlsx* sadrži sledeće informacije:

* *Ime*
* *Prezime*
* *Katedra* – na kojoj je posmatrani naučnik angažovan u oktobru 2020.
* *H indeks* – kao mera produktivnosti i citiranosti naučnika
* *Broj radova*

Ostali podaci iz primarnog skupa podataka, predstavljaju podatke o radovima dobijene iz indeksne baze načnih radova *Scopus,*  zaključno sa oktobrom 2020. Radovi su podeljeni u četri dadoteke gde svaka odgovara jednoj od prethudno pomenute katedre, *epidemiologija.xlsx, imunologija.xlsx, infektivne\_bolesti.xlsx, mikrobiologija.xlsx.*

Podaci se sastoje od sledećih kolona:

1. *Author* – ime zaposlenog kome je pridružena odgovarajuća publikacija u Scopus bazi,

* *Authors* – svi autori rada,
* *Title* – naslov publikacije,
* *Year* – godina izdavanja,
* *Source title* – ime časopisa u kome je rad objavljen,
* *Volume* – volumen časopisa u kome je rad objavljen,
* *Issue* – redni broj sveske časopisa u kome je rad objavljen,
* *Art. No.* – identifikacioni broj rada,
* *Page start* – početna stranica rada u časopisu,
* *Page end* – poslednja stranica rada u časopisu,
* *Page count* – ukupan broj stranica rada,
* *Cited by* – broj citata drugih autora prema Scopus bazi,
* *Link* – link ka izvoru,
* *Document Type* – tip rada,
* *Source* – indeksna baza koja predstavlja izvor podataka (Scopus).

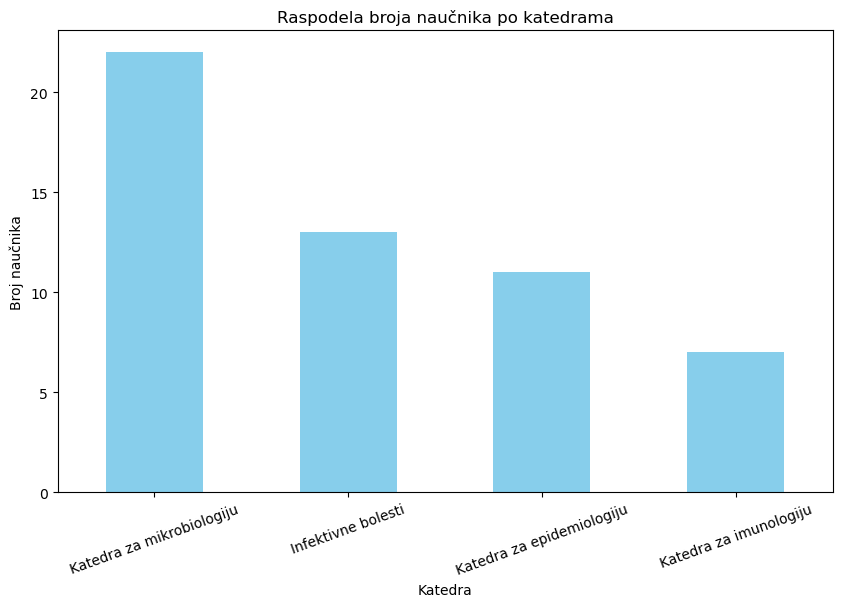
U analizi su od interesa samo publikacije tipa Article, Article in Press, Review, Book Chapter, Letter, Note, dok se ostale mogu zanemariti

1. Baza podataka

U ovom poglavlju biće dat opis baze podataka korišćene u ovom istraživanju.

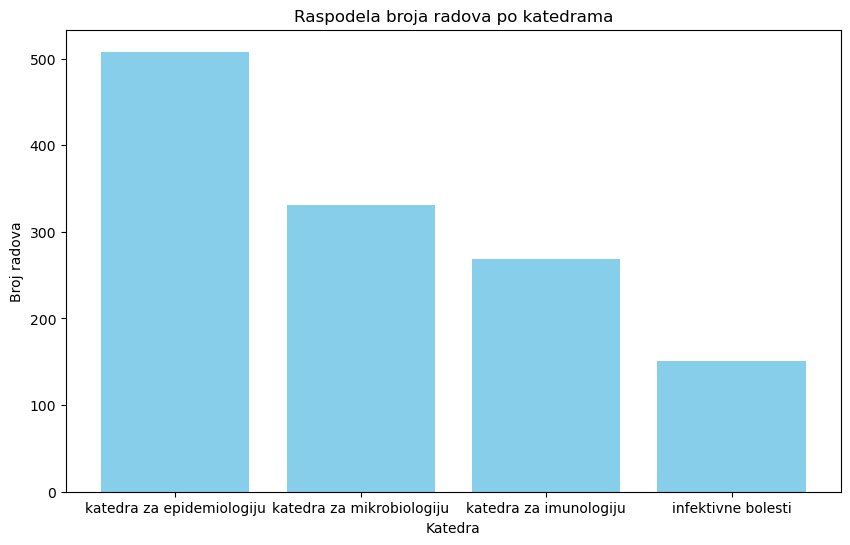
* 1. Opšte informacije

Fajl koji sadrži informacije o naučnicima, od interesa inicijalno sadrži informacije o 59 naučnika, analizom sadržaja fajla primećujemo da od toga postoji 6 naučnika koji ne sadrže informaciju u H indeksu i broju radova. Predmet ovog istraživanja je analiza naučne produkcije, samim tim nam zaposleni za koje nemamo informacije o radovima u posmatranom periodu nisu od interesa, i otklanjamo ih iz inicijalnog skupa podataka. Nakon čega je analizirana raspodela posmatranih naučnika po katedrama, koja je prikazana na slici 3.1.1. Sa koje primećujemo da aktivni naučnici koji se bave proučavanjem epidemiologije i zaraznih bolesti nisu ravnomerno raspoređeni po katedrama od interesa.



Slika 3.1.1. Raspodela broja naučnika po katedrama

Vršimo analizu radova prikupljenih iz *Scopus* baze, primećujemo da postoje duplikati. Konačan skup podataka je dobijen otklanjanjem među duplikatima, onih vrednosti koji sadrže manje informacija, fokusirajući se na broj citiranosti. Nakon uklanjanja podataka navedenim principom u bazi su ostala još dva para duplikata. Kod jednog od parova duplikata imamo dva puta identične informacije, tako da je nasumično vršeno uklanjanje jednog zapisa, dok je za drugi par izvršena provera dostupnosti izvora u *Scopus* bazi, i otklonjen je onaj duplikat čiji izvorni link ne vodi na traženi rad. Na slici

Slika 3.1.2. Raspodela broja radova po katedrama

Možemo primetiti da iako se na katedri za epidemiologiju nalazi manje autora u odnosu na katedru za mikrobiologiju (slika 3.1.1.), oni su učestvovali u pisanju većeg broja radova, čije vrednosti možemo videti u tabeli 3.1.1.

Tabela 3.1.1. Raspodela broja radova po katedrama

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Infektivne bolesti | Katedra za epidemiologiju | Katedra za imunologiju | Katedra za mikrobiologiju |
| Broj radova | 151 | 508 | 269 | 331 |

1. Statistička obrada podataka

Izvršena je statistička obrada podataka, sa fokusom na istraživačka pitanja data u postavci projekta [1].

4.1 . Analiza produktivnosti autora

Analiziran je očišćeni skup radova, u tabeli 4.1.1. vidimo 10 istraživača sa najvećim brojem radova na osnovu analiziranih podataka.

Tabela 4.1.1. Prikaz autora sa najviše radova celovito brojanje

|  |  |  |
| --- | --- | --- |
| Autor | Katedra | Broj radova |
| Tatjana Pekmezović | Katedra za epidemiologiju | 262 |
| Vladimir Trajković | Katedra za imunologiju | 155 |
| Vera Pravica | Katedra za imunologiju | 87 |
| Ljiljana Marković-Denić | Katedra za epidemiologiju | 84 |
| Tatjana Gazibar | Katedra za epidemiologiju | 77 |
| Maja Stanojević | Katedra za mikrobiologiju | 70 |
| Darija Kisić-Tepavčević | Katedra za epidemiologiju | 62 |
| Valentina Arsić-Arsenijević | Katedra za mikrobiologiju | 49 |
| Ivana Ćirković | Katedra za mikrobiologiju | 46 |
| Dušan Popadić | Katedra za imunologiju | 46 |

Uviđamo da Tatjana Pekmezović prednjači sa brojem radova, nakon nje se nalazi Vladimir Trajković sa 107 radova manje u odnosu na svoju koleginicom. Tatjana pripada katedri za epidemiologiju, koja kao što smo videli u analizi u prethodnom poglavlju broji najviše radova, njihova 4 autora se nalaze u prvih 10 po broju radova, kao i 3 u prvih 5 mesta. Dok zanimljiva je činjenica da je druga katedra po broju radova iz prethodne analize katedra za mikrobiologiju, iako se njihovi autori ne nalaze u prvih 5. Prethodna tabela je dobijena celovitim brojanjem radova po autoru, dok ako uzmemo u obzir i broj koautora sa Medicinskog fakulteta koji su se bavili istim radom, dobijemo tabelu 4.1.2.

Tabela 4.1.2. Prikaz autora sa najviše radova frakcijalno brojanje

|  |  |  |
| --- | --- | --- |
| Autor | Katedra | Broj radova |
| Tatjana Pekmezović | Katedra za epidemiologiju | 222.96 |
| Vladimir Trajković | Katedra za imunologiju | 135.87 |
| Vera Pravica | Katedra za imunologiju | 73.08 |
| Ljiljana Marković-Denić | Katedra za epidemiologiju | 52.98 |
| Maja Stanojević | Katedra za mikrobiologiju | 46.46 |
| Tatjana Gazibar | Katedra za epidemiologiju | 40.50 |
| Valentina Arsić-Arsenijević | Katedra za mikrobiologiju | 34.08 |
| Jovan Ranin | Infektivne bolesti | 33.92 |
| Sandra Spasić-Grujičić | Katedra za epidemiologiju | 33.42 |
| Ivana Ćirković | Katedra za mikrobiologiju | 29.04 |

Gde možemo primetiti drugačiju raspodelu autora, gde su prve 4 pozicije ostale ne promenjene, kao i da je u prvih 10 isplivala i katedra za Infektivne bolest, gde se autor Jovan Ranin pri celovitom brojanju radova nalazio na 13. mestu. Na osnovu čega znamo da on ima više radova koji su pisani u saradnji sa manje kolega, u odnosu na osobe koje su se nalazile iznad njega primenom celovitog brojanja radova.

4.2 . Analiza broja koautora po svakom autoru

Analizom polja *Authors* u kome se nalaze informacije o kautorima na posmatranom radu, odvojenih zarezom. Utvrđen je prosečan broj koautora po svakom autoru koji nam je od interesa i čije ime i prezime se nalazi u *Author* koloni, treba napomenuti da nije vršena analiza stringa koji predstavlja koautore, već je uzeta pretpostavka o ispravnosti, da je svaki autor odvojen zarezom i da se posmatrani autor sa Medicinskog fakulteta univerziteta u Beogradu ne nalazi u toj listi. U tabeli 4.2.1. se nalazi 5 autora sa najvećim brojem koautora u proseku na svojim radovima.

Tabela 4.2.2. Prikaz autora sa najviše koautora u proseku na svojim radovima

|  |  |  |
| --- | --- | --- |
| Autor | Katedra | Broj koautora |
| Elenora Dubljanin | Katedra za mikrobiologiju | 390.74 |
| Isidora Vujičić | Katedra za epidemiologiju | 343.81 |
| Valentina Arsić-Arsenijević | Katedra za mikrobiologiju | 71.82 |
| Maja Stanojević | Katedra za mikrobiologiju | 32.34 |
| Vladimir Trajković | Katedra za imunologiju | 32.00 |

Možemo primetiti da se prva dva autora iz tabele 4.2.2. ne nalaze u tabelama koje prikazuju broj autora sa najvećim brojem radova (4.1.1. i 4.1.2.), dok se ostali nalaze u sva tri posmatrana rezultata. Što znači da ova dva autora imaju manji broj radova, a oni su pisani u saradnji sa velikim brojem koautora. Sa slike 4.2.1. možemo primetiti da najveći broj autora ima manje od 10 koautora na svojim radovima, za koje možemo pretpostaviti da su oni više doprineli tim radovima.

A graph with blue squares

Description automatically generatedSlika 4.2.1. Raspodela broja prosečnog broja koautora po autoru

4.3 . H-indeks

H-indeks je način da se u jednom podatku objedini informacija o broju publikacija autora i broju citata. Naučnik poseduje j-indeks od h ukoliko ima najmanje h radova, od kojih je svaki barem h puta citiran. [2]. U primarnom skupu podataka koji sadrži informaciju o autorima od interesa, imamo informaciju o njihovom H-indeksu, za te podatke znamo samo da su dobijeni sa prezentacija odgovarajućih katedri, vršeno je upoređivanje vrednosti tog H-indeksa sa izračunatim H-indeksom na osnovu prethodne definicije podacima o radovima iz *Scopus* baze. U tabeli 4.3.1. su prikazani dobijeni i izračunati H-indeks samo za autore kod kojih postoji razlika između te dve vrednosti.

Tabela 4.3.1. Prikaz autora za koje postoji razlika između izračunatog i dobijenog H-indeksa

|  |  |  |
| --- | --- | --- |
| Autor | H-indeks dobijen | H-indeks izračunat |
| Vera Pravica | 35 | 33 |
| Nataša Vučković-Opavski | 1 | 0 |
| Aleksandar Džamić | 6 | 5 |
| Vladimir Trajković | 37 | 36 |
| Emina Milošević | 8 | 7 |

Primećujemo da je izračunati indeks manji od dobijenog, razlog odstupanja može biti nepoklapanje baza radova na osnovu kojih je računat H-indeks za autore, u ovom istraživanju je korišćena *Scopus* baza, dok nemamo informaciju na osnovu koje baze je dobijen izloženi H-indeks na sajtu fakulteta.

4.4 . Produktivnost katedri

Posmatramo produktivnost katedri na osnovu produkcije i citiranosti u časopisima.

1. Модели машинског учења

У овом поглављу биће представљена четири алгоритма машинског учења коришћена за класификацију датог текста. Након трансформације текста извршена је подела на одлике (енгл. *features*) и мете (енгл. *targets*). Одлике представља колоне *cleaned\_text,* док мете података представља колона аутор. Затим је извршена подела на податке за тренирање (80%) и податке за тестирање (20%), помоћу *train\_test\_split* методе из *sckit-learn* библиотеке.

* 1. Векторизација

Подаци за тренирање који представљају одлике се налазе у текстуалном формату, који није погодан за алгоритме класификације. Потребно га је претворити у нумерички облик. Ова трансформација је извршена помоћу *TfidfVectorizer (Term Frequency-Inverse Document Frequeny Vectorizer)* класе из *scikit-learn* библиотеке. Наведена класа има методу *fit\_transform* која речима из улазног текста додељује одговарајуће тежине, чиме се приказује колику важност нека реч има у датом текстуалном документу.

* 1. Логистичка регресија

Логистичка регресија представља математички модел којим се описује веза између атрибута и категоричке зависне променљиве. Она враћа вероватноћу да дати скуп атрибута, са специфичним вредностима, припада свакој од категорија. У оквиру овог истраживања, примењена је бинарна логистичка регресија, јер је потребно утврдити да ли особа пати од дијабетес или не. Овај модел се ослања на употребу логистичке функције како би повезао улазне атрибуте и пружио вероватноћу припадности одређеној категорији.

* 1. КНН – к најближих комшија

Алгоритам к најближих суседа скр. КNN (енгл. *k-nearest neighbors*) представља

непараметарски метод, надгледног учења. Циљ КНН класификације је додељивање нових узорака одређеним класама на основу сличности са већ познатим тренинг скупом података. Користи метрику удаљености како би мерио сличности између узорака у простору атрибута. Ова метрика се користи за проналажење *к* најближих суседа за сваки нови узорак. Параметар *к* представља број најближих суседа који ће бити узети у обзир приликом класификације новог узорка. Промену прецизности алгоритма у зависности од вредности параметра можемо видети на слици 5.3.1.

Slika na kojoj se nalazi tekst, snimak ekrana, Plot, linija

Opis je automatski generisan

Слика 5.3.1. Прецизност КНН алгоритма за различите вредности параметра

* 1. Стабло одлучивања

Стабло одлучивања представља алгоритам машинског учења који се базира на концепту доношења одлука у складу са условима и акцијама. Карактерише га јасна графичка структура, у облику стабла. У стаблу одлучивања, корен представља почетни услов који одређује избор параметра на основу кога се врши раздвајање података. Гране стабла представљају различите услове који се примењују на податке, а листови садрже финалне одлуке, тј. излазне резултате алгоритма. Процес раздвајања на гране захтева меру информационог добитка. Информациони добитак је колико раздвајање на одређени начин помаже у смањењу неодлучности. За сваки могући услов врши се рачунање информационог добитка и одабира се најбољи пут за раздвајање, тј. стабло се изграђује тако да минимизује неодлучност и максимизује информациони добитак.

* 1. Мултиномијални наивни Бајесов класификатор

Представља један од алгоритама за класификацију у области машинског учења, који се заснива на Бајесовој теореми [2]. Термин наивни потиче од претпоставке да су сви атрибути независни, што често није случај. Мултиномијални се користи када су подаци представљени као бројност (фреквенција), као што је често случај у анализи текста.

1. Модели неуралних мрежа

У овом поглављу биће представљена три модела неуралних мрежа коришћена за класификацију датог текста. Као и њихови резултати. Након трансформације текста извршена је подела на одлике (енгл. *features*) и мете (енгл. *targets*). Одлике представља колоне *cleaned\_text,* док мете података представља колона аутор. Затим је извршена подела на податке за тренирање (80%) и податке за тестирање (20%), помоћу *train\_test\_split* методе из *sckit-learn* библиотеке.

* 1. Токенизација и *GloVe*

Пре самог коришћена модела неуралних мрежа, потребно је извршити додатне припреме текста. Прво је извршена токенизација коришћењем класе *Tokenizer* из *keras* библиотеке, на овај начин су одређене јединствене речи које постоје у текстовима, и њима је додељен одговарајући индекс. Ови индекси су корисни за представљање текста као низа бројева. За даљу обраду нам је потребно да сви вектори добијени токенизацијом имају једнаку дужину. Извршено је додавање нула, (енгл. *Padding)*, код вектора са мањом дужином од потребне.

Јединствене речи, добијене токенизацијом, преставићемо матрицом која садржи информације о семантичким везама између речи, базираној на статистици заједничког понављања у великим текстовима. То је урађено користећи већ креирани фајл *GloVe*, који има 100 димензија. *GloVe (Global Vectors for Word Representation)*, је алгоритам не надзорног учења за добијање векторске репрезентације речи. У овом семинарском учитан је фајл са резултатима већ истренираног алгоритма, на великој бази података. Врши се проналазак сваког од токена из базе података, у овом фајлу и ако он постоји мења се одговарајућом матрицом, у супротном замењује се матрицом попуњеном нулама, примена овог алгоритма види се на слици 6.1.1.

Slika na kojoj se nalazi tekst, snimak ekrana, Font, softver

Opis je automatski generisan

Слика 6.1.1. Учитавање и примена *GloVe* фајла

За разлику од неких других метода, где се фокусирају на локални контекст, нпр. *Word2Vec skip-gram, GloVe* узима у обзор глобални контекст речи и њихова заједничка појављивања. На овај начин ухвати шире семантичке везе, и речи које имају слична значења или употребе такође имају сличне матричне репрезентације.

* 1. *LSTM*

*LSTM (Long Short-Term Memory)* је рекурентна неуронска мрежа (РНН), направљена да превазиђе ограничења традиционалне РНН мреже у чувању дуготрајних зависности секвенци података и настајању градијента. У стању су да обрађују секвенце података, као што су временске серије, текст и говор. Користе меморијске ћелије и капије за контролу протока информација, на овај начин селективно задрже или одбаце информације по потреби и тако избегавају проблем настајања градијента.[3]

* 1. *BiLSTM*

*BiLSTM (Bidirectiona Long Short-Term Memory)* је унапређење традиционалнe *LSTM* архитектуре. Кључна разлика је у томе како се информације обрађују унутар мреже. Код стандардне *LSTM* се уналзни низ обрађује у једном смеру, и чува зависности само од претходних информација, док код *BiLSTM* се то врши у оба смера, и чува се зависност и од претходних и будућих информација.

* 1. *CNN*

*CNN (Convolution Neural Network)* конволуциона неуронска мрежа, која кроз своје слојеве примењује конволуционе операције на улазне податке. Конволуција представља прелаз филтером преко улазних података, да би се извукле карактеристике. Ови филтери науче да детектују потребне карактеристике за решавање одређеног проблема. Најчешћа примена је у обради слика и снимака, али овакав тип мреже налази своју примену и у обради природног језика. Дизајнирана је да обрађује структурине матричне податке, па као улаз може да прими матрице које представљају одређене токене извучене из текста, добијене нашом обрадом.

* 1. Креирање слојева код модела неуралних мрежа

Креирање слојева неуралних мрежа извршено је коришћењем *Sequential* класе из *keras* библиотеке. Код које је могуће једноставно додавање слојева користећи *add* методу.

Први слој код сваког модела, јесте *Embedding layer* из *keras* библиотеке. Примарна улога овог слоја је да мапира бројевну репрезентацију речи у вектор густине фиксне величине. Као улазне параметре прихвата величину речника тј. број јединствених речи, затим величину вектора помоћу кога је представљена свака реч, тежине одређене помоћу *GloVe* фајла, и на крају величину најдуже речи. Додавање овог слоја можемо видети на слици 6.5.1.

Slika na kojoj se nalazi tekst, snimak ekrana, Font

Opis je automatski generisan

Слика 6.5.1. Додавање *Embedding layer*

Други слој представља изабрани модел и служи за процесирање речи добијених од претходног слоја. За креирање сваког од слоја коришћена је одговарајућа класа из *keras* библиотеке. Након тога су додата два скривена слоја, праћена са *dropout layer* за регулацију, и на крају излазни слој за класификацију у једну од три класе. *Dropout* спречава *overfitting* током тренирања, одбацивањем дела података који су дошли од претходног слоја.

1. Резултати

У овом поглављу биће представљени кључни резултати семинарског. Анализираћемо перформансе различитих модела класификације на описаном скупу података у поглављу 3, као и перформансе неуронских мрежа. Приказати њихове резултате коришћењем одговарајућих метрика.

* 1. Метрика

За евалуацију резултата класификације коришћене су различите метрике:

* Тачност
* Прецизност
* Одзив
* Ф1-мера
* Матрица конфузије
* *Multi-class Log Loss*

Свака метрика пружа јединствени увид у перформансе класификационих модела, омогућавајући дубље разумевање њихових способности и недостатака.

* + 1. Тачност

Тачност даје информацију о генералној успешности предикције модел, кроз све класе. Рачуна се по следећој формули:

(6.1.2.)

где представља број тачних предикција, а , укупан број предикција.

* + 1. Прецизност

Прецизност је метрика која процењује тачност позитивних предвиђања модела. Даје информацију о томе колико је модел склон тачном класификовању позитивних примера. Рачуна се по следећој формули:

(7.1.2.)

где *Tp*  представља број правих позитивних вредности, тј. узорака које је модел класификације означио као позитивне, и они то заиста и јесу, а *Fp* представља број лажно позитивних, модел је негативне вредности погрешно класификовао.

* + 1. Одзив

Одзив је метрика која процењује способност модела да тачно идентификује све стварно позитивне примере. Пружа информацију о томе колико је модел ефикасан у откривању позитивних примера. Рачуна се по следећој формули:

(7.1.3)

где *Tp*  представља број правих позитивних вредности и *Fn* представља број лажно негативних вредности.

* + 1. Ф1-мера

Ф1-мера комбинује прецизност и одзив модела у једну вредност, и представља њихову хармоничну средину. Користи се када је потребно узети у обзир равнотежу између тачности позитивних предвиђања и способности модела да идентификује све позитивне примере. Рачуна се по формули 6.1.4. где *P* представља вредностпрецизности, а *O* одзива.

(7.1.4.)

* + 1. Матрица конфузије

Табела која сумира број тачно позитивних, тачно негативних, лажно позитивних и лажно негативних предикција за сваку класу. Даје детаљан пресек перформанси модела за сваку класу.

* + 1. Multi-class Log Loss

Процењује перформансе класификације модела, чији излаз представља вероватноћу између 0 и 1. У поставци проблема за такмичење, захтевана је евалуација решења коришћењем ове метрике. Добијене по следећој формули:

(7.1.6.)

где је *N* број посматрања у тест подскупу, *M* број класа, у овом случају 3,  *y*  да ли тренутни улаз одговара датој класи или не, узима вредност 0 или 1, и  *p*  представља вероватноћу да дати улаз одговара посматраној класи. Коришћена је за процену перформанси само неуронских мрежа, зато што оне директно враћају вероватноћу. Могуће је прилагодити и алгоритме машинског учења тако да враћају вероватноћу припадања одређеној класи, али ова вредност није значајна у процесу класификације датих алгоритама.

* 1. Резултати добијени алгоритмима машинског учења

Резултате класификације алгоритама машинског учења можемо видети у табели 7.2.1.

Најбоље се показао Мултиномијални наивни Бајесов класификатор, по свим параметрима, одмах после њега је Логистичка регресија. Док стабло одлучивања има доста слабе перформансе.

Табела 7.2.1. Резултати класификације алгоритама машинског учења

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Алгоритми | Тачност [%] | Прецизност [%] | Одзив [%] | Ф1-мера [%] |
| КНН | 68.9 | 74.45 | 68.897 | 67.588 |
| Логистичка регресија | 80.4 | 80.568 | 80.388 | 80.352 |
| Стабло одлучивања | 58.5 | 58.412 | 58.495 | 58.421 |
| МНБ | 81.3 | 82.203 | 81.324 | 81.246 |

У матрицама конфузије датим у наставку, лабеле су кодиране на следећи начин:

* 0 = Едгар Алан По
* 1 = Хаурд Филипс Лафкрафт
* 2 = Мери Шели

На слици 7.2.1. налази се матрица конфузије КНН алгоритма, можемо приметити да он има велике вредности дуж дијагонале, што се поклапа са перформансама алгоритма. Такође видимо да лажно детектује 0 класу, што значи да преписује По-у дела која нису његова, до овога долази због благог дизбаланса класа, где постоји више текстова овог аутора у односу на друга два.

Slika na kojoj se nalazi tekst, snimak ekrana, dijagram, linija

Opis je automatski generisan

Слика 7.2.1. Матрица конфузије КНН алгоритма

На слици 7.2.3. налази се матрица конфузије логистичке регресије, на њој примећујемо доста мање вредности лажних позитива у односу на КНН алгоритам.

Slika na kojoj se nalazi tekst, snimak ekrana, dijagram, kvadrat

Opis je automatski generisan

Слика 7.2.2. Матрица конфузије Логистичке регресије

Матрица конфузије стабла одлучивања приказана је на слици 7.2.3., на њој видимо знатно више вредности лажних позитива, као и ниже вредности на дијагонали. Овакви резултати одговарају перформансама алгоритма из табеле, који се показао као најлошији избор за решавање датог проблема.

Slika na kojoj se nalazi tekst, snimak ekrana, dijagram, kvadrat

Opis je automatski generisan

Слика 7.2.3. Матрица конфузије стабла одлучивања

На слици 7.2.4. приказана је матрица конфузије мултиномијалног наивни Бајесов класификатор, који се показао најбоље у класификације. Али има јако сличне перформансе са логистичком регресијом, што се види и са упоређивањем њихових матрица конфузије.

Slika na kojoj se nalazi tekst, snimak ekrana, dijagram, kvadrat

Opis je automatski generisan

Слика 7.2.4. Матрица конфузије МултиномијалногНБ

* 1. Резултати добијени неуронским мрежама

Резултате класификације неуронским мрежама можемо видети у табели 7.3.1.

Најбоље се показао Мултиномијални наивни Бајесов класификатор, по свим параметрима, одмах после њега је Логистичка регресија. Док стабло одлучивања има доста слабе перформансе.

Табела 7.3.1. Резултати класификације неуронским мрежама

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Алгоритми | Тачност [%] | Прецизност [%] | Одзив [%] | Ф1-мера [%] |
| *LSTM* | 74.41 | 74.46 | 74.41 | 74.38 |
| *BiLSTM* | 74.95 | 74.95 | 74.95 | 74.94 |
| *CNN* | 69.87 | 70.11 | 69.87 | 69.76 |

На слици 7.3.1. налази се матрица конфузије *LSTM* модела, примећујемо високе вредности дуж дијагонале, али не занемарив број лажних позитива.

Slika na kojoj se nalazi tekst, snimak ekrana, kvadrat, dijagram

Opis je automatski generisan

Слика 7.3.1. Матрица конфузије *LSTM*

На слици 7.3.2. приказани су графици *Log-Loss* функције, на тренинг и валидационом сету, можемо приметити да обе функције опадају што представља индикацију да не постоји *overfitting.*

Slika na kojoj se nalazi tekst, linija, snimak ekrana, Plot

Opis je automatski generisan

Слика 7.3.2. *Log-Loss* фунција *LSTM*

На слици 7.3.3. налази се матрица конфузије *BiLSTM* модела, који има сличне вредности као и *LSTM* модел, тако да бидирекционални приступ не доводи до успешније класификације.

Slika na kojoj se nalazi tekst, snimak ekrana, kvadrat, dijagram

Opis je automatski generisan

Слика 7.3.3. Матрица конфузије *BiLSM*

На слици 7.3.4. видимо *Log-Loss* функције за овај модел, које су опадајуће. Такође примећујемо више вредности валидационе Log-Loss функције што указује на веће грешке и већи проблем *overfitting-a* у односу на *LSM* модел.

Slika na kojoj se nalazi tekst, linija, Plot, snimak ekrana

Opis je automatski generisan

Слика 7.3.4. *Log-Loss* функција *BiLSTM*

Матрица конфузије *CNN* мреже приказана је на слици 7.3.5., на њој видимо ниже вредности на дијагонали за не доминанту класу у односу на претходна два алгоритма.

Slika na kojoj se nalazi tekst, snimak ekrana, kvadrat, dijagram

Opis je automatski generisan

Слика 7.3.5. Матрица конфузије *CNN* мреже

Слика 7.3.3. Матрица конфузије *BiLSM*

На слици 7.3.6. видимо *Log-Loss* функције за овај модел, које су опадајуће. Такође примећујемо више вредности валидационе Log-Loss функције што указује на веће грешке и већи проблем *overfitting-a,* али чињеница да ова функција има вредност мању од 0.5 указује на успешност алгоритма.

Slika na kojoj se nalazi tekst, Plot, dijagram, linija

Opis je automatski generisan

Слика 7.3.6. *LogLoss* фунција *CNN*

1. Закључак

На основу резултата приказаних у претходном поглављу, примећујемо да најбоље перформансе има мултиномијалног наивни Бајесов класификатор код кога све метрике су у распону 81.2% - 82.2%, након њега следи Логистичка регресија чије метрике достижу 80%. Након ова два алгоритма машинског учења, следе неуралне мреже, *LSTM I BiLSTM* дају јако сличне резултате, мало испод 75% али када погледамо график log-loss функције код *BiLSTM* модела примећујемо већи *overfitting.* Након њих следе *CNN* модел и КНН алгоритам, неурална мрежа има нижу прецизност, док се по осталим параметрима показала за нијансу боља од КНН алгоритма. Стабло одлучивања има најлошије резултате класификације.

Литература

1. Analiza socijalnih mreža, Projektni zadatak
2. H-indeks, https://nauka.cpn.rs/sadrzaj/vrednovanje-bibliografskih-podataka/

(20.08.2024.)

1. Introduction to Long Short-Term Memory(LSTM), <https://www.simplilearn.com/tutorials/artificial-intelligence-tutorial/lstm> (17.01.2024.)

Списак слика

[Слика 1.1. Едгар Алан По 1](#_Toc156427506)

[Слика 1.2. Хаурд Филипс Лафкрафт 2](#_Toc156427507)

[Слика 1.3. Мери Шели 2](#_Toc156427508)

[Слика 1.3. Мери Шели 5](#_Toc156427509)

[Слика 3.2.1. 50 Најчешће коришћених речи на целој бази података 6](#_Toc156427510)

[Слика 3.2.2. 50 Најчешће коришћених речи Едгар Алан По 6](#_Toc156427511)

[Слика 3.2.3. 50 Најчешће коришћених речи Хаурд Филипс Лафкрафт 7](#_Toc156427512)

[Слика 3.2.4. 50 Најчешће коришћених речи Мери Шели 8](#_Toc156427513)

[Слика 4.1.1. Процес лематизације 9](#_Toc156427514)

[Слика 4.2.1. Процес отклањања *stopwords* 10](#_Toc156427515)

[Слика 4.3.1. Излаз након трансформације текста 10](#_Toc156427516)

[Слика 5.3.1. Прецизност КНН алгоритма за различите вредности параметра 12](#_Toc156427517)

[Слика 6.1.1. Учитавање и примена *GloVe* фајла 15](#_Toc156427518)

[Слика 6.5.1. Додавање *Embedding layer* 16](#_Toc156427519)

[Слика 7.2.1. Матрица конфузије КНН алгоритма 20](#_Toc156427520)

[Слика 7.2.2. Матрица конфузије Логистичке регресије 20](#_Toc156427521)

[Слика 7.2.3. Матрица конфузије стабла одлучивања 21](#_Toc156427522)

[Слика 7.2.4. Матрица конфузије МултиномијалногНБ 22](#_Toc156427523)

[Слика 7.3.1. Матрица конфузије *LSTM* 23](#_Toc156427524)

[Слика 7.3.2. *Log-Loss* фунција *LSTM* 23](#_Toc156427525)

[Слика 7.3.3. Матрица конфузије *BiLSM* 24](#_Toc156427526)

[Слика 7.3.4. *Log-Loss* функција *BiLSTM* 25](#_Toc156427527)

[Слика 7.3.5. Матрица конфузије *CNN* мреже 25](#_Toc156427528)

[Слика 7.3.3. Матрица конфузије *BiLSM* 26](#_Toc156427529)

[Слика 7.3.6. *LogLoss* фунција *CNN* 26](#_Toc156427530)

Списак табела

[Табела 7.2.1. Резултати класификације алгоритама машинског учења 19](#_Toc156427531)

[Табела 7.3.1. Резултати класификације неуронским мрежама 22](#_Toc156427532)