Univerzitet u Nišu
Elektronski fakultet

Seminarski rad

Analiza sentimenata Twitter postova korišćenjem open-source biblioteka u okviru Python programskog jezika

Veb majning, 2020/2021.

Mentor:
Doc. dr Miloš Bogdanović

Student:
Vladimir Janjić, 1283

Sadržaj

Uvod

Krajem 90ih i početkom 2000ih, rasprostanjenost i dostupnost Interneta značajno raste povećanjem brzine protoka. Veliki broj ljudi dobija pristup Internetu čime se znatno uvećava količina korisničkog materijala u vidu raznih postova, blogova, komentara, recenzija. Eksponencijalni rast količine podataka zadaje velike probleme analitičarima podataka koji nisu u mogućnosti da ostanu u korak sa ubrzanim rastom, gde se količina dostupnih podataka uvećava za 20% svake godine. [1] Stoga početkom 2000ih godina kreće sa razvojem disciplina pod nazivom analiza sentimenata.

Analiza sentimenata predstavlja skup zadataka čija su osnova subjektivne informacije dobijene obradom prirodnog jezika. Drugi naziv za disciplinu jeste „rudarenje mišljenja“ (eng. *Opinion Mining*), međutim originalni naziv je ispravniji zbog razlike u značenju reči sentiment (osećanje) i mišljenje. Sentiment uključuje upotrebu osećanja pri izražavanju stavova, mišljenja ili suda (npr. „Voleo bih da danas bude sunčan dan“), dok mišljenje uključuje rezonovanje zasnovano na činjenicama (npr. „Danas će možda biti sunčan dan“).

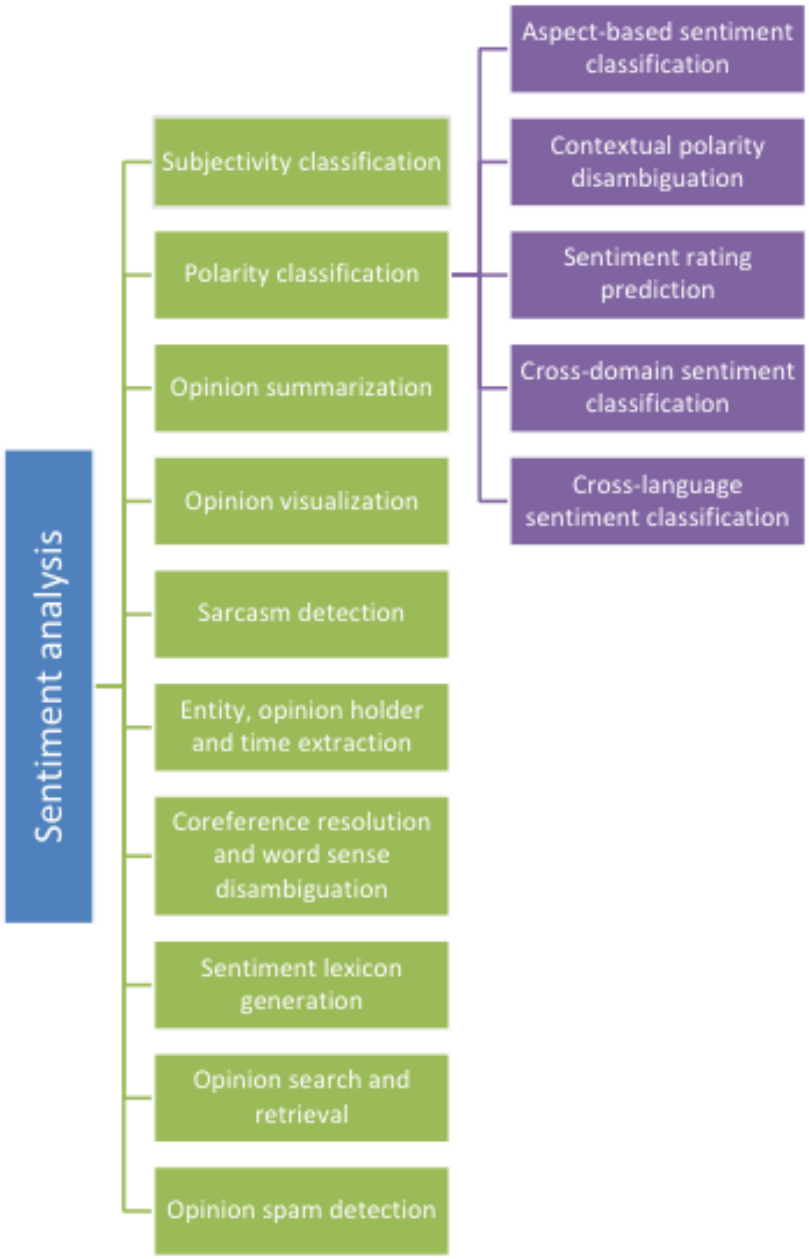
Disciplina dobija na značaju pojavom i rastom popularnosti socijalnih mreža. Veliki broj korisnika dobija mogućnost izražavanja mišljenja, što predstavlja dobru podlogu za analizu i izvlačenje znanja. Primer jedne od najpopularnijih socijalnih mreža jeste Twitter. U tekućoj godini (2022.) broj tvitova svakog sekunda prešao je broj od 10,000 što na dnevnom nivou iznosi 867 miliona. [2] Nažalost, veliki broj tvitova donosi i brojne probleme. Najveći problem odnosi se na tip jezika koji se koristi pri pisanju tvitova. Korišćeni jezik je veoma nestruktuiran sa nedefinisanim pravilima koji se razlikuje od korisnika do korisnika, dok se razlike konstantno uvećavaju. Zbog toga se uspešna analiza sentimenata socijalnih mreža zasniva na kreiranju robusnog sistema koji će prevazići razlike u jeziku i uspešno izvući znanje iz subjektivnih postova.

Razvojem discipline, uvećao se i broj dostupnih alata. Zbog složenosti discipline, alat koji uspešno obavlja posao na zadovoljavajućem nivou prilično je skup i složene izrade. Samo mali broj alata dostupan je širem broju korisnika, uz veoma ograničene mogućnosti. Međutim, naprednijim korisnicima su dostupni brojni alati otvorenog koga, čijim se kombinovanjem ostvaruje potrebna funkcionalnost. Najširi ekosistem ovih alata dostupan je u okviru programskog jezika Python.

U okviru rada biće predstavljeni osnovni koncepti analize sentimenata kroz karakteristike, upotrebu, pogodnosi, ali i probleme, kao i tipove alata. Praktični deo rada rasniva se na upotrebi biblioteka u okviru Python programskog jezika za analizu skupa Twitter postova, koristeći razne tehnike pri obradi i analizi podataka.

Analiza sentimenata i zadaci discipline

Cilj analize sentimenata je definisanje automatskih alata koji će moći da izvuku subjektivne informacije iz tekstova u prirodnom jeziku (kao što su mišljenja i osećanja) i stvoriti strukturirano i upotrebljivo znanje koje se može koristiti pri donošenju odluka. [3] Predstavlja jednu od najznačajnijih disciplina u obradi prirodnog jezika, kao i jednu od najsloženijih. Glavni zadatak analize sentimenata jeste klasifikacija polariteta. Osim toga, obuhvata i niz podzadataka predstavljenih na slici 1.



Slika 1. Zadaci analize sentimenata

Prema slici 1, izvršena je podela na sledece zadatke:

- **Klasifikacija subjektivnosti** – odnosi se na klasifikovanje teksta kao subjektivnog ili objektivnog. Objektivni tekst ne sadrži sentiment i samim tim nije podložan daljoj analizi. Subjektivni tekst sadrži tekst zasnovan na emocijama i može se dalje obrađivati.
- **Klasifikacija polariteta** – glavni zadatak u okviru kog se subjektivni tekst klasifikuje na pozitivni, negativni ili neutralni. Detaljnija analiza obuhvata dodeljivanje brojčane vrednosti sentimentu (npr. realne vrednosti od -1 do 1, celobrojne vrednosti ili procenti 0 do 100, broj zvezdica od 1 do 5 ili -2 do 2, veoma negativno do veoma pozitivno...) u zavisnosti od jačine. Analiza se vrši na nivou poruke, rečenice ili pojedinih aspekata. U okviru klasifikacije polariteta mogu se izvojiti:
    - **Klasifikacija sentimenata na osnovu aspekata** – najniži nivo klasifikacije u kome učestvuje entitet i aspekt vezan za isti (npr. „Video igre su zabavne“, „video igre“: entitet, „su zabavne“: aspekt). Kompanije koriste ovaj nivo klasifikacije jer imaju uvid u probleme vezane za određeni deo proizvoda (npr. „Loša grafika“, „Dobar gameplay“, „Odlična priča“, itd.).
    - **Razjašnjenje kontekstualnog polariteta** – zasniva se na upotrebi dodatnog znanja za preciznije klasifikovanje sentimenta. Rečenica „Bilo je jako strašno i jezivo!“ predstavlja negativni sentiment u svakodnevnoj situaciji, dok u kontekstu horor video igara predstavlja pozitivno iskustvo.
    - **Predviđanje ocene sentimenta** – zasniva se na upotrebi znanja dobijenih iz prethodnih klasifikacija za klasifikovanje novih tekstova. Time se mogu kreirati pravila za klasifikovanje ili trenirati veštačke neuronske mreže. Još jedna od upotreba jeste korigovanje ili dopuna ocena ukoliko se recenzija značajno razlikuje od dodeljene ocene (npr. hvaljenje proizvoda u okviru teksta i dodeljivanje ocene 1/5 zvezdica)
    - **Klasifikacija sentimenata različitih domena** – zasniva se na uspešnoj klasifikaciji tekstova koji dolaze iz različitih izvora i imaju različite strukture, kao što su recenzije proizvoda i postovi sa socijalnih mreža. Problem se zasniva u razlici u govoru, nivou formalnosti, značenju reči itd.
    - **Klasifikacija sentimenata različitih jezika** – najveću upotrebu ima kod klasifikacije polariteta sa postova socijalnih mreža gde se reči slobodno kombinuju iz više jezika. Problem se zasniva u pravilnom prepoznavanju reči stranog jezika koje mogu izgledati kao niz nasumičnih simbola ili greške u pisanju. Dodatni problem predstavlja veći rečnik koji se upotrebljava.
- **Sumiranje mišljenja** – predstavlja izvlačenje konteksta i najznačajnijih delova iz teksta. Time se otklanja šum iz teksta i omogućuje lakše izvlačenje uvida.

- **Vizuelizacija mišljenja** – vizuelno predstavljanje sentimenata u raznim oblicima: stubičasti grafova po broju pozitivnih, negativnih i neutralnih tekstova; mape reči po najkorišćenijim rečima iz većeg broja tekstova; vizuelizacija uticaja određenih reči na sentiment teksta...
- **Detekcija sarkazma** – uz detekciju ironije (koje se u okviru teksta na Internetu koristi pod istim imenom kao i sarkazam) predstavlja jedan od najvećih izazova klasifikacije teksta, kako za ljude, tako i za mašine. Uspešno otkrivanje sarkazma značajno utiče na uspešnost klasifikacije polariteta.
- **Ekstrakcija entiteta, nosioca mišljenja i vreme** – bitne stavke koje se odnose na pravilno donošenje odluka. Ekstrakcijom entiteta dobija se uvid o domenu sentimenta, nosioc mišljenja je bitan u okviru grupa ljudi (npr. rasa, socialni status, starost...) za preciznu i kontrolisanu akciju (npr. marketing prema određenoj grupi), dok je vreme objavljivanja teksta bitno zbog veoma brze promene trendova na Internetu.
- **Razrešavanje koreferenci i razjašnjenje značenja reči** – odnosi se na precizno definisanje entiteta na koji se sentiment odnosi (npr. „Moja omiljena igra je Dark Souls. Prešao sam je više puta.“ Druga rečenica sadrži „je“ koje se odnosi na video igru) i razjašnjenje reči koje mogu biti dvosmislene (npr. „kosa“ može značiti alat ili deo tela).
- **Generisanje leksikona sentimenata** – odnosi se na kreiranje pravila za dodeljivanje sentimenata, kao i definisanje sentimenata pojedinih reči zasnovano na prethodnoj obradi. Postoji više različitih leksikona koja sadrže pravila nastala ljudskim radom (sopstveno subjektivno mišljenje ljudi) ili pravila nastala mašinskim učenjem.
- **Traženje i pronalaženje mišljenja** – zasniva se na pronalaženju i izdvajanju subjektivnih rečenica i reči u okviru teksta koji može sadržati i objektivne činjenice.
- **Otkrivanje spam mišljenja** – zasniva se na otkrivanju mišljenja koji se ponavljaju u velikoj meri, a koja sadrže isto ili slično mišljenje. Primer korišćenja jeste detekcija „review bombing“-a, gde korisnici masovno pišu recenzije vezane za istu stavku uz niže ocene.

Vredi napomenuti korišćenje analize sentimenata pri kreiranju „troll“ filtera (ljudi koji sa namerom pišu kontroverzne izjave ili one koje se kose sa opšteprihvaćenim sa ciljem izazivanja burne reakcije drugih korisnika), detekciji marketinga na socijalnim mrežama (reklama ne mora da se poklapa sa stvarnim stavom korisnika), detekciji pokretača mišljenja (otkrivanje centra mišljenja i način širenja), detekciji maltretiranja i zlostavljanja, kao i u svrhe predlaganja materijala koji će biti zanimljiv korisniku zasnovan na prethodnom iskustvu.

Kao napredna upotreba analize sentimenata izdvaja se detekcija emocija. Predstavlja nivo iznad klasifikacije polariteta. Zadatak je izdvojiti konkretnu emociju koju nosi tekst (npr. sreća, tuga...). Čak ni najbolji sistemi nisu u mogućnosti da na zadovoljavajuć način ispune ovaj zadatak zbog razlika u načinu izražavanja emocija ljudi, ali i zbog različitih definicija i shvatanja emocija. Dodatan problem predstavlja potreba da ljudi prethodno obeleže tekstove da bi bio upotrebljiv pri mašinskom učenju.

Upotreba analize sentimenata

Iako i dalje u razvoju, analiza sentimenata našla je primenu u raznim oblastima Internet poslovanja. Osim navedene upotrebe u okviru nadgledanja socijalnih mreža, analiza sentimenata se uspešno primenjuje i u sledećim oblastima:

- **Nadgledanje brendova** – brojne kompanije koriste pogodnosti analize sentimenata da prate zadovoljstvo korisnika njihovim brendom. Time se stvara uvid u trenutne trendove, kvalitet poslovanja brenda, i omogućava brza reakcija na negativne sentimente i nezadovoljstvo korisnika.
- **Korisnički servis** – korišćenjem analize sentimenata moguće je razvrstati zahteve za podrškom prema tipu i glavnim temama. Takođe je moguće brzo identifikovati nezadovoljstvo korisnika određenim proizvodom (ili delom proizvoda) i brzo reagovati kako bi se ispravila greška.
- **Istraživanje tržišta** – korišćenjem analize sentimenata stvara se uvid u potrebe određenih grupa ljudi i na taj način omogućava ciljno reklamiranje. Istraživanjem tržišta dolazi se do informacija o tekućim trendovima, što može uticati na tip i sadržaj reklama. Ciljanjem ljudi, koji su izrazili pozitivan sentiment prema proizvodu, sličnim proizvodima, poboljšava se šansa za ponovnu kupovinu.

Pogodnosti korišćenja analize sentimenata

U odnosu na tradicionalni način obrade podataka u potrazi za sentimentom, analizom sentimenata postižu se određene prednosti:

- **Veća pouzdanost** – pri klasifikovanju sentimenata, ljudi se u najboljem slučaju slažu u 80% slučajeva zbog razlika u mišljenju, odrastanju, okruženju, itd. Eliminisanjem ljudi iz procesa, stvara se pouzdanija klasifikacija koja je lišena učešća ljudske greške.
- **Veća snaga** – broj sentimenata koje je moguće obraditi u jedinici vremena daleko premašuje mogućnosti ljudi. Broj sentimenata može da varira (npr. veća količina podataka sa sentimentima nakon izbacivanja novog proizvoda), tako da je od ključne važnosti da oni budu obrađeni na vreme.
- **Čuva vreme** – ljudi koji su bili zaduženi za analizu sentimenata se mogu rasporediti na drugim poslovima. Takođe, algoritmi za analizu sentimenata se usavršavaju svakodnevno, čime se dodatno ubrzava proces i podiže kvalitet obrade.
- **Brza reakcija** – nastaje kao rezultat gorenavedenih stavki. Brzom obradom velike količine sentimenata omogućuje se trenutna reakcija koja može znatno uticati na trenutni kvalitet poslovanja.

Problemi korišćenja analize sentimenata

Za samo dvadesetak godina, analiza sentimenata uspešno je rešila brojne probleme obrade prirodnog jezika. Međutim, konstantna promena i evolucija jezika znatno otežavaju napredak komplikujući stvaranje pravila i struktura koja se koriste za analizu. Neki od problema su deo samog jezika i načina upotrebe, dok je deo problema nastao evolucijom računarstva i Interneta. Najvažniji problemi koji usporavaju napredak discipline jesu:

- **Subjektivnost i ton** – problem određivanja koja od rečenica sadrži sentiment. Razlike u izražavanju sentimenata kod ljudi.
- **Kontekst i polaritet** – određivanje polariteta je teško samo po sebi, posebno ako se radi o složenim rečenicama. Polaritet dodatno zavisi od konteksta koji često nije dostupan.
- **Ironija i sarkazam** – teško ih je detektovati u pisanom obliku jer izgledaju isto kao i rečenice suprotnog polariteta. Negativni sentimenti se izražavaju pozitivnim rečima.
- **Poređenja** – sentiment je moguće izraziti kao poređenje naspram nekog drugog sentimenta (npr. kvalitet jednog proizvoda naspram drugog). Problem se javlja kada nisu svi akteri poređenja dostupni za analizu.
- **„Emoji“** – izražavanje emocija posebnim znakovima i kombinacijom znakova interpunkcije. Često utiču na sentiment čitavog teksta.
- **Definisanje neutralnog** – problem pri izdvajanju tekstova koji su objektivni (ne sadrže sentiment), pri izdvajanju nebitnih informacija kao i kod tekstova sa željama.
- **Ljudske greške** – problemi kod slaganja ljudi oko polariteta sentimenta, posebno kada je potrebno mišljenje više ljudi.

Posebni tipovi problema javljaju se u okviru socijalnih mreža kao platforme na kojoj se najbrže razvija i menja govor:

- **Kratke poruke** – poruke sa socijalnih mreža često imaju dužinu od par reči ili rečenica, što može biti nedovoljno za ispravnu analizu. Reči i delovi rečenica mogu biti izostavljeni kako bi se uštedelo na prostoru, što utiče na razumljivost.
- **Šum u sadržaju** – veoma neformalno pisanje poruka (loš format, skraćenje ili produženje reči, upotreba žargona, loša upotreba velikih i malih slova, itd.) otežava razumevanje ne samo za mašinu, već i za ljude. Pojedine reči mogu imati posebno značenje za jednu, a drugo značenje za druge grupe ljudi. Takođe, postoji veliki broj načina za pisanje jedne iste reči, često uz kombinovanje sa brojevima.
- **Dinamičnost** – brza promena trendova na socijalnim mrežama utiče na to da vreme igra važnu ulogu u pravilnom određivanju polariteta sentimenata. Sentiment koji je jednog dana bio pozitivan, već sledećeg može biti negativan i obrnuto.

- **Eksplicitne i implicitne informacije** – znanje o pojedinim osobinama osobe vlasnika sentimenta (npr. godine, pol, mesto) može da utiče na kategorizaciju, međutim, najčešće ove informacije nisu dostupne za analizu.
- **Višejezičnost** – kombinovanje više jezika česta je pojava u okviru poruka. Takođe, problem predstavlja obučavanje algoritama mašinskog učenja čiji rezultat analize umnogome zavisi od jezika na kojem je obučavan. Iako najrasprostranjeniji, Engleski nije jedini jezik koji ljudi masovno koriste.
- **Odnosi** – polaritet poruke može zavisiti od odnosa između ljudi ili grupacija ljudi. Dvoje ljudi mogu koristiti negativne sentimente u šali ili je sentiment određene grupe prema drugog zasnovan na pripadnosti toj grupi.

Uspešnim rešavanjem ovih problema ostvaruje se napredak u disciplini. Time nastaju brojna korisnička rešenja i biblioteke čijim se kombinovanjem rešava određeni podskup problema. Za klasifikovanje polariteta Twitter postova upotrebljen je programski jezik Python kao i skup biblioteka potrebnih za obradu, predstavljanje i rešavanje problema.

Analiza sentimenata Twitter postova u Python ekosistemu

Praktični deo rada sastoji se u predviđanju polariteta sentimenata Twitter postova korišćenjem raznih biblioteka i tehnika. Za potrebe rada korišćen je set podataka iz repozitorijuma Kaggle (https://www.kaggle.com/datasets/yasserh/twitter-tweets-sentiment-dataset). U okviru seta sadržano je 27480 postova raspoređenih po sentimentima (negative 28%, neutral 40%, positive 31%), kao i unikatni identifikator i deo teksta koji je korišćen za obeležavanje polariteta. Na slici 2 prikazane su osnovne informacije o učitanom setu, dok je na slici 3 prikazan izgled dela podataka. Set je predstavljen korišćenjem Pandas biblioteke.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 27480 entries, 0 to 27480
Data columns (total 4 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   textID         27480 non-null  object
 1   text           27480 non-null  object
 2   selected_text  27480 non-null  object
 3   sentiment      27480 non-null  object
dtypes: object(4)
memory usage: 1.0+ MB
```

Slika 2. Osnovne informacije o setu podataka

	textID	text	selected_text	sentiment
0	cb774db0d1	I'd have responded, if I were going	I'd have responded, if I were going	neutral
1	549e992a42	Sooo SAD I will miss you here in San Diego!!!	Sooo SAD	negative
2	088c60f138	my boss is bullying me...	bullying me	negative
3	9642c003ef	what interview! leave me alone	leave me alone	negative
4	358bd9e861	Sons of ****, why couldn't they put them on t...	Sons of ****,	negative
...	...	...	...	...
27476	4eac33d1c0	wish we could come see u on Denver husband l...	d lost	negative
27477	4f4c4fc327	I've wondered about rake to. The client has ...	, don't force	negative
27478	f67aae2310	Yay good for both of you. Enjoy the break - y...	Yay good for both of you.	positive
27479	ed167662a5	But it was worth it ****.	But it was worth it ****.	positive
27480	6f7127d9d7	All this flirting going on - The ATG smiles...	All this flirting going on - The ATG smiles. Y...	neutral

27480 rows × 4 columns

Slika 3. Prikaz podataka učitanog seta podataka

Za dalji tok rada identifikator i odabrani deo teksta nisu od koristi, tako da se mogu odstraniti iz seta podataka. Novonastali set podataka prikazan je na slici 4.

	text	sentiment
0	I'd have responded, if I were going	neutral
1	Sooo SAD I will miss you here in San Diego!!!	negative
2	my boss is bullying me...	negative
3	what interview! leave me alone	negative
4	Sons of ****, why couldn't they put them on t...	negative
...	...	...
27476	wish we could come see u on Denver husband l...	negative
27477	I've wondered about rake to. The client has ...	negative
27478	Yay good for both of you. Enjoy the break - y...	positive
27479	But it was worth it ****.	positive
27480	All this flirting going on - The ATG smiles...	neutral

27480 rows × 2 columns

Slika 4. Izmenjeni set podataka korišćen u nastavku rada

Tehnike koje se primenjuju za klasifikovanje polariteta nad setom podataka mogu se podeliti u dve grupe: klasifikacija zasnovana na **pravilima** i klasifikacija **mašinskim učenjem**. I jedna i druga tehnika određuju polaritet u obliku brojeva, tako da je potrebno prekodirati vrednosti. Opcije na raspolaganju su enkodiranje jedinicom za pripadnost određenoj klasi (eng. *One Hot Encode*) ili kodiranje brojčanim vrednostima. Izabrana je druga opcija zbog tipa klasa koje su raspoređene oko nule, tako da su korišćene vrednosti -1, 0 i 1 (greška je upotrebiti 0, 1 i 2 jer bi treća klasa bila duplo bolja od druge, što nije slučaj!). Kolona je dodata setu podataka kao na slici 5.

	text	sentiment	sentiment_num
1315	it was a biligual sweatshop LOL I talk 2 him ...	negative	-1
6422	On the bus to NYC http://yfrog.com/08kaifj	neutral	0
17769	: Ok its suppose 2b followfriday not unfollow ...	positive	1
16344	had such a fun time with allegra tonite!!! we ...	positive	1
15836	Very cute - I don't think I can make it to Ma...	neutral	0
18695	dunno. Maybe the flu. I feel a bitbetter now.	positive	1
11403	ya i did i seen all them but Robert	neutral	0
22643	I was the blue lol http://twitpic.com/67zgz	neutral	0
1332	Waking up early to go to the gym	neutral	0
19474	it drained my energy	negative	-1

Slika 5. Enkodiranje vrednosti setimenata brojčanim vrednostima

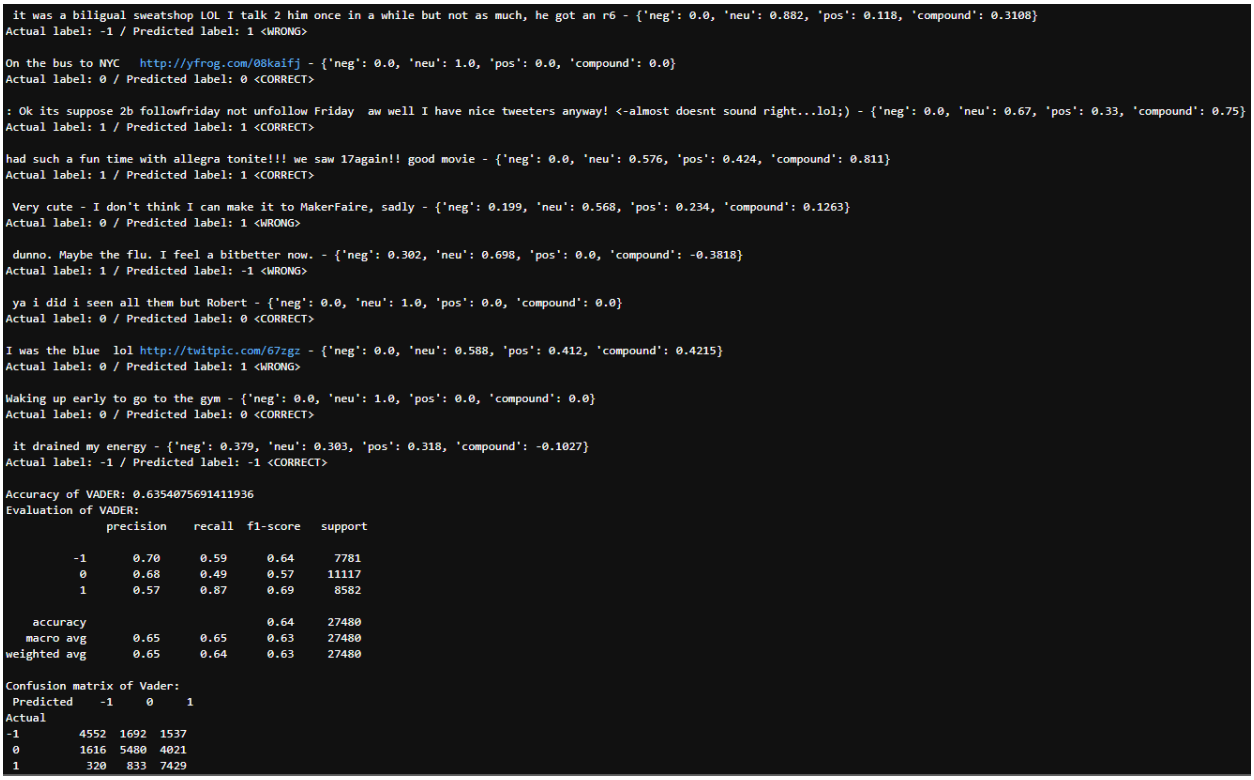
Tehnike klasifikacije korišćenjem pravila

Klasifikacija polariteta korišćenjem pravila (eng. *Rule based*) odvija se korišćenjem rečnika, zbirke ili leksikona koji sadrži pravila za određivanje polariteta, kao i polaritete i jačinu reči koje sadrže sentimente. Postoji veliki broj različitih leksikona, od kojih se mnogi nadograđuju svakodnevno. Raniji leksikoni nastaju definisanjem pravila od strane ljudi, dok noviji leksikoni koriste mašinsko učenje za definisanje i nadograđivanje pravila. U okviru rada iskorišćeni su leksikoni:

- VADER,
- TextBlob,
- AFINN.

VADER

Predstavlja akronim za „Valence Aware Dictionary and sEntiment Reasoner“. Sadrži pravila i vrednosti sentimenata u vrednostima [-4, 4] koji se skaliraju na vrednosti [-1, 1] korišćenjem VADER-a iz NLTK („Natural Language Tool Kit“) biblioteke. Određuje se vrednost posebno za negativne, neutralne i pozitivne sentimente, a onda se prikazuje zbirni sentiment. Treba napomenuti da nije vršeno nikakvo procesiranje postova jer leksikon radi na nivou rečenica i kontekst je od značaja. Primer rada VADER-a, kao i rezultati klasifikacije, dati su na slici 6.



```
 it was a biligual sweatshop LOL I talk 2 him once in a while but not as much, he got an r6 - {'neg': 0.0, 'neu': 0.882, 'pos': 0.118, 'compound': 0.3108}
Actual label: -1 / Predicted label: 1 <WRONG>

On the bus to NYC   http://yfrog.com/08kaifj - {'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}
Actual label: 0 / Predicted label: 0 <CORRECT>

: Ok its suppose 2b followfriday not unfollow Friday  aw well I have nice tweeters anyway! <-almost doesnt sound right...lol;) - {'neg': 0.0, 'neu': 0.67, 'pos': 0.33, 'compound': 0.75}
Actual label: 1 / Predicted label: 1 <CORRECT>

had such a fun time with allegra tonite!!! we saw 17again!! good movie - {'neg': 0.0, 'neu': 0.576, 'pos': 0.424, 'compound': 0.811}
Actual label: 1 / Predicted label: 1 <CORRECT>

 Very cute - I don't think I can make it to MakerFaire, sadly - {'neg': 0.199, 'neu': 0.568, 'pos': 0.234, 'compound': 0.1263}
Actual label: 0 / Predicted label: 1 <WRONG>

 dunno. Maybe the flu. I feel a bitbetter now. - {'neg': 0.302, 'neu': 0.698, 'pos': 0.0, 'compound': -0.3818}
Actual label: 1 / Predicted label: -1 <WRONG>

 ya i did i seen all them but Robert - {'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}
Actual label: 0 / Predicted label: 0 <CORRECT>

I was the blue  lol http://twitpic.com/67zgz - {'neg': 0.0, 'neu': 0.588, 'pos': 0.412, 'compound': 0.4215}
Actual label: 0 / Predicted label: 1 <WRONG>

Waking up early to go to the gym - {'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}
Actual label: 0 / Predicted label: 0 <CORRECT>

 it drained my energy - {'neg': 0.379, 'neu': 0.303, 'pos': 0.318, 'compound': -0.1027}
Actual label: -1 / Predicted label: -1 <CORRECT>

Accuracy of VADER: 0.6354075691411936
Evaluation of VADER:
              precision    recall  f1-score   support

          -1       0.70      0.59      0.64      7781
           0       0.68      0.49      0.57     11117
           1       0.57      0.87      0.69      8582

    accuracy                           0.64     27480
   macro avg       0.65      0.65      0.63     27480
weighted avg       0.65      0.64      0.63     27480

Confusion matrix of Vader:
 Predicted    -1     0     1
Actual
-1         4552  1692  1537
 0         1616  5480  4021
 1          320   833  7429
```

Slika 6. Primer rada VADER-a i dobijeni rezultati

TextBlob

Predstavlja samostalnu biblioteku za procesiranje prirodnog jezika (eng. *„Natural Language Processing“* – NLP). Polaritet sentimenata je predstavljen u obliku realnih brojeva [-1, 1]. Treba napomenuti da je i dalje u ranoj fazi razvoja. Prikaz rada i rezultati TextBlob-a dati su na slici 7.

```
 it was a biligual sweatshop LOL I talk 2 him once in a while but not as much, he got an r6 - 0.5
Actual label: -1 / Predicted label: 1 <WRONG>

On the bus to NYC   http://yfrog.com/08kaifj - 0.0
Actual label: 0 / Predicted label: 0 <CORRECT>

: Ok its suppose 2b followfriday not unfollow Friday  aw well I have nice tweeters anyway! <-almost doesnt sound right...lol;) - 0.475
Actual label: 1 / Predicted label: 1 <CORRECT>

had such a fun time with allegra tonite!!! we saw 17again!! good movie - 0.5385091145833333
Actual label: 1 / Predicted label: 1 <CORRECT>

 Very cute - I don't think I can make it to MakerFaire, sadly - 0.07500000000000001
Actual label: 0 / Predicted label: 0 <CORRECT>

 dunno. Maybe the flu. I feel a bitbetter now. - 0.0
Actual label: 1 / Predicted label: 0 <WRONG>

 ya i did i seen all them but Robert - 0.0
Actual label: 0 / Predicted label: 0 <CORRECT>

I was the blue  lol http://twitpic.com/67zgz - 0.4
Actual label: 0 / Predicted label: 1 <WRONG>

Waking up early to go to the gym - 0.1
Actual label: 0 / Predicted label: 1 <WRONG>

 it drained my energy - 0.0
Actual label: -1 / Predicted label: 0 <WRONG>

Accuracy of TextBlob: 0.5925036390101892
Evaluation of TextBlob:
              precision    recall  f1-score   support

          -1       0.70      0.40      0.51      7781
           0       0.57      0.59      0.58     11117
           1       0.57      0.77      0.66      8582

    accuracy                           0.59     27480
   macro avg       0.61      0.59      0.58     27480
weighted avg       0.61      0.59      0.58     27480

Confusion matrix of TextBlob:
 Predicted    -1     0     1
Actual
-1         3136  3139  1506
 0         1112  6512  3493
 1          261  1687  6634
```

Slika 7. Primer rada TextBlob-a i rezultati

AFINN

Predstavlja analizu sentimenata zasnovanu na listi reči. Svaka reč iz leksikona određena je polaritetom u celobrojnim vrednostima [-5, 5]. Podržava više jezika. Verzija za Python dolazi u obliku samostalne biblioteke. Iako je nastavak rada na projektu prekinut, i dalje daje dobre rezultate u odnosu na druge leksikone. Podržana je i opciona obrada emotikona. Na slici 8 prikazani su primeri kao i rezultati rada AFINN-a.

```
 it was a biligual sweatshop LOL I talk 2 him once in a while but not as much, he got an r6 - 3.0
Actual label: -1 / Predicted label: 1 <WRONG>

On the bus to NYC   http://yfrog.com/08kaifj - 0.0
Actual label: 0 / Predicted label: 0 <CORRECT>

: Ok its suppose 2b followfriday not unfollow Friday  aw well I have nice tweeters anyway! <-almost doesnt sound right...lol;) - 8.0
Actual label: 1 / Predicted label: 1 <CORRECT>

had such a fun time with allegra tonite!!! we saw 17again!! good movie - 7.0
Actual label: 1 / Predicted label: 1 <CORRECT>

 Very cute - I don't think I can make it to MakerFaire, sadly - 0.0
Actual label: 0 / Predicted label: 0 <CORRECT>

 dunno. Maybe the flu. I feel a bitbetter now. - -2.0
Actual label: 1 / Predicted label: -1 <WRONG>

 ya i did i seen all them but Robert - 0.0
Actual label: 0 / Predicted label: 0 <CORRECT>

I was the blue  lol http://twitpic.com/67zgz - 3.0
Actual label: 0 / Predicted label: 1 <WRONG>

Waking up early to go to the gym - 0.0
Actual label: 0 / Predicted label: 0 <CORRECT>

 it drained my energy - -2.0
Actual label: -1 / Predicted label: -1 <CORRECT>

Accuracy of AFINN: 0.6475254730713246
Evaluation of AFINN:
              precision    recall  f1-score   support

          -1       0.71      0.60      0.65      7781
           0       0.69      0.52      0.60     11117
           1       0.58      0.85      0.69      8582

    accuracy                           0.65     27480
   macro avg       0.66      0.66      0.65     27480
weighted avg       0.66      0.65      0.64     27480

Confusion matrix of AFINN:
 Predicted    -1     0     1
Actual
-1         4665  1633  1483
 0         1551  5828  3738
 1          319   962  7301
```

Slika 8. Primer rada AFINN-a i rezultati

Uporedni rezultat rada sva tri leksikona prikazan je na slici 9. U svrhe poređenja, izvršeno je zaokruživanje vrednosti svih leksikona na -1, 0 i 1. Sa slike se može uočiti da VADER i AFINN daju približne rezultate od 63% i 64%, dok je TextBlob neznatno lošiji sa 59%.

	Model	Accuracy
0	VADER	0.635408
1	TextBlob	0.592504
2	AFINN	0.647525

Slika 9. Uporedni prikaz rezultata rada leksikona

Predviđanje polariteta mašinskim učenjem

Mašinsko i duboko učenje predstavlja budućnost razvoja analize sentimenata. Veliki broj tehnika se sa lakoćom može primeniti na klasifikaciju polariteta jer predstavlja vid nadgledanog učenja. Međutim, rad algoritama mašinskog učenja zasniva se na pojedinačnim tokenima, tako da je tekst potrebno dodatno preprocesirati. Upotrebljeno je više različitih tokenajzera (algoritmi koji rastavljaju rečenice na sastavne delove – tokene), od kojih su neki samostalni, dok drugi dolaze u okviru NLTK biblioteke. Na slici 10 dat je uporedni prikaz rezultata 5 algoritama.

	text	sentiment	sentiment_num	Tweet Tokenizer	Tweet Tokenizer Proc	Word Tokenize	Casual Tokenize	Twokenize	Twikenizer
1315	it was a biligual sweatshop LOL I talk 2 him ...	negative	-1	[it, was, a, biligual, sweatshop, LOL, I, talk...	[biligual, sweatshop, lol, talk, not, much, get]	[it, was, a, biligual, sweatshop, LOL, I, talk...	[it, was, a, biligual, sweatshop, LOL, I, talk...	[it, was, a, biligual, sweatshop, LOL, I, talk...	[it, was, a, biligual, sweatshop, LOL, I, talk...
6422	On the bus to NYC http://yfrog.com/08kaifj	neutral	0	[On, the, bus, to, NYC, http://yfrog.com/08kaifj]	[bus, nyc]	[On, the, bus, to, NYC, http, :, //yfrog.com/0...	[On, the, bus, to, NYC, http://yfrog.com/08kaifj]	[On, the, bus, to, NYC, http://yfrog.com/08kaifj]	[On, the, bus, to, NYC, http, :, /, /, yfrog, ...
17769	: Ok its suppose 2b followfriday not unfollow ...	positive	1	[:, Ok, its, suppose, 2b, followfriday, not, u...	[suppose, folowfriday, not, unfolow, friday, w...	[:, Ok, its, suppose, 2b, followfriday, not, u...	[:, Ok, its, suppose, 2b, followfriday, not, u...	[:, Ok, its, suppose, 2b, followfriday, not, u...	[:, Ok, its, suppose, 2b, followfriday, not, u...
16344	had such a fun time with allegra tonite!!! we ...	positive	1	[had, such, a, fun, time, with, allegra, tonit...	[fun, time, alegra, tonite, saw, again, good, ...	[had, such, a, fun, time, with, allegra, tonit...	[had, such, a, fun, time, with, allegra, tonit...	[had, such, a, fun, time, with, allegra, tonit...	[had, such, a, fun, time, with, allegra, tonit...
15836	Very cute - I don't think I can make it to Ma...	neutral	0	[Very, cute, -, I, don't, think, I, can, make,...	[cute, think, make, makerfaire, sadly]	[Very, cute, -, I, do, n't, think, I, can, mak...	[Very, cute, -, I, don't, think, I, can, make,...	[Very, cute, -, I, don't, think, I, can, make,...	[Very, cute, -, I, don, ', t, think, I, can, m...
18695	dunno. Maybe the flu. I feel a bitbetter now.	positive	1	[dunno, ., Maybe, the, flu, ., I, feel, a, bit...	[duno, maybe, flu, feel, bitbeter]	[dunno, ., Maybe, the, flu, ., I, feel, a, bit...	[dunno, ., Maybe, the, flu, ., I, feel, a, bit...	[dunno, ., Maybe, the, flu, ., I, feel, a, bit...	[dunno, ., Maybe, the, flu, ., I, feel, a, bit...
11403	ya i did i seen all them but Robert	neutral	0	[ya, i, did, i, seen, all, them, but, Robert]	[see, robert]	[ya, i, did, i, seen, all, them, but, Robert]	[ya, i, did, i, seen, all, them, but, Robert]	[ya, i, did, i, seen, all, them, but, Robert]	[ya, i, did, i, seen, all, them, but, Robert]
22643	I was the blue lol http://twitpic.com/67zgz	neutral	0	[I, was, the, blue, lol, http://twitpic.com/67...	[blue, lol]	[I, was, the, blue, lol, http, :, //twitpic.co...	[I, was, the, blue, lol, http://twitpic.com/67...	[I, was, the, blue, lol, http://twitpic.com/67...	[I, was, the, blue, lol, http, :, /, /, twitpi...
1332	Waking up early to go to the gym	neutral	0	[Waking, up, early, to, go, to, the, gym]	[wake, early, gym]	[Waking, up, early, to, go, to, the, gym]	[Waking, up, early, to, go, to, the, gym]	[Waking, up, early, to, go, to, the, gym]	[Waking, up, early, to, go, to, the, gym]
19474	it drained my energy	negative	-1	[it, drained, my, energy]	[drain, energy]	[it, drained, my, energy]	[it, drained, my, energy]	[it, drained, my, energy]	[it, drained, my, energy]

Slika 10. Uporedni prikaz rezultata tokenajzera.

Na osnovu nekoliko proizvoljnih kriterijuma autora (obrada linkova, znakova interpunkcije, produženih reči, emotikona, itd.), izvršeno je rangiranje po kvalitetu i odabran najbolji za korišćenje u daljem toku rada:

1. **Tweet Tokenizer** (NLTK)
2. Twokenize
3. Casual Tokenize (NLTK)
4. Word Tokenize (NLTK)
5. Twikenizer

Sledeći korak u obradi sastoji se u prečišćavanju šuma u tokenima i pokušaju standardizacije. Idući redom, izvršene su sledeće izmene:

- Postavljanje svih slova na mala i uklanjanje znakova interpunkcije;
- Uklanjanje reči koje sadrže tačku u sebi (čime se uklanjanju URL adrese) ili počinju znakovima „@“, „#“, „_“;

- Uklanjanje brojeva iz reči;
- Uklanjanje ponovljenih karaktera iz reči, uz proveru za validnost reči za 1, 2 ili 3 ponovljena karaktera. Reči van rečnika koje su često u upotrebi („wanna“, „gonna“, „lemme“, itd.) svedene su na jedan ponovljen karakter, ali to ne utiče negativno na rad klasifikacije (nebitno je broj slova ima ukoliko je isti token);
- Lematizacija reči, odnosno vraćanje reči na osnovni oblik. Za razliku od prethodnih koraka gde su izmene vršene na nivou reči, lematizacija se vrši na nivou skupa tokena zbog toga što je kontekst rečenice bitan za pravilno označavanje dela govora (eng. *„Part Of Speech“* - POS);
- Podela reči koje sadrže apostrof ' i uklanjanje drugog dela koji se sastoji od jednog ili dva slova (u najvećem broju slučajeva);
- Uklanjanje reči koje su kraće od 3 slova (ne nose veliko značenje) ili previše dugih reči (upotrebljen je broj od 45 prema najdužoj reči u velikim rečnicima). Izostavljena je reč „no“ zbog negacije;
- Uklanjanje „stop“ reči iz tokena (reči koje ne nose značenje same po sebi, već dopunjuju rečenicu) kao i token „...“ koji je zaostao posle uklanjanja interpunkcija.

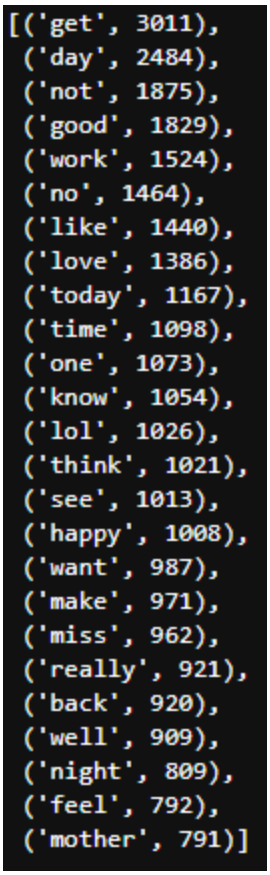
Lista dopunskih reči iz NLTK biblioteke sadrži veliki broj negacija koje su uklonjene iz liste tokena. To bi moglo predstavljati problem, tako da je izvršeno testiranje van okvira rada nad algoritmom koji je dao najbolje rezultate pri klasifikaciji kako bi se utvrdio uticaj negativnih dopunskih reči na rezultat. Došlo je do poboljšanja, međutim poboljšanje je bilo minimalno na trećoj decimali (max. 0.5% poboljšanje), tako da je nastavljen rad bez dopunskih reči.

Na slici 11 može se videti rezultat obrade tokena, kao i uporedno stanje pre obrade.

	text	sentiment	sentiment_num	Tweet Tokenizer	Tweet Tokenizer Proc
1315	it was a biligual sweatshop LOL I talk 2 him ...	negative	-1	[it, was, a, biligual, sweatshop, LOL, I, talk...	[biligual, sweatshop, lol, talk, not, much, get]
6422	On the bus to NYC http://yfrog.com/08kaifj	neutral	0	[On, the, bus, to, NYC, http://yfrog.com/08kaifj]	[bus, nyc]
17769	: Ok its suppose 2b followfriday not unfollow ...	positive	1	[:, Ok, its, suppose, 2b, followfriday, not, u...	[suppose, folowfriday, not, unfolow, friday, w...
16344	had such a fun time with allegra tonite!!! we ...	positive	1	[had, such, a, fun, time, with, allegra, tonit...	[fun, time, alegra, tonite, saw, good, movie]
15836	Very cute - I don't think I can make it to Ma...	neutral	0	[Very, cute, -, I, don't, think, I, can, make,...	[cute, think, make, makerfaire, sadly]
18695	dunno. Maybe the flu. I feel a bitbetter now.	positive	1	[dunno, ., Maybe, the, flu, ., I, feel, a, bit...	[duno, maybe, flu, feel, bitbeter]
11403	ya i did i seen all them but Robert	neutral	0	[ya, i, did, i, seen, all, them, but, Robert]	[see, robert]
22643	I was the blue lol http://twitpic.com/67zgz	neutral	0	[I, was, the, blue, lol, http://twitpic.com/67...	[blue, lol]
1332	Waking up early to go to the gym	neutral	0	[Waking, up, early, to, go, to, the, gym]	[wake, early, gym]
19474	it drained my energy	negative	-1	[it, drained, my, energy]	[drain, energy]

Slika 11. Rezultat procesiranja tokena

Iako nepotrebno za dalji rad klasifikacije polariteta, izvršeno je izračunavanje frekventne distribucije reči, tj. koliko se određena reč ponavlja u tekstu. Rezultat najkorišćenijih reči dat je na slici 12.



Slika 12. Najkorišćenije reči u setu podataka po broju ponavljivanja

Sledeći korak jeste pretvaranje reči (tokena) u oblik koji je razumljiv algoritmima mašinskog učenja. Korišćene su dve tehnike**: „Bag of Words“** (BoW) i **„Term Frequency-Inverse Document Frequency“** (TfIdf). BoW tehnikom se kreira rasuta matrica koja sadrži 1 na mestu postojanja reči u određenom postu, dok TfIdf kreira rasutu matricu koja sadrži frekvenciju reči određene formulom. Vrednost je veća ukoliko se reč javlja češće u postu (važna reč za post), a istovremeno se smanjuje ukoliko se javlja u većem broju postova (generalno manja važnost reči). Obe matrice su dopunjene nulama kako bi bile upotrebljive. TfIdf reprezentacija sadrži i ngrame (skupove reči) od 2 i 3 reči (koji se pojavljuju najmanje dva puta), kako bi se poboljšao rezultat, ali ne sadrži reči koje se pojavljuju jednom. Tehniku ngrama kod BoW reprezentacije (novonastala „Bag of Ngrams“ – BoN reprezentacija) nije bilo moguće primeniti zbog ograničenosti sistemskih resursa. Na slici 13 dat je primer izgleda rečnika BoW reprezentacije koji sadrži par reč i broj (jednoznačno reprezentuje reč), dok je na slici 14 dat primer izgleda TfIdf rečnika koji sadrži par reč i broj (fekvencija dobijena formulom). Broj reči u okviru BoW rečnika iznosi 17964, dok broj reči i ngrama u okviru TfIdf rečnika iznosi 23658.

```
BoW - Number of sentences: 27480, Unique words: 17964
Example:
[('bylaurenluke', 2260),
 ('superpower', 15216),
 ('brings', 1995),
 ('lexi', 9026),
 ('krathong', 8695),
 ('enlgland', 4970),
 ('muzik', 10467),
 ('nba', 10605),
 ('cade', 2279),
 ('shelter', 13950)]
```

Slika 13. Primer BoW rečnika

```
TfIdf - Number of sentences: 27480, Unique words and ngrams: 23658
Example:
[('get hang', 9.42949066654802),
 ('bikini', 10.122637847107965),
 ('goto', 9.834955774656184),
 ('small', 7.414587646005755),
 ('hahahaha', 8.043196305428129),
 ('thought would', 10.122637847107965),
 ('shape', 9.024025558439854),
 ('school day', 9.42949066654802),
 ('sunshine gona', 10.122637847107965),
 ('ant', 9.42949066654802)]
```

Slika 14. Primer TfIdf rečnika

Kreiranjem brojčanih reprezentacija, ostvareni su uslovi za korišćenje algoritama mašinskog učenja. Za potrebe rada mašinskog učenja upotrebljena je biblioteka SKLearn. Na samom početku, izvršeno je predviđanje lažnim/maketa (eng. *„Dummy“*) klasifikatorom koji nasumično dodeljuje klase. Podaci su pre početka rada podeljeni u odnosu 80% prema 20%, na trening i test podatke. Rezultat je približno broj 1 podeljen brojem klasa, dok na rezultat utiče i broj pripadnika određene klase. Kako u konkretnom slučaju postoje tri klase, preciznost je u okviru očekivanih 33%, što je prikazano na slici 15. Rezultat je korišćen u svrhe poređenja sa ostalim algoritmima.

```
Base Dummy BoW   0.32787481804949054
Base Dummy TdIdf 0.32787481804949054
```

Slika 15. Rezultat rada „dummy“ klasifikatora

Algoritmi korišćeni u svrhe predviđanja jesu:

- **Multinomial Naive Bayes** (Naivni Bajesov za više klasa)
- **Logistic Regression** (Logistička regresija)
- **Linear SVC** (Linearni klasifikator potpornih vektora)
- **SGD Classifier** (Klasifikator stohastički gradijent opadanja nad potpornim vektorima)
- **Random Forest Classifier** (Klasifikator nasumične šume; šume slučajnih odluka)

U nastavku će biti prikazani rezultati rada svih algoritama, dok je diskusija o rezultatima ostavljena nakon prikazivanja svih rezultata. Svaki od klasifikatora treniran je 5 puta korišćenjem tehnike kros-validacije (prvi put se uzima u obzir prvih 20% podataka za test, zatim sledećih 20% i tako do pete iteracije gde se koristi poslednjih 20%). Za svaku kombinaciju klasifikator i reprezentacija, prikazana je preciznost nad trening podacima svake iteracija, srednja vrednost preciznosti (ukoliko je dostupna) i preciznost nad test podacima.

Multinomial Naive Bayes

```
Multinomial Naive Bayes scores BoW: [0.65476461 0.64225608 0.6472595  0.65863088 0.65468608]
Mean score: 0.6515194307972825
Test score: 0.6550218340611353
```

Slika 16. Rezultati algoritma MNB za BoW reprezentaciju podataka

```
Multinomial Naive Bayes scores TfIdf: [0.63133955 0.62952013 0.62065044 0.63179441 0.63717015]
Mean score: 0.6300949360998265
Test score: 0.6381004366812227
```

Slika 17. Rezultati algoritma MNB za TfIdf reprezentaciju podataka

Logistic Regression

```
Logisitc Regression scores BoW: [0.68796907 0.6906982  0.69092563        nan 0.70200182]
Mean score: nan
Test score: 0.6925036390101892
```

Slika 18. Rezultati algoritma LR za BoW reprezentaciju podataka

```
Logisitc Regression scores TfIdf: [0.68887878 0.69115306 0.68592222 0.69752104 0.69131028]
Mean score: 0.6909570757462851
Test score: 0.6932314410480349
```

Slika 19. Rezultati algoritma LR za TfIdf reprezentaciju podataka

Linear SVC

```
Linear SVM scores BoW: [0.66340687 0.666136   0.66704571 0.67295884 0.67470428]
Mean score: 0.6688503390619338
Test score: 0.6663027656477438
```

Slika 20. Rezultati algoritma SVC za BoW reprezentaciju podataka

```
Linear SVM scores TfIdf: [0.67068456 0.68046395 0.66499886 0.67455083 0.67993631]
Mean score: 0.6741269018105859
Test score: 0.6750363901018923
```

Slika 21. Rezultati algoritma SVC za TfIdf reprezentaciju podataka

SGD Classifier

```
Linear SVM (SGD) scores BoW: [0.69865818 0.69820332 0.69797589        nan        nan]
Mean score: nan
Test score: 0.7083333333333334
```

Slika 22. Rezultati algoritma SGD za BoW reprezentaciju podataka

```
Linear SVM (SGD) scores TfIdf: [0.69752104 0.69843075 0.69456448 0.70798272 0.69267516]
Mean score: 0.6982348271621213
Test score: 0.7079694323144105
```

Slika 23. Rezultati algoritma SGD za TfIdf reprezentaciju podataka

Random Forest Classifier

```
Random Forest scores BoW: [0.6784171  0.70138731 0.69774846 0.68796907 0.69517743]
Mean score: 0.6921398761625668
Test score: 0.7003275109170306
```

Slika 24. Rezultati algoritma RFC za BoW reprezentaciju

```
Random Forest scores TfIdf: [0.6881965  0.70025017 0.69274505 0.69115306 0.69494995]
Mean score: 0.6934589470072552
Test score: 0.7030567685589519
```

Slika 25. Rezultati algoritma RFC za TfIdf reprezentaciju

Diskusija o rezultatima

U tabeli 1 dat je uporedni prikaz svih algoritama sa srednjim vrednostima preciznosti klasifikacije trening i test podataka zbog nemogućnosti da rezultati budu predstavljeni u okviru Pandas tabele. Kod klasifikatora sa NaN vrednostima, srednja vrednost je izračunata na osnovu dostupnih rezultata. Najbolji rezultati označeni su podebljanim brojevima.

Algoritam	Trening BoW %	Test BoW %	Trening TfIdf %	Test TfIdf %
MNB	65.15	65.50	63.00	63.81
LR	69.29	69.25	69.10	69.32
SVC	66.89	66.63	67.41	67.51
SGD	**69.83**	**70.83**	**69.82**	**70.80**
RFC	69.21	70.03	69.34	70.30

Tabela 1. Uporedni prikaz rezultata rada algoritama mašinskog učenja

Analizom rezultata iz tabele dolazimo do zaključka da je algoritam **SGD** neznatno bolji od ostalih. Slede, gledano prema rezultatima preciznosti nad test podacima: RFC, LR, SVC, MNB. Treba napeomenuti da i prostiji algoritmi (u ovom slučaju logistička regresija) daju približno iste rezultate kao i mnogo složeniji modeli (u ovom slučaju skup stabala odluka).

Reprezentacija reči putem „Word Embedding“-a

Drugačiji pristup u prikazivanju tokena sastoji se u treniranju (ili korišćenju prethodno treniranih) modela „ugrađivanja reči“. Najpoznatiji model nosi naziv „**Word2Vec**“ (reč u vektor). Pristup se zasniva na predstavljanju reči u obliku vektora sa ciljem pronalaženja sličnosti između reči na osnovu pozicija vektora u prostoru. Reči koje su po značenju slične se nalaze bliže jedna drugoj i obrnuto. Postoje više varijanti od kojih se izdvajaju dve: „**Continuous Bag of Words**“ (CBoW) i „**Skip-gram**“ (SG). CBoW radi na principu predviđanja trenutne reči na osnovu prozora sa okolnim rečima (ne zavisi od redosleda reči), dok SG koristi princip predviđanja okolnih reči na osnovu trenutne reči (redosled i blizina reči utiču na težine).

Postoje i drugi modeli konverzije reči u vektore. Kompanija Facebook je objavila svoje rešenje u obliku „**fastText**“-a, dok „**GloVe**“ nastaje kao projekat na Stenford univerzitetu. Principi rada se neznatno razlikuju, ali je rezultat i dalje u vidu vektora. „fastText“ radi na principu sličnom modelu „Word2Vec“, dok „GloVe“ koristi tzv. globalne vektore (odatle i naziv) koji određuju slučnost reči u širem okviru (čitavoj zbirci reči u obrađenim dokumentima). Kao podvarijantu „Word2Vec“-a, treba napomenuti „**Doc2Vec**“ koja uzima u obzir sličnost reči na nivou paragrafa. Za potrebe rada korišćena je tehnika određivanja „srednjih vektora“ tj. deljenje vektora brojem reči, kao i „**Spacy**“ biblioteka koja sadrži veliki rečnik sa odgovarajućim vrednostima vektora.

Word2Vec

Za predstavljanje reči preko „Word2Vec“ modela korišćene su obe varijante (CBoW i SG). Minimalno pojavljivanje reči postavljeno je na 1 (uzima u obzir sve reči) dok je prozor postavljen na 2, što znači da uzima u obzir dve dodatne reči (ukupno tri). Razlog za mali prozor jeste postojanje velikog broja malih postova. Obavljeno je treniranje modela. Izvršeno je izdvajanje vektora za svaku reč na odgovarajućoj poziciji. Zbog nepravilne dimenzije novonastale matrice, vektori su svedeni na jednake veličine. Izvršeno je pokomponentno sabiranje vektora. Ukoliko su sve reči iz posta uklonjene, vrši se dodela nasumičnog vektora sa malim vrednostima [-0.2, 0.2] kako ne bi uticale na predviđanje. Dužina pojedinačnih vektora iznosi 100, tako da su dimenzije matrice (27480, 100).

fastText

Sličan princip rada kao kod prethodnog modela, tako da su iskorišćena ista podešavanja i dobijena matrica je istih dimenzija, samo se vrednosti vektora razlikuju.

Doc2Vec

Princip rada zahteva prethodno označivanje dokumenata (redni brojevi). Vrši se izgradnja rečnika, a zatim i treniranje modela. Rezultat predstavlja matrica istih dimenzija kao i prethodni modeli, sa različitim vrednostima vektora.

Averaged Word Vectors

Sastoji se u korišćenju „Word2Vec“ modela uz deljenje vektora brojem reči. Dobija se takođe matrica istih dimenzija sa različitim vrednostima vektora.

GloVe

Upotrebljena je prethodno trenirana verzija modela nad 27 milijardi Twitter postova sa veličinom vektora od 100 (dostupne i verzije sa manjim ili većim vektorima). Uzete su vrednosti vektora za svaku reč koja je dostupna u modelu, dok su reči van rečnika zamenjene vektorom jednake dužine sa vrednostima od [-0.2, 0.2]. Takođe je izvršeno sumiranje vektora na kraju.

Spacy

Korišćena je velika biblioteka reči koja sadrži preko 500 hiljada reči obeležene vektorima dužine 300. Ukoliko je reč van vektora, dopunjuje se nasumičnim vektorom dužine 300 sa vrednostima [-0.2, 0.2]. Jedina reprezentacija sa dimenzijama (27480, 300).

Treniranje modela mašinskog učenja nad vektorima reči

Izvršeno je treniranje istim modelima kao i kod prethodnih reprezentacija. Izvršene su minimalne izmene kako bi se što više poboljšao kvalitet predviđanja. Kao i kod prethodne sekcije treniranja modela, biće predstavljeni rezultati kros-validacije, dok je diskusija o rezultatima izvršena nad uporednim pregledom srednjih vrednosti preciznosti nad trening i test podacima.

Multinomial Naive Bayes

Pre upotrebe vektora potrebno je izvršiti skaliranje jer model zahteva nenegativne vrednosti. Upotrebom skaliranja se omogućuje rad modela, ali preciznost značajno opada u poređenju sa drugim modelima. Na slici 26 prikazani su rezultati treniranja modela nad svakom vektorskom reprezentacijom reči.

```
Naive Bayes scores Word2Vec CBoW:      [0.42915624 0.41960428 0.4296111  0.42574483 0.42902639]
Naive Bayes scores Word2Vec Skip-Gram: [0.43256766 0.42369798 0.4321128  0.42324312 0.43198362]
Naive Bayes scores FastText CBoW:      [0.42051399 0.40641346 0.42028656 0.41869456 0.42151956]
Naive Bayes scores FastText Skip-Gram: [0.43302252 0.41482829 0.42824653 0.41755743 0.42902639]
Naive Bayes scores Doc2Vec:            [0.44393905 0.43438708 0.43870821 0.43484194 0.44972702]
Naive Bayes scores Avg. Vectors:       [0.47304981 0.4746418  0.4664544  0.47327723 0.47520473]
Naive Bayes scores Glove:              [0.41642029 0.41687514 0.415738   0.41346373 0.41537762]
Naive Bayes scores Spacy:              [0.44575847 0.44211963 0.44712304 0.44575847 0.44836215]
```

Slika 26. Rezultati rada algoritma MNB za vektorske reprezentacije reči

Logistic Regression

Nisu vršene nikakve izmene na modelu. Zbog manjeg broja dimenzija omogućeno je lakše konvergiranje modela, tako da se ne javljaju NaN vrednosti. Rezultati rada prikazani su na slici 27.

```
Logistic Regression scores Word2Vec CBoW:      [0.60814191 0.60495793 0.60632249 0.60359336 0.60850773]
Logistic Regression scores Word2Vec Skip-Gram: [0.62178758 0.63202183 0.62451672 0.62337958 0.62420382]
Logistic Regression scores FastText CBoW:      [0.58107801 0.58357971 0.59040255 0.56925176 0.5878071 ]
Logistic Regression scores FastText Skip-Gram: [0.61269047 0.62178758 0.60996134 0.61769388 0.61260237]
Logistic Regression scores Doc2Vec:            [0.63042984 0.62883784 0.63679782 0.63588811 0.63466788]
Logistic Regression scores Avg. Vectors:       [0.62701842 0.61928588 0.62087787 0.61746645 0.61919927]
Logistic Regression scores Glove:              [0.66954742 0.65180805 0.6554469  0.65362747 0.65991811]
Logistic Regression scores Spacy:              [0.6743234  0.66931999 0.6800091  0.67318626 0.67470428]
```

Slika 27. Rezultati rada algoritma LR za vektorske reprezentacije reči

Linear SVM

Postavljen je parametar koji omogućuje brži rad algoritma u slučajevima kada broj uzoraka premašuje broj dimenzija. Rezultati rada prikazani su na slici 28.

```
Support Vector Machines scores Word2Vec CBoW:      [0.61428247 0.61109848 0.60836934 0.60382079 0.60873521]
Support Vector Machines scores Word2Vec Skip-Gram: [0.62337958 0.62883784 0.625199   0.62724585 0.62829845]
Support Vector Machines scores FastText CBoW:      [0.60200136 0.60950648 0.5990448  0.5867637  0.60395814]
Support Vector Machines scores FastText Skip-Gram: [0.6211053  0.62474414 0.61496475 0.62428929 0.62215651]
Support Vector Machines scores Doc2Vec:            [0.63315897 0.62770071 0.63247669 0.63861724 0.63193813]
Support Vector Machines scores Avg. Vectors:       [0.63384126 0.62474414 0.63293154 0.62997498 0.62420382]
Support Vector Machines scores Glove:              [0.66431658 0.65590175 0.66113259 0.65499204 0.66264786]
Support Vector Machines scores Spacy:              [0.68182852 0.67500569 0.68091881 0.67932681 0.67879891]
```

Slika 28. Rezultati rada algoritma SVM za vektorske reprezentacije reči

SGD Classifier

Povećan je broj iteracija za potrebe konvergiranja modela i eventualnog poboljšanja preciznosti. Rezultati rada prikazani su na slici 29.

```
Stohastic Gradient Descent (SVM) scores Word2Vec CBoW:      [0.55628838 0.59153969 0.60382079 0.51012054 0.58257507]
Stohastic Gradient Descent (SVM) scores Word2Vec Skip-Gram: [0.54241528 0.51648851 0.61678417 0.57630202 0.56619654]
Stohastic Gradient Descent (SVM) scores FastText CBoW:      [0.48987946 0.49556516 0.51284967 0.37411872 0.51819836]
Stohastic Gradient Descent (SVM) scores FastText Skip-Gram: [0.48669547 0.58585399 0.55310439 0.58585399 0.54299363]
Stohastic Gradient Descent (SVM) scores Doc2Vec:            [0.53013418 0.59472368 0.57175347 0.59153969 0.5843949 ]
Stohastic Gradient Descent (SVM) scores Avg. Vectors:       [0.61746645 0.61928588 0.61974073 0.62383443 0.616697  ]
Stohastic Gradient Descent (SVM) scores Glove:              [0.58517171 0.63452354 0.60859677 0.54650898 0.56574158]
Stohastic Gradient Descent (SVM) scores Spacy:              [0.63361383 0.58289743 0.58744599 0.59290425 0.60191083]
```

Slika 29. Rezultati rada algoritma SGD za vektorske reprezentacije reči

Random Forest Classifier

Broj stabala odluke je povećan na 100 radi poboljšanja preciznosti zbog toga što se treniranje modela odvija brže usled smanjenih dimenzija. Rezultati rada prikazani su na slici 30.

```
Random Forest scores Word2Vec CBoW:      [0.55060268 0.55378667 0.54946554 0.55424153 0.56619654]
Random Forest scores Word2Vec Skip-Gram: [0.56584035 0.57539231 0.5801683  0.57266318 0.5843949 ]
Random Forest scores FastText CBoW:      [0.49965886 0.48987946 0.49897657 0.51057539 0.50523203]
Random Forest scores FastText Skip-Gram: [0.54628156 0.54628156 0.55446896 0.55310439 0.56119199]
Random Forest scores Doc2Vec:            [0.58630885 0.59290425 0.59108483 0.58267    0.59190173]
Random Forest scores Avg. Vectors:       [0.63088469 0.62087787 0.62883784 0.62224244 0.62261146]
Random Forest scores Glove:              [0.63270412 0.623607   0.62337958 0.62406186 0.64058235]
Random Forest scores Spacy:              [0.62019559 0.62952013 0.61564703 0.60859677 0.63034577]
```

Slika 30. Rezultati rada algoritma RFC za vektorske reprezentacije reči

Diskusija o rezultatima

Na slici 31 (podeljena na dva dela zbog preglednosti) prikazani su uporedni rezultati obučavanja svih modela nad svim reprezentacijama reči. Kolone predstavljaju rezultate preciznosti nad trening i test podacima, dok redovi predstavljaju različite modele.

	Representation	MultinomoalNB MS	MultinomoalNB TS	Logistic Regression MS	Logistic Regression TS
0	Word2Vec CBoW	0.426629	0.433406	0.606305	0.619905
1	Word2Vec Skip-Gram	0.428721	0.440138	0.625182	0.627729
2	FastText CBoW	0.417486	0.425582	0.582424	0.594978
3	FastText Skip-Gram	0.424536	0.435953	0.614947	0.621179
4	Doc2Vec	0.440321	0.447780	0.633324	0.638464
5	Avg. Vector	0.472526	0.486536	0.620770	0.620451
6	Glove	0.415575	0.431405	0.658070	0.661936
7	Spacy	0.445824	0.462882	0.674309	0.683588

Support Vector Machines MS	Support Vector Machines TS	Stohastic Gradient Descent (SVM) MS	Stohastic Gradient Descent (SVM) TS	Random Forest MS	Random Forest TS
0.609261	0.618086	0.568869	0.547307	0.554859	0.569505
0.626592	0.632278	0.563637	0.555495	0.575692	0.586426
0.600255	0.614447	0.478122	0.547489	0.500864	0.505822
0.621452	0.629367	0.550900	0.582424	0.552266	0.558042
0.632778	0.639374	0.574509	0.595706	0.588974	0.596070
0.629139	0.628457	0.619405	0.620997	0.625091	0.626092
0.659798	0.665757	0.588109	0.579148	0.628867	0.631914
0.679176	0.685044	0.599754	0.575873	0.620861	0.631186

Slika 31. Uporedni prikaz rezultata treniranja modela nad vektorskim reprezentacijama reči

Rezultati označeni žutom bojom predstavljaju najbolje rezultate dobijene obučavanjem modela nad trening i test podacima. Kao najbolji model pokazao se model potpornih vektora koji je u svakoj vektorskoj reprezentaciji postigao najbolji rezultat, osim kod preciznosti nad test podacima za Word2Vec CBoW reprezentaciju i kod preciznosti nad trening podacima za Doc2Vec reprezentaciju. U oba slučaja je logistička regresija bila bolja.

Rezultati označeni zelenim okvirom predstavljaju najbolje rezultate za određene reprezenzaije. Predstavljanje srednjim vektorima se najbolje pokazalo kod Naivnog Bajesovog klasifikatora i kod klasifikatora sa stohastičnim gradijentom opadanja. Korišćenje utreniranog GloVe modela pokazalo se najbolje kod klasifikatora nasumične šume. Leksikon Spacy biblioteke se pokazao najbolje u logističkoj regresiji i kod modela potpornih vektora, a ujedno i najbolje ako se uzmu u obzir rezultati svih modela.

Treba napomenuti da su razlike u modelima i dalje veoma male i da jednostavniji modeli daju bolje rezultate od složenijih. Poređenjem sa reprezentacijama BoW i TfIdf dolazimo do zaključka da su se vektori u ovom slučaju pokazali kao lošiji izbor. Razlog može biti veliki broj nepravilnih reči kod Twitter postova ili sam način obrade teksta i dodeljivanja vrednosti vektora.

Korišćenje metoda dubokog učenja

Na kraju treba posvetiti pažnju modelima dubokog učenja kao jednim od najperspektivnijih pravaca u daljem razvoju discipline analize sentimenata. Sastoje se u korišćenju mreža sastavljenih od velikog broja perceptrona (osnovna jedinica mreže koja obavlja jednostavnu funkciju) raspoređenih u više slojeva (nivoa). Izdvajaju se ulazni i izlazni sloj, dok ostali predstavljaju „skrivene“ slojeve i čine jezgro rada. Korišćenje algoritama dubokog učenja u okviru Python programskog jezika omogućeno je bibliotekom Keras koja čini omotač za biblioteku TensorFlow.

Pre početka rada sa Keras bibliotekom, izvršen je osnovni test jednim od klasifikatora iz SKLearn bibliteke koja predstavlja rad višeslojnog perceptrona (eng. *Multi-Layer Perceptron*). Predstavlja jednostavnu implementaciju potpuno povezane mreže gde je moguće definisati dimenzije skrivenih slojeva, stopu učenja, rano zaustavljanje, aktivacione funkcije slojeva kao i optimizator. Dimenzije mreže određuju broj slojeva i broj perceptrona u svakom sloju. Stopa učenja predstavlja brzinu promene težina veza nakon svake iteracije. Previše velika stopa učenja utiče na lošije fino podešavanje mreže (vrši se prevelika korekcija težina), dok premala stopa učenja utiče na lošije konvergiranje mreže. Rano zaustavljanje predstavlja mogućnost mreže da prekine ranije sa radom na osnovu posmatranja određene vrednosti (npr. preciznost). Pogodno je u slučajevima kada dolazi do zastoja u promeni promenljivih ili promene u lošem pravcu (npr. ista preciznost u više epoha ili smanjenje). Aktivaciona funkcija je funkcija koja se izvršava na svakom sloju za dobijanje ulaza za sledeći sloj, dok je optimizator zadužen za podešavanja težina.

Na slici 32 prikazani su rezultati rada višeslojnog perceptrona. Rezultati su veoma bliski onima dobijenim algoritmima mašinskog učenja. Korišćena su sledeća podešavanja:

- **solver: adam** – predstavlja najbolji optimizator za rad sa velikim brojem podataka koji koristi stohastički gradijent opadanja;
- **alpha: 0.00001** – vrednost regularizacije kod L2 (lasso);
- **learning_rate: adaptive** – održava konstantno učenje dok god se gubitak na slojevima smanjuje (dok se smanjuje greška u treniranju). Ukoliko se između iteracija ne poboljšava gubitak ili validacioni rezultat za određenu vrednost (ukoliko je rano zaustavljanje omogućeno), trenutna stopa učenja se deli sa 5 (radi boljeg podešavanja);
- **early_stopping: True** – izdvaja 10% iz trening seta koji služi za validaciju. Ukoliko se rezultat ne poboljša u narednih 10 (osnovna vrednost) epoha, vrši se prekid obučavanja.
- **activation: relu** – korišćenje ReLU aktivacione funkcije koja daje najbolje rezultate. Definisana je funkcijom f(x) = max(0, x). Ukoliko je vrednost ispod nule zamenjuje se nulom, do se sve pozitivne vrednosti samo prosleđuju dalje;
- **hidden_layer_sizes: (512, 512)** – mreža sa 512 slojeva gde svaki sloj čine 512 potpuno povezanih neurona;
- **max_iter: 200** – osnovno podešavanje za broj iteracija (epoha).

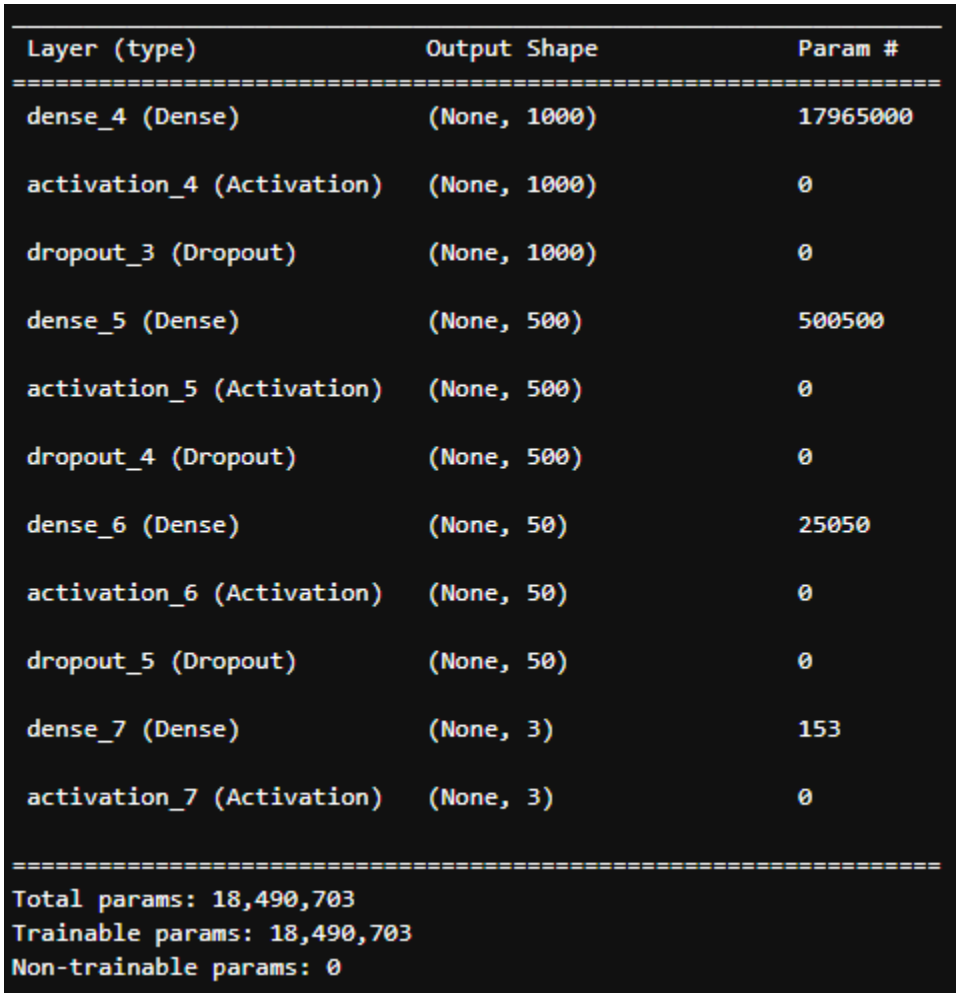
```
Multi Layer Perceptron BoW:   0.7019650655021834
Multi Layer Perceptron TfIdf: 0.6883187772925764
```

Slika 32. Rezultati rada višeslojnog perceptrona

Rad u okviru Keras biblioteke sastoji se u definisanju tipa mreže, definisanju slojeva, kompajliranju i na kraju treniranju mreže nad podacima. Pre početka rada definisana je povratna funkcija (eng. *Callback*) koja posmatra validacionu preciznost i zaustavlja obučavanje mreže nakon 3 iteracije bez poboljšanja, uz povratak najboljih težina. Takođe, potrebno je prekodirati vrednosti klasa u kategorijske gde se dobijaju tri kolone sa jedinicom na mestu pripadnosti klasi.

Model koji se koristi je sekvencijalni, koji predstavlja osnovni model u okviru biblioteke (jedan ulaz/jedan izlaz neurona). Prvi sloj predstavlja potpuno povezani ulazni sloj sa 1000 izlaza i odgovarajućim brojem ulaza u zavisnosti od toga da li se koristi BoW ili TfIdf reprezentacija podataka. Za aktivacionu funkciju sloja iskorišćena je ReLU funkcija, koja se koristi i u svim ostalim slojevima izuzev izlaznog. Sledi definisanje odbacivanja, čime se odbacuje 50% svih težina između slojeva. Cilj je sprečiti prenaučenost mreže, tj. slučaj kada preciznost raste nad trening podacima, ali opada na test podacima. U nastavku sledi potpuno povezani sloj od 500 neurona, odbacivanje od 50%, potpuno povezani sloj od 50, odbacivanje od 50%. Na kraju izlazni sloj sadrzi potpuno povezani sloj sa 3 ulaza i izlazom. Aktivaciona funkcija koja se koristi je „softmax“. Izlaz funkcije jeste realni broj koji predstavlja šansu da predviđanje pripada određenoj klasi (mreža nije u potpunosti sigurna kojoj klasi pripada, već daje predviđanje sa određenom sigurnošću).

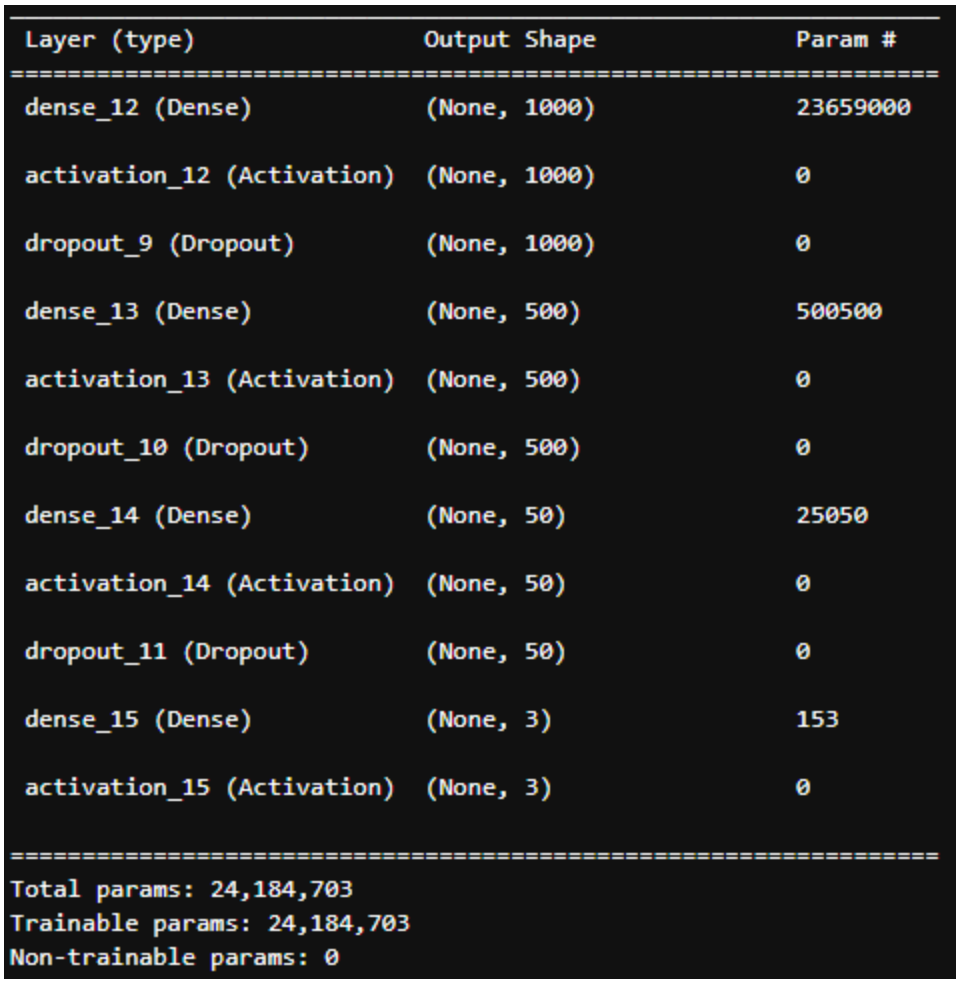
Kompajliranje klase se ostvaruje pozivom funkcije gde se definišu ostali parametri kao što su funkcija gubitka i optimizator. Za funkciju gubitka se koristi kategorijska krosentropija koja određuje kolika je razlika između predviđene i stvarne vrednosti, ukoliko se radi o više kategorija. Manja razlika znači manju funkciju gubitka i samim tim bolje rezultate. Za optimizator se koristi Adam, dok je nadgledana metrika obučavanja preciznost. Izvršeno je obučavanje za BoW i TfIdf reprezentacije. Izgled definisane mreže za BoW reprezentaciju podataka prikazan je na slici 33, dok je na slici 34 dat prikaz mreže za TfIdf reprezentaciju podataka.



Layer (type)	Output Shape	Param #
dense_4 (Dense)	(None, 1000)	17965000
activation_4 (Activation)	(None, 1000)	0
dropout_3 (Dropout)	(None, 1000)	0
dense_5 (Dense)	(None, 500)	500500
activation_5 (Activation)	(None, 500)	0
dropout_4 (Dropout)	(None, 500)	0
dense_6 (Dense)	(None, 50)	25050
activation_6 (Activation)	(None, 50)	0
dropout_5 (Dropout)	(None, 50)	0
dense_7 (Dense)	(None, 3)	153
activation_7 (Activation)	(None, 3)	0

Total params: 18,490,703
Trainable params: 18,490,703
Non-trainable params: 0

Slika 33. Izgled duboke mreže kod BoW reprezentacije podataka



Slika 34. Izgled duboke mreže kod TfIdf reprezentacije podataka

Obučavanje samih mreža izvršeno je kroz 20 iteracija sa podešavanjem težina na svakih 100 uzoraka. Korišćena je funkcija ranog zaustavljanja uz validacioni set od 10% uzoraka. Nakon obučavanja mreže izvršeno je predviđanje klasa nad test podacima i poređenje sa stvarnim vrednostima. Na slici 35 i 36 date su preciznost i metrike obučavanja mreža za BoW i TfIdf reprezentacije podataka. Dobijeni rezultati su približni vrednostima iz prethodnih obučavanja (70% za BoW, 69% za TfIdf). Treba naglasiti da se prenaučenost javlja veoma brzo nakon pokretanja rada mreža i zbog toga se mreže zaustavljaju posle svega nekoliko iteracija (5 za BoW, 4 za TfIdf). Ukoliko mreže nastave sa radom, kroz 20 iteracija, preciznost nad trening podacima postaje gotovo stoprocentna, ali preciznost nad test podacima značajno opada ispod vrednosti sa ranijih iteracija.

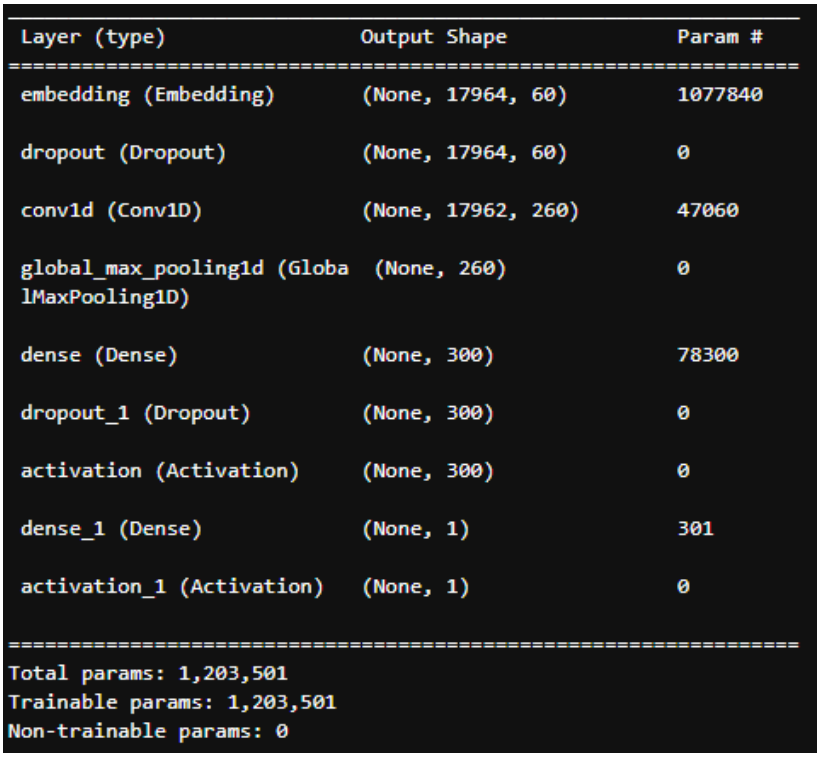
```
Deep Neural Network - Train accuracy: 0.8714519650655022

Deep Neural Network - Test accuracy: 0.7019650655021834

Deep Neural Network - Train Classification Report
              precision    recall  f1-score   support

           0       0.86      0.85      0.85      8815
           1       0.88      0.91      0.90      6863
           2       0.87      0.86      0.87      6306

    accuracy                           0.87     21984
   macro avg       0.87      0.87      0.87     21984
weighted avg       0.87      0.87      0.87     21984


Deep Neural Network - Test Classification Report
              precision    recall  f1-score   support

           0       0.67      0.69      0.68      2302
           1       0.75      0.75      0.75      1719
           2       0.69      0.66      0.67      1475

    accuracy                           0.70      5496
   macro avg       0.70      0.70      0.70      5496
weighted avg       0.70      0.70      0.70      5496
```

Slika 35. Rezultati rada duboke mreže nad BoW reprezentacijom podataka

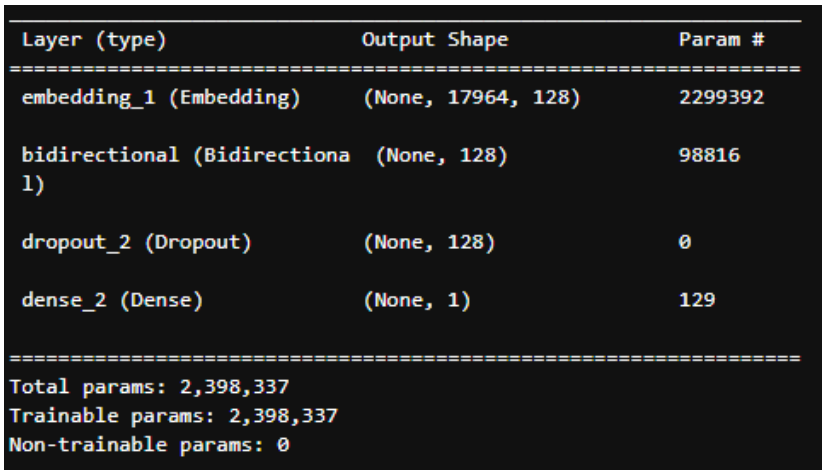
```
Deep Neural Network - Train accuracy: 0.8135462154294032

Deep Neural Network - Test accuracy: 0.6850436681222707

Deep Neural Network - Train Classification Report
              precision    recall  f1-score   support

           0       0.75      0.84      0.79      8815
           1       0.85      0.87      0.86      6863
           2       0.89      0.70      0.79      6306

    accuracy                           0.81     21984
   macro avg       0.83      0.81      0.81     21984
weighted avg       0.82      0.81      0.81     21984


Deep Neural Network - Test Classification Report
              precision    recall  f1-score   support

           0       0.63      0.74      0.68      2302
           1       0.73      0.74      0.74      1719
           2       0.75      0.53      0.62      1475

    accuracy                           0.69      5496
   macro avg       0.70      0.67      0.68      5496
weighted avg       0.69      0.69      0.68      5496
```

Slika 36. Rezultati rada duboke mreže nad TfIdf reprezentacijom podataka

Za potrebe analize sentimenata, moguće je koristiti još nekoliko tipova mreža dubokog učenja, od kojih treba izdvojiti konvolucione mreže i rekurentne mreže. U okviru rada postavljen je kod za primere izgleda ovakvih mreža, međutim, nije bilo moguće ostvariti obučavanje mreža zbog velike potrebe za resursima i značajnog vremena potrebnog za pokretanje. Poređenjem preciznosti obučavanja dobijenih testovima sa onima iz izvornog materijala, predviđa se ostvarivanje poboljšanja u preciznosti od oko 5%, dok bi se štelovanjem mreža mogao ostvariti dobitak u preciznosti i do 10%. Izgledi mreža za BoW reprezentaciju podataka dati su na slikama 37 (konvoluciona) i 38 (rekurentna).



Slika 37. Izgled konvolucione mreže za BoW reprezentaciju podataka



Slika 38. Izgled rekurentne mreže za BoW reprezentaciju podataka

Zaključak

Iako veoma mlada disciplina, analiza sentimenata postigla je veoma brz razvoj tako da nalazi primenu u velikom broju oblasti Internet poslovanja. Brz i lak uvid u veliki broj sentimenata korisnika omogućuje bolje donošenje odluka i blagovremeno otklanjanje problema i nezadovoljstva korisnika. Razvoj discipline dobija dodatno na značaju popularizacijom socijalnih mreža. Korišćenjem analize sentimenata nad socijalnim mrežama ostvaruje se uvid u sentiment najšireg broja korisnika. Glavni cilj je određivanje polariteta sentimenta, tj. da li sadrži pozitivana, negativna ili neutralna osećanja, stavove i mišljenja. Međutim, neformalnost i nestruktuiranost jezika na Internetu i dalje predstavlja velike probleme u uspešnom rešavanju zadatka.

Pristupi u radu menjali su se razvojem discipline. Na početku je za određivanje sentimenata bio zadužen čovek, što je unosilo veliki broj grešaka zbog ličnih sklonosti. Kao pokušaj prevazilaženja uticaja ljudske greške na analizu, nastaju brojni leksikoni koji sadrže pravila za analizu sentimenata kao i vrednosti sentimenata pojedinih reči. Najveći problem leksikona predstavlja statičnost i nepromenljivost. Pravila pisanja i značenje pojedinih reči se veoma brzo menjaju na Internetu. Zbog toga se razvojem moći računara sve više upotrebljavaju algoritmi mašinskog učenja za definisanje i izmenu pravila. Budućnost razvoja discipline sastoji se u korišćenju dubokih mreža koje nalaze sve veću primenu uvećanjem dostupnosti računarskih resursa.

Analiza sentimenata Twitter postova u okviru Python programskog jezika odvija se upotrebom i kombinacijom raznih biblioteka otvorenog koda. Dostupni su svi principi rada – određivanje sentimenata pravilima, mašinsko kao i duboko učenje. Korišćenje pravila iz leksikona ostvaruje se direktnom primenom nad tekstom postova, dok je za potrebe mašinskog i dubokog učenja potrebna reprezentacija reči u obliku brojeva, nakon obrade samog teksta i uklanjanja šuma. Najprostije reprezentacije uključuju označavanje reči jedinicom pri pojavljivanju u tekstu ili određivanje frekvencija pojavljivanja, dok napredniji uključuju predstavljanje reči u obliku vektora, čija blizina određuje sličnost reči. Obučavanjem algoritama mašinskog i dubokog učenja dobija se preciznost u predviđanju sentimenata u granicama od 60% do 70%, što predstavlja solidan rezultat, koji bi se mogao poboljšati finijom obradom podataka, podešavanjem mreža ili kombinovanim pristupom uz korišćenje leksikona.

Literatura

1. https://www.statista.com/statistics/871513/worldwide-data-created/
2. https://www.renolon.com/number-of-tweets-per-day/
3. Knjiga: Sentiment Analysis in Social Networks (Federico Alberto Pozzi, Elisabetta Fersini etc.)
4. https://getthematic.com/sentiment-analysis/
5. https://monkeylearn.com/sentiment-analysis/

Ostali materijal:

- Applied Natural Language Processing with Python - Implementing Machine Learning
- and Deep Learning Algorithms for Natural Language Processing (Taweh Beysolow II)
- Natural Language Processing in Action (Hobson Lane, Cole Howard, Hannes Hapke)
- Natural Language Processing with Python Cookbook - Over 60 recipes to implement text analytics solutions using deep learning principles (Krishna Bhavsar, Naresh Kumar, Pratap Dangeti)
- Natural Language Processing with PyTorch - Build Intelligent Language Applications Using Deep Learning (Delip Rao, Brian McMahan)
- Practical Natural Language Processing - A Comprehensive Guide to Building Real-World NLP Systems (Sowmya Vajjala, Bodhisattwa Majumde, Anuj Gupta, Harshit Surana)
- Text Analytics with Python - A Practitioners Guide to Natural Language Processing (Dipanjan Sarkar)