UDK 004.774.1:316.774]:578.834
DOI 10.7251/ZRSNG2223021L

Uticaj predprocesiranja na tačnost klasifikacije objava na društvenim mrežama o korona virusu

The influence of preprocessing on the accuracy of the classification of posts on social networks about the corona virus

Jelena Lazić, Elektrotehnički fakultet Univerzitet u Beogradu

***Sažetak*—Početkom 2020. godine nastupila je pandemija izazvana virusom korona. Na globalnom nivou uvedene su mjere restrikcija u cilju sprječavanja daljeg širenja virusa, nakon kojih su skoro svi aspekti života svedeni na rad od kuće. Ograničenje komunikacije uživo, dovelo je do povećane aktivnosti korisnika na društvenim mrežama. Analiza objavljenih sadržaja na njima može dati uvid o osjećanjima i stavovima koji preovladavaju među korisnicima. U ovom radu vršena je klasifikacija objava o korona virusu na društvenoj mreži Tviter napisanih na engleskom jeziku. Korištena je javno dostupna baza podataka sa Kaggle platforme. Tvitovi su klasifikovani na osnovu svog sentimentalnog značenja u jednu od pet klasa: ekstremno pozitivni, pozitivni, neutralni, negativni i ekstremno negativni tvitovi. Cilj rada je ispitati na koji način preprocesiranje podataka utiče na tačnost klasifikacije. Korišteni su klasifikator Naivni Bajes, KNN i vještačke neuralne mreže. Rezultati ukazuju na to da način predprocesiranja linkova i tagovanja ne utiče na tačnost klasifikacije, ali način procesiranja heštagova može imati uticaj na tačnost.**

***Ključne riječi* – procesiranje prirodnih jezika; sentimentala analiza tekesta; klasifikacija objava na tviteru**

***Abstract* – At the beginning of 2020, there was a pandemic caused by the corona virus. At the global level, restrictive measures were introduced in order to prevent the further spread of the virus, after which almost all aspects of life were reduced to working from home. The restriction of live communication has led to increased user activity on social networks. Analysis of the published content on them can provide insight into the feelings and attitudes that prevail among users. In this paper, a classification of announcements about the corona virus on the Twitter social network written in English was made. A publicly available database from the Kaggle platform was used. Tweets are classified based on their sentimental meaning into one of five classes: extremely positive, positive, neutral, negative, and extremely negative tweets. The goal of the paper is to examine how data preprocessing affects classification accuracy. The Naive Bayes classifier, KNN, and artificial neural networks were used. The results indicate that the method of preprocessing links and tagging does not affect the classification accuracy, but the method of processing hashtags can have an effect on accuracy.**

***Keywords* – natural language processing; sentimental text analzsis; tweets classification**

I. UVOD

Početkom 2020. godine nastupila je pandemija izazvana virusom korona, koja se eksponencijalnom brzinom proširila na čitav svijet. Na globalnom nivou uvedene su mjere restrikcije u cilju sprječavanja daljeg širenja virusa, nakon kojih su skoro svi aspekti života svedeni na rad od kuće. Društvene djelatnosti za koje je samo nekoliko mjeseci ranije bilo skoro nemoguće zamisliti da se obavljaju samo putem računara, poput obrazovanja ili trgovine, svele su se potpuno na digitalni oblik. Od prvog pojavljivanja virus korona zahvatio je preko 180 zemalja. Širom svijeta prouzrokovani su ogromni gubici u privredi i ekonomiji. Osim velikog broja preminulih i materijalnih gubitaka, pandemija je dovela i do narušavanja kvaliteta društvenog života, što je za posledicu imalo narušavanje mentalnog zdravlja ljudi. Istraživanje o osjećanjima ljudi tokom vanrednih situacija, od suštinskog je značaja za očuvanje mentalnog zdravlja.

Ograničenje komunikacije uživo, dovelo je do povećane aktivnosti korisnika na društvenim mrežama. Milioni ljudi koristili su platforme kao što su Fejsbuk, Tviter, Gugl, Redit, Snepčet i Tik-tok, da bi izrazili svoje stavove i emocije koje su osjećali tokom perioda pandemije. Analiza objavljenih sadržaja na ovim platformama može dati uvid o osjećanjima i stavovima koji preovladavaju među korisnicima ovih platformi. Među pomenutim društvenim mrežama posebno se izdvaja Tviter, jer su na njemu najviše zastupljeni tekstualni sadržaji, a sama mreža osim što važi za društvenu mrežu, važi i za mrežu koja pruža informativne sadržaje. Osim informacija o emocijama i osjećanjima, na osnovu analize tvitova određenih korisnika, moguće je dobiti i procjenu o stepenu informisanosti datih korisnika o određenoj temi.

Oblast nauke u čiji domen spada analiza tekstualnih sadržaja i ekstrakcija njihovog sentimentalnog značenja zove se obrada prirodnih jezika. To je multidiciplinarna oblast, koja se služi metodama i tehnikama vještačke inteligencije i lingvističkim pravilima, da bi proučavala probleme

automatskog prevođenja i razumijevanja prirodnih jezika. U poređenju sa formalnim jezicima, kao što su programski jezici, priordni jezici imaju kompleksniju gramatiku, značajno veće vokabulare, a samo značenje riječi zavisi od konteksta u kome su one upotrebljene. Sve ovo obradu prirodnih jezika čini veoma zahtjevnim problemom.

U ovom radu na osnovu trening skupa podataka obučavani su algoritmi koji za novi tvit, tvit iz testirajućeg skupa podataka, određuju kojoj klasi pripada. Postoji pet klasa, ekstremno pozitivni tvitovi, pozitivni, neutralni, negativni i ekstremno negativni. U svakodnevnoj komunikaciji susrećemo se sa iskazima koji se mogu razvrstati u ovih pet klasa. Prilikom određivanja kakav stav iznosi određen iskaz često obraćamo pažnju na ponašanje, stav, tonalitet i facijalnu ekspresiju govornika. Prilikom analize tekstualnih sadržaja, nemamo te mogućnosti. Dodatno, korisnici Tvitera često koriste neformalne izraze, a u njima mogu biti zastupljeni i ironija i sarkazam koji dodatno otežavaju sentimentalnu analizu [1]. Ne postoji jasna definicija kakav tvit se smatra ekstremno pozitivnim/negativnim ili pozitivnim/negativnim, i razlikovanje ovih klasa težak je zadatak i za ljude.

Postoji mnogo radova koji se bave klasifikacijom tvitova. U zavisnoti od toga na koju temu se odnose napisani tvitovi i koja vrsta klasifikatora se koristi, zavisi i način predobrade podataka. U odnosu na klasifikaciju drugih iskaza, klasifikacija objava na Tviteru razlikuje se po tome što tvitovi sadrže heštagove, linkove, tagovanja i retvit oznake, a način njihovog predprocesiranja razlikuje se od autora do autora. Cilj ovog rada jeste ispitati na koji način predprocesiranje posebnih tviter oznaka, utiče na tačnost klasifikacije. Da bi se otklonio uticaj klasifikatora, korištene su tri različite vrste klasifikatora na istom skupu podataka.

Za realizovanje algoritma korištena je javno dostupna baza podataka sa Kaggle platforme **Error! Reference source not found.**. Tokom predprocesiranja i klasifikacije korištene su tehnike procesiranja prirodnih jezika opisane u knjigama **Error! Reference source not found.** i **Error! Reference source not found.**. Kao dodatna literatura korišteni su rad o sentimentalnoj analizi tvitova o korona virusu [5], rad o klasifikaciji COVID19 tvitova metodama mašinskog učenja **Error! Reference source not found.**, kao i rad o detekciji ironije i sarkazma u tvitovima na engleskom jeziku [1]. Autori rada [5] ističu da su primjetili da su negativni tvitovi češće retvitovani. U istom radu korištena je baza podataka sa 226668 tvitova, raspoređenih u tri klase, pozitivni, neutralni, negativni. Autori smatraju da to nije dovoljan broj primjera za obučavanje klasifikatora. U ovom istraživanju korištena je baza podataka sa manjih brojem tvitova, a klasifikacija je podrazumijevala veći broj klasa, što je moglo uticati na dobijanje manje tačnosti klasifikacije. Analiza značenja tvitova o raznim temama je popularan pristup [7] i najčešće se koristi za proučavanje emocija koje pokazuju korisnici socijalnih medija, a na osnovu kojih se vrše predviđanja rezultata političkih izbora [8], predikcije cijena na berzi [9] ili drugo.

II. DRUŠTVENA MREŽA TVITER

Tviter (Twitter) je mikrobloging platforma, osnovana 21. marta 2006. godine u San Francisku, Kalifornija. Danas je jedna od najrasprostranjenijih društvenih mreža, sa preko 229 miliona registrovanih korisnika širom svijeta. Cilj mreže jeste širenje informacija i mišljenja korisnika putem najčešće tekstualnih objava, tvitova (tweets).

Korisnici Tvitera su tipično anonimni, odnosno mogu izabrati proizvoljno korisničko ime. Nakon što se registruje na Tviter, korisnik bira koje naloge želi da prati. Praćenje korisnika ne mora biti obostrano. Korisnici na sajt postavljaju sadržaj u vidu teksta, linkova i fotografija, a drugi korisnici taj sadržaj mogu ocijeniti (like), komentarisati (comment) ili podijeliti na svom nalogu (retweet). Svakom korisniku na početnoj strani izlaze objave korisnika koje on prati, ili objave sa kojima su interagovali korisnici koje on prati. Osim objava korisnici mogu razmjenjivati i privatne poruke. Na Sl. 1 prikazana je jedna Tviter stranica.

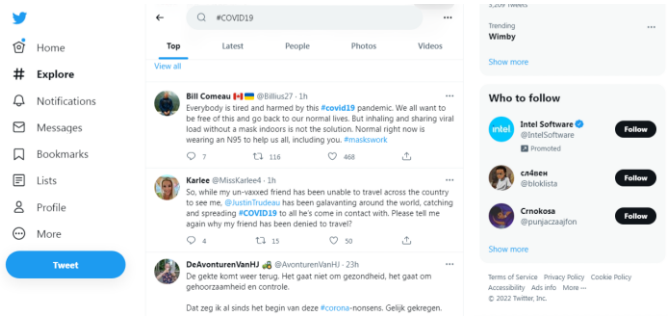
Korisnici Tvitera imaju mogućnost korištenja heštagova, oznaka # uz frazu specifičnog značenja koja može sadržati jednu ili više riječi napisanih bez razmaka, koji služe grupisanju sadržaja. Osim heštagova korisnici mogu pozivati druge korisnike u svojoj objavi upotrebom oznake @ i korisničkog imena, ili koristiti linkove koji vode na druge sajtove upotrebom url adrese sajta.

Analiza sadržaja objavljenih tvitova veoma je korisna za ispitivanje javnog mnenja o određenoj temi. Tokom pandemije virusom korona 2020. godine, zabilježen je značajan rast broja korisnika Tvitera [10]. U ovom periodu zabilježeno je i širenje netačnih informacija o COVID-19 pandemiji na ovoj platformi, zbog čega je Tviter najavio posebno filtriranje objava vezanih za COVID-19 [10].

Kao društvena mreža, Tviter danas zauzima važnu ulogu u svim društvenim, ekonomskim i političkim dešavanjima, zbog čega je i očekivan porast interesovanja za analizu sadržaja objavljenog na Tviteru. Analiza tvitova posebno je zahtjevna jer predstavlja obradu teksta koji sadrži neformalne oblike izražavanja, lokalizme, emotikone, heštagove, linkove, skraćenice ili izraze koji često nemaju semantičko značenje.

III. PROCESIRANJE PODATAKA

Da bi se tvitovi moglli koristiti kao ulazi klasifikatora potrebna je njihova predobrada ili predprocesiranje, koje podrazumijeva nekoliko koraka koji se primjenjuju na svakom pojedinačnom tvitu obučavajućeg i testirajućeg skupa



Slika 1 Prikaz jedne Tviter stranice – na njoj se vide izdvojeni tvitovi sa heštag oznakom COVID19

podataka. U ovom radu ispitivan je utica predobrade na tačnost klasifikacije te je korišteno nekoliko različitih načina predobrade, osnovna obrada, obrada sa linkovima, obrada sa tagovanjima i obrada bez heštagova. Koraci predrocesiranja su:

- Uklanjanje znakova koji ukazuju da tvit nije originalno napisan nego da se radi o dijeljenju već napisanog tvita, retweet oznaka. Ovaj korak karakterističan je za procesiranje tvitova, i korišten je u svim verzijama predobrade.
- Predobrada hiper linkova korištenih u tvitovima. Osnovna obrada podrazumijava uklanjanje hiperlinkova, a obrada sa linkovima potrazumijeva zamjenu linkova u svim tvitovima unaprijed definisanom jedinstvenom riječju.
- Uklanjanje oznake # koja se obično javlja na kraju tvita uz riječi koje predstavljaju referencu na određenu temu, heštag oznaka. U osnovnoj obradi vrši se uklanjanje samo oznake # ne i čitave riječi ispred koje se nalazila oznaka, u obradi bez heštagova vrši se uklanjanje čitave riječi.
- Obrada tagovanja drugih korisnika tvitera, riječi koje počinju znakom @, osnovna obrada podrazumijva potpuno uklanjanje riječi koje počinju oznakom @, obrada sa tagovanjima podrazumijva njihovu zamjenu unaprijed definisanom jedinstvenom riječju.
- Uklanjanje stopwords, riječi koje nemaju značenje. Koristi se na isti način u svim verzijama predobrade. S obzirom da su svi tvitovi napisani na engleskom jeziku, koristi se nltk python biblioteka u kojoj stopwords obuhvataju razne vrste zamjenica kao što su:

I, me, my, myself, we, our, you've, you'll, you'd, your, she's, her, hers, herself...

- Uklanjanje znakova punktuacije. Koristi se na isti način u svim verzijama predobrade. Primjeri znakova punktuacije koje treba otkloniti su:

!"#$%&'()*+,-./:;<=>?@[]^`{|}~...

- Uklanjanje riječi koje u sebi sadrže znakove koji ne pripadaju alfabetu, većina ovih riječi su posledica grešaka tokom prikupljanja podataka. Koristi se na isti način u svim verzijama predobrade.
- Izdvajanje pojedinačnih riječi koje se pojavljuju u tvitu, pri čemu se sva velika slova pretvaraju u mala. U ovom koraku se riječi svode na svoj osnovni oblik, odnosno uklanjaju se nastavci koji odgovaraju množini imenica, glagolskim nastavcima određenih glagolskih vremena, ili gradaciji pridjeva. Tako se glagolima otklanjaju nastavci ing i ed, imenicama nastavak s za množinu, pridjevima nastavci komparativa i superlativa er i est ... Ovaj korak takođe se primjenjuje u svim verzijama predobrade tvitova.

Rezultat predobrade je pretvaranje svakog pojedinačnog tvita u listu riječi, tokena, koje se u njemu pojavljuju. Na kraju, s obzirom da imaju isto značenje, riječ coronavirus zamijenjena je riječju covid. U nastavku je dat primjer originalnog tvita i njemu odgovarajućeg niza riječi nakon predprocesiranja.

Originalni tvit:

Morning everyone have a great and safe day. ??? coronavirus StopPanicBuying BeKind mufc MUFC_Family

Lista izdvojenih riječi tvita:

morn, everyon, great, safe, day, covid, stoppanicbuy, bekind, mufc

Nedostatak poslednjeg koraka algoritma, jeste to što ne postoji mogućnost prepoznavanja tipa riječi, te se dešava da nekada dođe do uklanjanja nastavka iako se zapravo ne radi o nastavku nego o osnovi riječi. Tako se riječ coronavirus svodi na coronaviru, riječ advice na advic ... Ovakve zamjene ne utiču na dalje rezultate obrade i klasifikacije tvitova.

Nakon predobrade, uklonjeni su svi tvitovi čija je lista riječi bila prazna. Preostalo je 41121 obučavajućih tvitova. Isti postupak predobrade izvršen je i na testirajućem skupu. Nakon uklanjanja svih tvitova čija je lista riječi prazna, preostalo je 3795 testirajućih tvitova.

IV. KLASIFIKATORI

A. Naivni Bajes

Za klasifikaciju tvitova koriste se razni modeli. U zavisnosti od toga koji tip klasifikatora se koristi zavisi i način na koji se klasifikator obučava kao i obilježja koja se koriste. Jedan od često korištenih klasifikatora u ovoj oblasti je Naivni Bajes. Ovaj model uzima u obzir aprirone vjerovatnoće pojavljivanja klasa, kao i aposterione vjerovatnoće pojavljivanja riječi u klasama.

Razlog zašto se ovaj klasifikator smatra naivnim je njegova pretpostavka da su sva obilježja uslovno nezavisna jedna od drugog (ukoliko je poznata klasa), što u realnosti nije

tačno. Obilježja su skoro uvijek, manje ili više zavisna jedna od drugih. Prilikom procesiranja tvitova, ova pretpostavka ogleda se u tome, što se zanemaruje redoslijed riječi koje se pojavljuju u tvitu. Naivni Bajes posmatra samo koje riječi postoje u datom tvitu, ne i koja riječ se javlja pored koje riječi, na ovaj način može da se izgubi kontekst u kome su protumačene riječi.

Druga riječ u nazivu klasifikatora, potiče od toga što se za klasifikaciju koristi Bajesova teorema za aposterionu vjerovatnoću. U problemu klasifikacije tvitova o korona virusu postoji pet klasa. Apriorna vjerovatnoća pojavljivanja

	Unnamed: 0	Extremely Positive	Positive	Neutral	Negative	Extremely Negative	Total	1	2	3	4	5
23	covid	6284.0	11372.0	8301.0	9791.0	5430.0	41178.0	0.051606	0.061055	0.088590	0.061050	0.054252
232	price	1052.0	2295.0	1365.0	2747.0	1605.0	9064.0	0.008639	0.012322	0.014568	0.017128	0.016036
67	store	1487.0	2441.0	1592.0	1854.0	837.0	8211.0	0.012212	0.013105	0.016990	0.011560	0.008363
49	supermarket	1186.0	2127.0	1446.0	1958.0	1064.0	7781.0	0.009740	0.011420	0.015432	0.012209	0.010631
33	food	1042.0	1866.0	707.0	2079.0	1598.0	7292.0	0.008557	0.010018	0.007545	0.012963	0.015966
...	...	...	...	...	...	...	...	...	...	...	...	...
4709	businessown	1.0	3.0	2.0	0.0	0.0	6.0	0.000008	0.000016	0.000021	0.000006	0.000009
4707	arrog	0.0	1.0	0.0	2.0	3.0	6.0	0.000008	0.000005	0.000010	0.000012	0.000030
4705	dynata	2.0	3.0	1.0	0.0	0.0	6.0	0.000016	0.000016	0.000011	0.000006	0.000009
4686	havent	1.0	2.0	1.0	1.0	1.0	6.0	0.000008	0.000011	0.000011	0.000006	0.000010
6532	quiroga	6.0	0.0	0.0	0.0	0.0	6.0	0.000049	0.000005	0.000010	0.000006	0.000009

6533 rows × 12 columns

Slika 2 Riječnik riječi trening skupa podataka - izdvojene su sve riječi koje u se pojavljivale u trening skupu podataka, za svaku od njih izdvojeno je koliko puta se pojavila u kojoj klasi i kolike su njoj odgovarajuće aposterione vjerovatnoće klasa

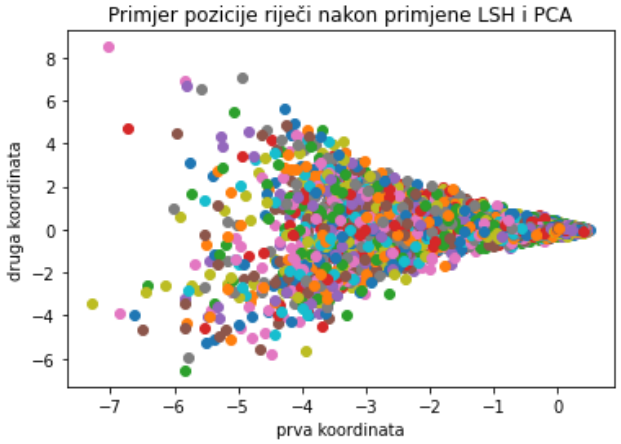
svake klase računa se na osnovu frekvencije klase u trening skupu podataka. Aposteriona vjerovatnoća pojavljivanja riječi u određenoj klasi računa se na osnovu frekvencije date riječi u odnosu na ukupan broj riječi koje su se pojavile u svim tvitovima date klase. Riječnik predstavlja skup svih riječi trening skupa podataka sa njihovim aposterionim vjerovatnoćama za svaku klasu. Ukoliko se neka riječ nije pojavila u nekoj klasi koristi se metod Laplasovog izglađivanja. Riječi koje su se pojavile manje od 6 puta u tvitovima testirajućeg skupa smatraju se rijetkim riječima i ne koriste se u daljoj obradi.

Klasifikacija tvitova vrši se pomoću formiranog rječnika. Na osnovu Bajesove formule za aposterionu vjerovatnoću računaju se vjerovatnoće da dati tvit pripada određenoj klasi. Tvit se svrstava u klasu čija je vjerovatnoća najveća.

A. KNN klasifikator

K najbližih susjeda (k nearest neighbors, KNN), spada u tip negenralizovanih klasifikatora, što znači da nema mougćnost učenja parametara klasa, ne postoji interni model klasifikatora, nego se odluka donosi pamćenjem trening podataka i upoređivanjem test primjera sa njima. Konačna odluka klasifikatora donosi se većinskim glasanjem k najbližih susjeda, odnosno k najsličnijih primjera trening skupa podataka. Klasifikator je implementiran korištenjem ugrađene funkcije sklearn python biblioteke. Parametar k je hiper-parametar klasifikatora i može značajno da utiče na rezultate klasifikacije. U ovom radu vrijednost parametra k određena je metodom krosvalidacije.

Da bi se izvršila klasifikacija tvitova pomoću KNN klasifikatora, potrebno je predstaviti tvitove u vektorskom prostoru riječi. Osnovna ideja primjene vektorskih prostora jeste da se riječima dodijele numeričke vrijednosti, odnosno nizovi brojeva, takvi da je na osnovu njih moguće odrediti u kakvim odnosima su riječi međusobno. Za razliku od rječnika, u kome se svaka riječ posmatra nezavisno od svih ostalih riječi, vektorski prostor pokušava da modeluje i međusobne veze riječi. Što se riječi pojavljuju češće u istoj rečenici, to će biti bliže u vektorskom prostoru. Ispostavlja se takođe, da se riječi koje su sinonimi javljaju blizu jedna drugoj unutar vektorskog prostora, ali i da riječi imaju tendenciju da budu bliže riječima iste vrste, imenice se međusobno grupišu, kao i pridjevi, predlozi i glagoli.



Slika 3 Prikaz riječi trening skupa u vektorskom prostoru - nakon predstavljanja riječi u vektorskom prostoru, LSH (local sensitive hashing) izvršena je PCA (principal component analysis) redukcija dimenzija da bi e riječi vizuelno predstavile

U python biblioteci gensim postoji ugrađen model Word2Vec, koji računa vektorsku reprezentaciju riječi na zadatoj dužini. Na Sl. 3 prikazane su riječi trening skup podataka predstavljene u vektorskom prostoru.

Nakon izdvajanja vektora riječi, vrši se računanje vektora koji odgovaraju dokumentima, sumiranjem ili usrednjavanjem vektora svih riječi koje se pojavljuju u dokumentu (tvitu). U ovom radu korišteno je sumiranje. Sličnost dva dokumenta određuje se računanjem distance njima odgovarajućih vektora. Za distancu se najčešće koriste Euklidska i kosinusna distanca.

Kao obilježja dokumenata za KNN klasifikaciju korištene su vrijednosti vektora dokumenta u prostoru 100 dimenzija, a parametar k je imao vrijednost 20. Kao mjera međusobne udaljenosti odbiraka koristi se kosinusna distanca. Tvitu se dodijeljuje ona klasa koja je najzastupljenija među njegovih k susjeda u vektorskom prostoru.

B. Vještačke nerualne mreže

Pored Naivnog Bajesa i KNN klasifikatora, za klasifikaciju tvitova korištene su i vještačke neuralne mreže. Da bi se izvršila klasifikacija tvitova pomoću neuralnih mreža potrebna je njihova posebna obrada tako da se dati tvitovi mogu dovesti kao ulazi mreža. Nakon, predprocesiranja podataka, tvitovi su svedeni na oblik u kome se mogu predstaviti kao liste normalizovanih riječi, koje su se pojavile u njima. Sledeći korak u pripremi tvitova za klasifikaciju pomoću neuralnih

mreža jeste tokenizacija riječi, koja podrazumijeva numerisanje riječi. Svakoj riječi dodijeli se jedinstven broj, a tvit je niz brojeva riječi koje su se u njemu pojavile. Da bi svi tvitovi bili iste dužine, nizovi kraći od najdužeg niza se dopunjavaju nulama. Za pretvaranje tvitova u sekvence tokena korištena je funkcija keras python biblioteke.

Nizovi tokena koriste se kao ulazi neuralne mreže, u kojoj je prvi sloj embeding sloj, koji računa embeding za dati niz tokena, odnosno za svaki token računa vektorsku reprezentaciju na zadatom broju dimenzija. U ovom radu korišteni su vektori dimenzije 100.

```
train.iloc[5]['TweetTokens']

['cashier', 'groceri', 'store', 'share', 'insight', 'covid', 'prove', 'credibl', 'comment', 'civic', 'class', 'know', 'talk']

padded_sequence_train[5]

array([ 548,    7,    4,  157,  506,    2, 1404, 4581, 1248, 5959,  959,
         65,  320,    0,    0,    0,    0,    0,    0,    0,    0,    0,
          0,    0,    0,    0,    0,    0,    0,    0,    0,    0,    0,
          0,    0,    0,    0,    0,    0], dtype=int32)
```

Slika 4 Prikaz tvita nakon tokenizacije

Korištena je mreža čija arhitektura je prikazana na Sl. 5. Pored svakog sloja naznačene su dimenzije ulaznih i izlaznih podataka, pri čemu None vrijednost odgovara broju trening i test primjera. Regularizacioni parametar dropout sloja je 0.2, aktivaciona funkcija predposlednjeg sloja je rely, a poslednjeg sloja softmax.

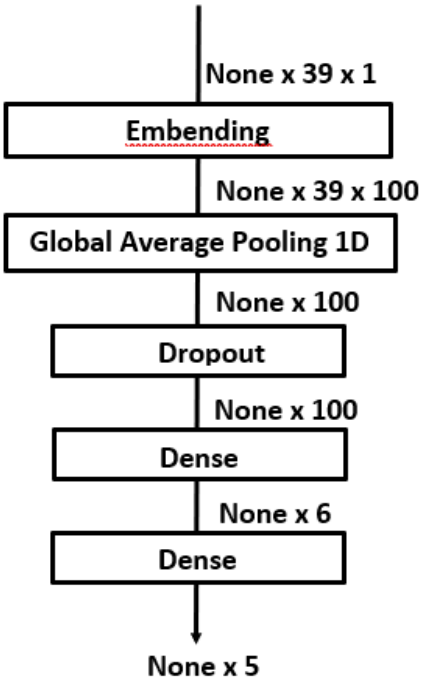
Za obučavanje korištena je categorical crossentropy funkcija gubitka, Adam optimizator, kao metrika korištena je tačnost klasifikacije. Za sprječavanje preobučavanja korišteno je rano zaustavljanje, pri čemu se obučavanje zaustavlja ukoliko se četiri puta za redom desi opadanje tačnosti klasifikacije na validacionom skupu. Kao validacioni skup korišten je test skup podataka, a za realizovanje ranog zaustavljanja korištena je funkcija keras biblioteke. Ukoliko dođe do ranog zaustavljanja, parametri mreže se vraćaju na najbolje postignute rezultate tokom obučavanja. Obučavanje je vršeno tokom 10 epoha.

V. REZULTATI

U trening skupu podataka linkovi su upotrebljeni ukupno 23982, a broj tagovanja drugih korisnika bio je 17575. U tabeli prikazani su tačnost klasifikacije za trening i test skup korištenjem tri različita klasifikatora. Osnovna obrada podrazumijevala je potpuno uklanjanje linkova, uklanjanje tagovanih osoba i otklanjanje oznake # ispred heštaga ali zadržavanje samog heštaga. Ostale obrade se od osnovne razlikuju samo po načinu procesiranja jednog elementa. Obrada sa linkovima podrazumijeva da se svaki link zamjeni jedinstvenom riječju, čime se uvodi razlikovanje tvitova sa i bez upotrebljenih linkova. Obrada sa tagovanjima koristi zamjenu takovanih korisnika specijalnom riječju, a obrada bez heštagova podrazumijeva potpuno uklanjanje heštaga, a ne samao oznake # kao što je to bio slučaj u osnovnoj obradi. Tačnost dobijena pomoću neuralne mreže najveća je tačnost do sada postignuta za dati skup podataka.

VI. ZAKLJUČAK

U ovom radu ispitivan je uticaj načina predprocesiranja na tačnost klasifikacije tvitova o korona virusu na osnovu njihovog sentimenta. Da bi se otklonio uticaj klasifikatora, korištena su tri različita načina klasifikacije, Naivnim Bajesom, KNN klasifikatorom i vještačkom neuralnom mrežom. Rezultati istraživanja ukazuju na to da način predprocesiranja linkova i tagovanja ne utiče na rezultate klasifikacije, ali način procesiranja heštagova može imati uticaj na tačnost. Ovakav rezultat je i očekiva, s obzirom da jedino heštagovi imaju sentimentalno značenje. Najveći uticaj načina obrade heštagova na tačnost klasifikacije je korištejem neuralne mreže. U nastavku rada na ovom projektu trebalo bi razmotriti načine predprocesiranja linkova tako da se iz njih izdvoji dodatna informacija na koju



Slika 5 Arhitektura korištene vještačke neuralne mreže

stranicu ili kakv tip stranice link vodi, kao i poseban način obrade heštagova tako da se u njima vrši razdvajanje na pojedinačne riječi.

LITERATURA

[1] J. Dimovska, M. Angelovska, D. Gjorgjevikj, Gj. Madjarov (2018). "Sarcasm and irony detection in english tweets", 10th International Conference, ICT Innovations, Macedonia

[2] Baza podataka preuzeta sa: https://www.kaggle.com/datasets/datatattle/covid-19-nlp-text-classification Pristupano: 10. Oktobra 2022.

[3] D. Jurafsky, H.J. Martin (2008). Speech and Language Processing, Prentice Hall

[4] S. J. Russell, P. Norvig (2010). Artificial Intelligence: A modern approach

[5] K. Chakraborty, S. Bhatia, S. Bhattacharyya, J. Platos, R. Bag, A.E. Hassanien (2020). "Sentiment analysis of COVID-19 tweets by deeplearning classifier - a study to show how popularity is affecting accuracy in social media" . Elsevier, Applied Soft Computing Journal, vol. 97.

[6] A. Mondal, S. K. Mahata, M. Dey, D. Das (2021). "Classification of COVID9 tweets using machine learning approaches", Proceedings of the Sixth Social Media Mining for Health Workshop 2012, Association for Computational Linguistics

[7] B. Lui (2012). "Sentiment analysis and opinion mining", Syntesis lecutres on human language technologies, vol. 5, number 1.

[8] J. E. Chung, E. Mustafaraj (2011). "Can collective sentiment expressed on twitter predict political elections?" , AAAI, vol.11

[9] M. Makrehchi, S. Shah, W. Liao (2013). "Stock prediciton using event-based sentiment analysis", IEEE/WIC/ACM International Joint Conferences on Web Intelignece (WI) and Inteligence Agent Technologies (IAT), online

[10] Wikipedia: Twitter https://en.wikipedia.org/wiki/Twitter Pristupano: 10. Oktobra 2022.

Tabela 1 Prikaz rezultata - tačnost klasifikacije za trening i test skup , korištenjem tri različita klasifikatora, za različite načine predprocesiranja, tačnost je prikazana u procentima

	Naivni Bajes		KNN		ANN	
	Trening	test	trening	Test	trening	test
Osnovna obrada	61,95	47,83	46,53	36,65	82,97	66,06
Sa linkovima	61,91	47,95	46.50	36,10	79,78	67,36
Sa tagovanjima	61,85	47,85	46,30	36,45	80,76	65,10
Bez heštagova	61,87	49,21	47,00	36,53	82,67	68,34