Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 007.5
DOI: https://doi.org/10.24867/27BE12Zdelar

SENTIMENT ANALIZA OBJAVA SA DRUŠTVENE MREŽE *TWITTER* O COVID-19 VAKCINAMA

SENTIMENT ANALYSIS OF TWITTER POSTS ON COVID-19 VACCINES

Filip Zdelar, *Fakultet tehničkih nauka, Novi Sad*

Oblast – RAČUNARSTVO I AUTOMATIKA

Kratak sadržaj – *U ovom radu predstavljen je razvoj sistema za analizu Tviter podataka o vakcinama protiv COVID-19 virusa u Srbiji. Analiza uključuje sistem za detekciju sentimenta i detekciju teme kojom se tvit bavi. Za potrebe izrade projekta, preuzet je označen skup engleskih tvitova i prikupljen skup srpskih tvitova koju su prevedeni na engleski jezik. Za detekciju sentimenta isprobana su dva pristupa obučavanje modela nadgledanog učenja na označenom skupu podataka engleskih tvitova o COVID-19 virusu. Na zadatku detekcije sentimenta najbolje performanse ostvarene su kovolucione neuronske mreže metodom koja je postizala tačnost 59% i identifikovano da postoji ukupno je 38% pozitivnih, 37% negativnih i 25% neutralnih srpskih tvitova. Model tema ostvario je koherentnost od 0.45 i identifikovano je petnaest tema.*

Ključne reči: *sentiment, topic , twitter, COVID-19, NLP, neuronske mreže*

Abstract *– This paper presents the development of a system for analyzing Twitter data about COVID-19 vaccines in Serbia. The analysis includes a sentiment detection system and a topic detection system related to the tweets. For the project's development, a labeled set of English tweets was used, along with a collection of Serbian tweets that were translated into English. Two approaches were tested for sentiment detection: supervised learning by training a model on a labeled dataset of English tweets about COVID-19. The best performance in sentiment detection was achieved using convolutional neural networks, which achieved an accuracy of 59%. It was identified that there are 38% positive, 37% negative, and 25% neutral Serbian tweets in the dataset. The topic model achieved a coherence score of 0.45, identifying fifteen topics.*

Keywords*: sentiment, topic, Twitter, COVID-19, NLP, neural networks*

1. UVOD

Planeta se suočila sa velikom pandemijom korona virusa krajem 2019. godine. Izbijanje bolesti korona virusa (COVID-19) u velikoj meri je uticalo na ljudski život. Zdravlje ljudi ugroženo je ne samo zbog vanredne situacije, već i zbog naknadnih društvenih ishoda kao što su nezaposlenost, nedostatak resursa i finansijska kriza. U ovakvim okolnostima kao sredstvo za borbu javljaju se vakcine i od sve veće važnosti postaje politika javnog zdravlja. Svetska zdravstvena organizacija (SZO) navodi da je ravnopravan pristup sigurnim i efikasnim vakcinama ključan za okončavanje pandemije COVID-19 i neumorno radi sa partnerima na razvoju, proizvodnji i primeni bezbednih i efikasnih vakcina [1]. Do 13. aprila 2022. godine na listi SZO za hitnu upotrebu je odobreno 10 vakcina protiv COVID-19, dok je ukupno 37 vakcina koje su licencirane i odobrene za upotrebu u hitnim slučajevima u bar jednoj državi, a 195 vakcina je u procesu razvoja i odobravanja [2].

Prvi slučaj korona virusa u Srbiji potvrđen je 6. marta 2020. godine, nakon čega zemlju zahvata epidemija i uvode se mere za suzbijanje sirenja virusa. Država Srbija 19. januara 2021. godine predstavlja sistem za primenu, zakazivanje i praćenje vakcina i imunizacije stanovništva. Upućen je poziv građanima da se vakcinišu u što većem broju i započet je proces masovne vakcinacije [3]. U naporima za postizanjem imunizacije stanovništva, nadležnim organima postaje sve neophodniji uvid u stavove šire javnosti povodom vakcinacije. Uspešno sprovođenje imunizacije stanovništva podrazumeva pozitivan stav građana prema vakcinama, pa jedan od zadataka učesnika u javnom zdravlju predstavlja odgovor na stavove ljudi koji se protive imunizaciji.

Za sistem javnog zdravlja može biti od značaja informacija o razlozima negativnog stava prema vakcinama, kako bi se kroz aktivnosti sistema javnog zdravlja moglo uticati na formiranje pozitivnog mišljenja javnosti. Dobra okolnost čini to što društvene mreže predstavljaju uobičajeno mesto za ljude da izraze svoje emocije i stavove, pa je velika dostupnost informacija o stavovima javnosti o značajnim temama, kao što je i milionski broj javno dostupnih objava na temu vakcina na društvenoj mreži kao što je Twitter. Analiza sentimenta može pružiti pravovremeni uvid u stav i mišljenje javnosti prema COVID-19 vakcinama i obezbediti smernice za kreiranje politike javnog zdravlja i dizajniranje prilagođenih programa edukacije o vakcinama.

Uz uvid u teme koje se javljaju u diskusijama koje sadrže negativan ili pozitivan stav prema vakcinaciji, moguće je poboljšati interakciju nadležnih organa sa širom javnošću i delovati u pravcu uspešne imunizacije stanovništva. U ovom radu će biti predstavljeno jedno rešenje za analizu podataka sa Twittera o COVID-19 vakcinama. Rešenje omogućava razumevanje stavova javnosti povodom vakcinacije i identifikaciju glavnih tema u objavama negativnog i pozitivnog sentimenta. Detaljniji opis

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

podataka, izazova i rešenja izložen je u ostatku rada. Naredno poglavlje se bavi srodnim istraživanjima na ovu temu. U trećem poglavlju će biti opisan skup podataka, način pripreme podataka za obučavanje i validaciju modela. Nakon toga, biće predstavljena metodologija koja je korišćena za rešavanje problema detekcije sentimenta i tema. Potom sledi prikaz rezultata i diskusija. Na samom kraju biće izvedeni zaključci samog rada.

2. METODOLOGIJA

U ovom poglavlju je predstavljeno pretprocesiranje tvitova, implementacija sistema za analizu sentimenata i analizu tema. Moduli za sentiment analizu su opisani u dva podpoglavlja, jedno koje je zasnovano na nadgledanom učenju i drugo koje je zasnovano na nenadgledanom učenju.

Modeli za klasifikaciju imaju na ulazu *TF-IDF* i *BERT* vektore tvitova, a izlaz klasifikatora su tri moguće klase; tvit je za vakcinaciju, tvit je neutralan povodom vakcinacije ili ne može da se odredi, i tvit je protiv vakcinacije. Analizom tema bi se dinamički odredile teme koje postoje u skupu podataka i izvršilo bi se prebrojavanje broja tvitova koje pripadaju jednoj temi.

2.1. Pretprocesiranje

Pre upotrebe podataka urađenu su sledeće transformacije, kako bi se skup podataka mogao koristiti u modelima i kako bi se odstranili uzorci koji nisu pogodni za obučavanje. U slučaju supa podataka za treniranje, kategorička obeležija *very negative* i *very positive* su prebačena kao *postive* i kao *negative*, što je odrađeno radi smanjenja kompleksnosti modela. Svim tvitovima su izbačeni simboli *hashtag*-a što je simbol @, dok je sama njegova vrednost ostavljena u tvtiu. Izbačeni su svi linkovi koji sadrže https ili http prefiks. Emodži ikonice su prvo razdvojene razmakom kako bi se zatim pretvorile u svoje semantičko značenje korišćenjem biblioteke *emoji*. Izbačeni su svi znaci interpunkcije kao što su ‘‘.’’, ‘‘,’’, ‘‘?’’, ‘‘!’’ i druge slične znakove. Nakon toga, izbačeni su specijalni znakovi koji ne pripadaju latiničnim slovima. Na samom kraju svi karakteri su prebačeni u mala slova latinice kako se reči sa istim semantičkim značenjem ne bi razlikovale u modelu.

Posle obrade podataka koji je prvobitno bio u tekstualnom obliku, izvršava se tokenizacija teksta kako bi se reči unutar teksta normalizovale. Između procesa uzimanja osnove leksikografske forme (lemitizacije) i oduzimanje sufiksa (stematizcaije), odbrana je lemtizacije zbog toga što je poznatija po davanju boljih rezultata, uprkos tome što je sporija. Prilikom ovog proces odbačene su zaustavne reči kao što su ‘‘*the*’’, ‘‘*is*’’, ‘‘*as*’’, ‘‘*to*’’, ‘‘*a*’’, ‘‘*that*’’ i mnoge druge slične reči koje se često pojavljuju, a pri tome ne daju semantičko značenje.

Na kraju je za svaku reč određena njena vrsta reči i ostavljene su samo one koje su pridevi, imenice, glagoli i prilozi. Urađene su obe vrste ranije spomenutih tokenizacija; *TFIDF* i *BERT*. Za dužinu TFIDF vektora dobijena je vrednost 126575. BERT reprezentacije podataka dobijene su korišćenjem *DestilBert* prethodno obučenog modela, dok su podaci pripremljeni za model korišćenjem DistilBertTokenizer.

Preuzet je prvi vektor poslednjeg skrivenog stanja, koji predstavlja reprezentaciju CLS tokena dužine 73.

2.2. Skup podataka

Na Preuzet je označeni skup engleskih tvitova i prikupljen skup srpskih tvitova koji su prevedeni na engleski jezik. Označeni skup engleskih tvitova korišćen je za treniranje i optimizaciju modela nadgledanog učenja.

Stratifikovanom podelom 0.7 podataka je izdvojeno za obučavanje, a 0.3 za validaciju i optimizaciju modela. Nasumičan deo srpskih tvitova ručno je označen i predstavlja test skup podataka. Srpski tvitovi su odabrani nasumično kroz vreme zato što je primećeno da su određeni događaji uticali da se javi velika količina tvitova sa istim sentimentom u tom vremenskom periodu, npr. vest o otvaranju fabrika za proizvodnju vakcina povukla je veliki broj pozitivnih tvitova u tom vremenskom periodu. Na ručno označenim srpskim tvitovima, nakon što su prevedeni na engleski, biće testirani optimizovani modeli, kao i nenadgledani pristupi za određivanje sentimenta. Podaci na kojima se vrši obuka modela nadgledanog učenja su preuzeti u CSV formatu sa *openICPRS* repozitorijuma javno dostupnih podataka o društvenim, bihevioralnim i zdravstvenim istraživanjima.

Set podataka “*COVID-19 Twitter Dataset with Latent Topics, Sentiments and Emotions Attributes*” formiran je sa ciljem istraživanja javnih konverzacija na Tviteru na temu pandemije COVID-19 virusa [4].

Set podataka sadrži preko 198 miliona Tviter objava na engleskom jeziku prikupljenih u periodu od 28. januara 2020. do 1. septembra 2021. koristeći ključne reči “corona”, “wuhan”, “nCov” i “covid”. Podaci su označeni korišćenjem CrystalFeel [5], kolekcije pet unapred obučenih algoritama mašinskog učenja za ekstrakciju sentimenta i ocenu emocije, koji su trenirani na manuelno označenim tvitovima. Za potrebe projekta, preuzeti su CSV fajlovi sa tvitovima sa teritorija sa engleskim kao govornim jezikom, Australije i Velike Britanije. Najpre je prikupljeno 149186 australijskih tvitova (87819 negativnih, 37033 pozitivnih, 24334 neutralnih), a zatim je set podataka proširen tvitovima iz Velike Britanije kojih ima 477410 tvitova (148443 negativnih, 163884 pozitivnih, 165083 neutralnih).

Konačan balansiran skup podataka cini 420000 tvitova, pri čemu nije uključen celokupan skup za obučavanje zbog ograničenih resursa. Svaki tvit u setu podataka označen je sa sedamnaest semantičkih atributa, ali pošto je obim ovog projekta ograničen na detekciju sentimenta i ne uključuje detekciju konkretnih emocija, za potrebe obučavanja modela korišćen je samo kvalitativni atribut koji ukazuje na kategoriju sentimenta (*very negative, negative, neutral or mixed, positive, very positive*). Takođe, model je obučavan da detektuje sentiment samo na teksta objave, odnosno nije obučavan na atributima koji sadrže informaciju o korisniku, njegovoj lokaciji, vremenu postavljanja tvita i interakcijama (broj retvitova, lajkova). Preuzeti skup podataka sadrži jedinstven ID tvita na osnovu kog je korišćenjem Tviter API-ja dobavljen tekst tvita. Modeli su trenirani da prepoznaju kategoriju, ali ne i intenzitet sentimenta, pa su tvitovi označeni sa *very positive* i *very negative* prilikom pripreme podataka

za treniranje modela označeni kao *positive* i *negative*, respektivno.

Cilj projekta jeste analiza sentimenta srpskih tvitova, pa priprema podataka uključuje prikupljanje Tviter objava o vakcinama na srpskom jeziku. Tvitovi su prikupljeni korišćenjem Tviter API-ja koristeći *academic research* pristup, koji omogućava pristup istorijski javnim podacima uz dodatne funkcionalnosti koje podržavaju prikupljanje preciznijih i potpunijih skupova podataka [6]. Prikupljeni su tvitovi na srpskom jeziku postavljeni u periodu od 31. januara 2021. do 31. januara 2022. godine, koristeći kombinaciju svih morfoloških oblika ključnih reči vakcina, vakcinisati i kovid, na ćirilici i latinici.

Zbog ograničenja dužine upita koji se prosleđuje Tviter API-ju inicijalni upit koji sadrži kombinacije svih morfoloških oblika ključnih reči izdeljen je na četiri kraća upita. Konačan skup podataka dobijen je spajanjem prikupljenih podataka različitih upita, pri čemu je vođeno računa da se jedna objava prikupi samo jedanput. Izostavljeni su tvitovi koji *repost*-uju originalnu objavu (*retweets*, *quotes*), kao i tvitovi koji predstavljaju *reply*, nakon sto je analizom utvrđeno da su ovi tvitovi često kratkog teksta i da je njihov sentiment zavistan od konteksta samog tvita na koji odgovaraju. Prikupljeno je 93435 tvitova, koji su zatim prevedeni sa srpskog na engleski jezik korišćenjem *GOOGLETRANSLATE* koji je dostupan u *google sheets* aplikaciji u kojoj su podaci čuvani. Nasumičnim izborom izdvojeno je 6000 tvitova koji su manuelno anotirani sistemom trostrukog glasanja – svakom tvitu su dve osobe nezavisno davale ocenu sentimenta i u slučaju neslaganja je treća osoba dodeljivala konačnu ocenu sentimenta.

2.3. Sentiment analiza i analiza tema

U ovom poglavlju su grupisani modeli svih klasifikacionih sistema za dobijanje sentimenta. Ulaz u sistem su engleski tvitovi koji su korišteni kao skup podataka za obučavanje. Skup podataka koji je zasnovan na srpskim tvitovima se koristi za validaciju modela istreniranih na obučavajućem engleskom skupu podataka. Sakupljen skup podataka se prevodi na engleski jezik gde se dalje pretprocesira na isti način na koje se pretprocesiraju tvitovi koji su sačinjeni od engleskog skupa podataka. Izvršava se tokenizacija korišćenjem *TF-IDF* i *BERT* modela. Svaki od ovih modela su istrenirani na modelima mašinskog učenja: naivni Bajes, metod nasumične šume, višeslojne neuronske mreže i mašine na bazi vektora nosača. Kod nenadgledane sentiment analize, izlaz je isti kao kod nadgledane analize, sa tim da kod nenadgledane analize nema skupa podataka koji je sačinjen od engleskih tvitova. Takođe na ulaz su dovedeni preobučeni modeli *TextBlob, Vader* i model leksičkih osobina. Ulaz za analizu tema čini skup podataka sačinjen od srpskih tvitova, dok je izlaz sačinjen od tema, koje će biti dobijene metodom lakta i izlaz će biti obučeni klasifikacioni model. Sama analiza tema će biti urađena sa dva moguća pristupa: *LDA* i *NMF*.

2.4. Rezultati

Konvolucione neuronske mreže su se pokazale kao tačniji i najprecizniji model za dobijanje sentimenta tvita vrednosti tačnosti od 52.3% sa TF-ITF teoknizatorom i 50,7% sa BERT tokenizatorom. Vrednosti tačnosti modela naivnog Bajesa, nasumične šume, višeslojne neuronske mreže i konvolucione neuronske mreže su prikazani u tabeli 1, gde su sve vrednosti kolona vrste tokenizatora. U tabli 2 su zadate vrednosti tačnosti i preciznosti TextBlob-a Vader-a i leksičke analize osobina. Za nenedgledano obučavanje, najbolje se pokazao model leksičke analize osobina (Afinn) koji je ostvario tačnost od 39,4% i preciznost od 44%. U tabeli 3 dati su rezultati metrika konvolucione neuronske mreže obučene sa BERT tokenima. Na osnovu visoke preciznosti za klasu neutralnih sentimenata, može da se zaključi da model generalno predviđa više predikcija za neutralnu klasu. Najveći odziv je ostvarila negativna klasa sa vrednosti od 65.7%, dok je za vrednost F-mere klasa neutralnih sentimenta ostvarila najbolju vrednost koja je 61,7%

Tabela 1. *Tačnost za svaki nadgledani klasifikacioni model upotrebom TFIDF i LDA tokenizatora*

	Tačnost modela upotrebom TF-IDF	Tačnost modela upotrebom BERT
Mašine na bazi vektora nosača	0.346	0.346
Naivni Bajes	0.333	0.333
Nasumične šume	0.447	0.485
Višeslojne neuronske mreže	0.448	0.587
Konvolucione neuronske mreže	0.523	0.597

Tabela 2. *Vrednost tačnosti i preciznosti za modele TextBlob, Vader i leksičke analize*

	Tačnost	Preciznost
TextBlob	0.376	0.380
Vader	0.371	0.374
Afinn	0.395	0.440

Tabela 3. *Vrednost preciznosti, odziva i F-mere za svaki sentiment dobijen upotrebom modela konvolucione neuronske mreže*

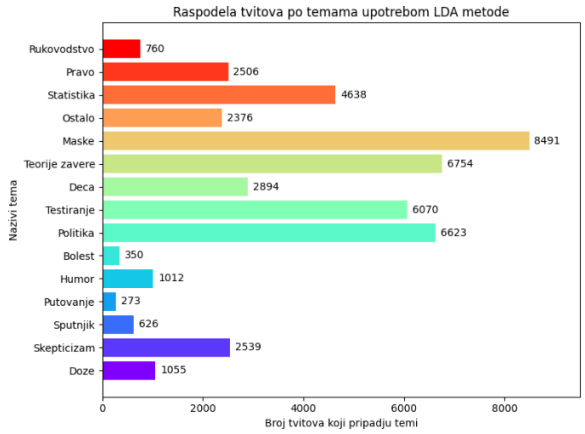
	Preciznost	Odziv	F-mera
Negativni sentiment	0.553	0.657	0.601
Neutralni sentiment	0.692	0.558	0.617
Pozitivni sentiment	0.544	0.607	0.574
Prosek	0.596	0.607	0.597

U radu [7] dobijeni su rezultati sa većom tačnosti i sa većom preciznosti. Razlog je u tome što su radili

detaljnije filtriranje relevantnih tvitova i što su kombinovali pristupe tokenizacije. Primenom modela dobijeni rezultati ukazuju da je *38%* pozitivnih, *37%* negativnih i *25%* neutralnih srpskih tvitova.

Za dobijanje tema, odabiran je model *LDA* nad *TF-IDF* tvitovima, koji je imao najveći skor koherencije u odnosu na druge modele.

Odabran je model koji ima petnaest tema. Dobijene su reči svake teme modela i zatim su imenovane od strane male grupe ljudi. Teme su: doze, politika, putovanje, sputnjik, humor, bolest, skepticizam, testiranje, deca, teorije zavere, maske, ostalo, statistika, pravo i rukovodstvo.



Slika 1. *Raspodela tvitova po temama upotrebom LDA metode nad celokupnim skupom podataka srpskih tvitova.*

3. ZAKLJUČAK

Ovaj rad se bavio problemom analize Tviter podataka vezanih na temu COVID-19 vakcinaciju. Analiza je bila usmerena na dobijanje sentimenta pojedinačnih tvitova kao i grupisanje tvitova po leksičkom značenju i pridodavanje značenja grupama. Rezultati ovakve analize bi mogli biti iskorišćeni u zdravstvu kao mera mišljenja javnog mnjenja. Uz dodatnu analizu nadležni bi mogli imati dublji uvid u formiranje i praćenje stavova ljudi na društvenoj mreži Tviter. Za ovu analizu prikupljeni su tvitovi koji su sadržali ključne reči vezane za vakcinaciju u periodu od 31. Januara 2021. Go 21. Januara 2022. godine. Isprobana su dva pristupa mašinskog učenja za dobijanje sentimenta.

Jedan pristup je vezan za nadgledano, gde su se obučavali modeli na skupu engleskih tvitova i zatim primenjivali na srpske tvitove, dok drugi pristup vezan nenadgledano učenje uz upotrebu semantičke analize pojedinačnih tvitova.

Za dobijanje sentimenata najbolje rezultate ostvario je konvoluciona neuronska mreža sa tačnosti od 0,597 i sa preciznosti od 0.596. Za dobijanje tema, korišćena su dva modela, NMF i LDA. Krajnja analiza tema je izvršena sa modelom LDA koja je ostvarila koherentnost modela od 0,44. Tvitovi su podeljeni u 15 tema i svaka tema je imenovana zasebno.

Dobijene grupe su: doze, politika, putovanje, humor, bolest, testiranje, skepticizam, deca, maske, statistika, pravo, rukovodstvo, teorija zavere, sputnjik i ostalo. Potencijalni pravci za dalji rad na ovu temu bi bili: proširenje skupa podataka, filtriranje skupa podataka, predprocesiranje, detekcija sarkazma i odabir modela.

4. LITERATURA

[1] https://www.who.int/emergencies/diseases/novel-coronavirus-2019/covid-19-vaccines (pogledano 3. 10. 2023.)

[2] https://covid19.trackvaccines.org/agency/who/ (pogledano 15. 4. 2022.)

[3] https://www.srbija.gov.rs/vest/en/166398/mass-vaccination-in-serbia-starts-today.php (pogledano 15. 4. 2022.)

[4] Raj Kumar Gupta, Ajay Vishwanath i Yinping Yang. "*COVID-19 Twitter Dataset with Latent Topics, Sentiments and Emotions Attributes*". CoRR abs/2007.06954 (2020.)

[5] CrystalFeel. Multidimensional Emotion Intensity Analysis from Natural Language. Institute of High Performance Computing, A*STAR. URL: https://socialanalyticsplus.net/crystalfeel (pogledano 30. 8. 2022.)

[6] https:// developer.twitter.com/en/products/twitter-api/academic-research (pogledano 15. 4. 2022.).

[7] Alaa Khudhair Abbas i dr. "*Twitter sentiment analysis using an ensemble majority vote classifier*". en. Xi'nan Jiaotong Daxue xuebao

Kratka biografija:

Filip Zdelar rođen je u Sremskoj Mitrovici 1998. god. Filip Zdelar rođen je 23. novembra 1998. godine u Sremskoj Mitrovici, Republika Srbija. Osnovnu školu pohađao je u selu Šašinci, dok je srednju školu završio u Novom Sadu, gde je bio član odeljenja za darovite matematičare. Tokom školovanja isticao se na takmičenjima iz matematike, fizike i programiranja, plasirajući se na republičkom nivou. Godine 2017. upisao je Fakultet tehničkih nauka na Univerzitetu u Novom Sadu, sa smerom informacionog inženjeringa. Filip je takođe dobitnik Dositejeve nagrade. Na fakultetu je postigao izuzetne rezultate i diplomirao je sa prosekom ocena od 9,67. Master rad na Fakultetu tehničkih nauka iz oblasti Elektrotehnike i računarstva Inteligentni sistemi odbranio je 2023.god.
kontakt: filipzdelar@gmail.com