Klasifikacija tvitova prema sentimentu korišćenjem naivnog Bajesovog klasifikatora i metode potpornih vektora

Student: Teodora Mihajlov, projekat iz predmeta: *Analiza i vizualizacija podataka*, predmetni profesor: Dragan O. Đurić; master studije Računarstvo u društvenim naukama, Univerzitet u Beogradu, Jul 2022

Uvod

U okviru ovog projekta poredili smo performans multinominalne logističke regresije i metode potpornih vektora (*Support Vector Machine - SVM*) u klasfikaciji tvitova prema sentimentu. Na samom početku, opisaćemo ulazne podatke, daćemo pregled statistike varijabli, a zatim ćemo opisati korpus i način njegove pripreme, te prikazati statistiku teksta. Na kraju ćemo dati opis i evaluaciju dva testirana modela, te ih uporediti. Projekat ima za cilj da utvrdi koji je od dva navedena modela uspešniji pri klasifikaciji tvitova, kao i da li se tvitovi uspešnije klasifikuju na osnovu matrice termina i dokumenata (*Term-Document Matrix, TDM*) ili na osnovu relativnih frekvencija termina u tekstu (*termfrequency - inverse document frequency, tfidf*).

Za projekat je korišćen programski jezik R (verzija 4.2.1) u okruženju RStudio. Biblioteke koje su korišćene učitane su na samom početku koda. Pre započinjanja rada isključili smo naučnu notaciju korišćenjem funkcije scipen = 999, radi lakše interpretacije rezultata. Vualizacije su rađene korišćenjem bibliteka ggplot2, ggeasy, i wordcloud.

Opis ulaznih podataka

Baza podataka koja sarži tvitove o korona virusu u periodu od 2. marta do 14. aprila 2020. godine i njihov sentiment preuzeta je sa sajta Keggle (web1). Bazu podataka uvezli smo korišćenjem funkcije read.csv u okvir R-a.

Originalni skup podataka sastoji se od 8 varijabli, sedam nezavisnih i jedne zavisne, i 44, 995 opservacija od čega smo zadržali četiri varijable, tri nezavisne - UserName (ID), TweetAt (datum tvita), Location (lokaciju sa koje je tvitovano), OriginalTweet (tekst tvita), i zavisnu varijablu Sentiment (sentiment tvita), te 5000 opservacija.

Kako smo zadržali samo oko 10% opservacija, kategoriju Sentiment smo sa originalnog skora od 5 (Extremely Negative, Negative, Neutral, Positive, Extremely Positive) smanjili na 3 -

Negative, Neutral, Positive, pri čemu smo Extremely Negative zamenili sa Negative, a Extremely Positive sa Positive. Takođe smo i numeričku varijablu za sentiment, SentimentNum - skalu od -1 do 1, kako bismo videli koji je srednji sentiment tvitova u korpusu. Takođe smo dodali varijablu sa dužinom tvitova, TweetLen. Baza podataka spremna za rad prikazana je na Slika 1.

	UserName	Location	TweetAt	OriginalTweet	Sentiment	SentimentNum	TweetLen
1	1	NYC	2020-03-02	TRENDING: New Yorkers encounter empty supermarket shel...	Negative	-1	228
2	2	Seattle, WA	2020-03-02	When I couldn't find hand sanitizer at Fred Meyer, I turned t...	Positive	1	193
3	3		2020-03-02	Find out how you can protect yourself and loved ones from ...	Positive	1	73
4	4	Chicagoland	2020-03-02	#Panic buying hits #NewYork City as anxious shoppers stock...	Negative	-1	308
5	5	Melbourne, Victoria	2020-03-03	#toiletpaper #dunnypaper #coronavirus #coronavirusaustral...	Neutral	0	252
6	6	Los Angeles	2020-03-03	Do you remember the last time you paid $2.99 a gallon for r...	Neutral	0	205
7	7		2020-03-03	Voting in the age of #coronavirus = hand sanitizer ? #Super...	Positive	1	90
8	8	Geneva, Switzerland	2020-03-03	@DrTedros "We canÂ't stop #COVID19 without protecting #...	Neutral	0	213
9	9		2020-03-04	HI TWITTER! I am a pharmacist. I sell hand sanitizer for a livi...	Negative	-1	280
10	10	Dublin, Ireland	2020-03-04	Anyone been in a supermarket over the last few days? Went...	Positive	1	239
11	11	Boksburg, South Africa	2020-03-04	Best quality couches at unbelievably low prices available to ...	Positive	1	215
12	12	New Delhi	2020-03-04	Beware of counterfeits trying to sell fake masks at cheap pri...	Negative	-1	237
13	13	USA, PA	2020-03-04	Panic food buying in Germany due to #coronavirus has beg...	Negative	-1	261
14	14		2020-03-04	#Covid_19 Went to the Grocery Store, turns out all cleaning ...	Positive	1	278
15	15	Washington, DC	2020-03-04	While we were busy watching election returns and bracing f...	Positive	1	197
16	16	Bengaluru	2020-03-04	#AirSewa @flyspicejet is not providing #webcheckin custom...	Negative	-1	280
17	17	Mumbai	2020-03-05	What Precautionary measures have you all taken in your res...	Positive	1	242
18	18	Toronto, Ontario	2020-03-05	When youÂ're stockpiling food & other supplies, buy e...	Neutral	0	189
19	19		2020-03-05	That's about a week from now. A bit optimistic. Probably it ...	Positive	1	238

Slika 1 - *Baza podataka spremna za rad*

Deskriptivna i inferencijalna statistika

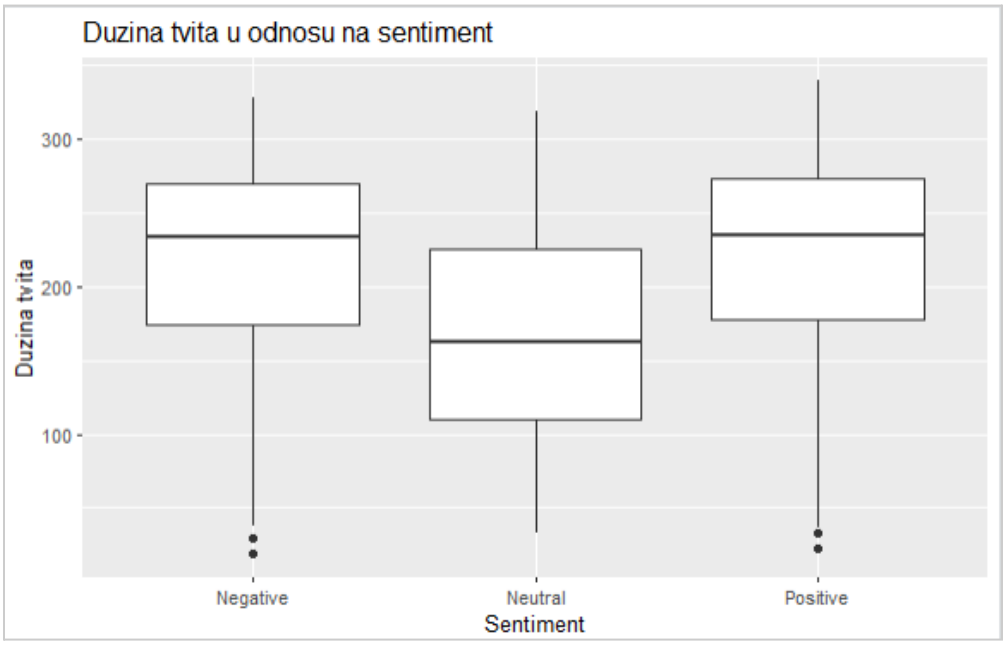
Statističku analizu baze podataka započeli smo deskriptivnom numeričkih varijabli SentimentNum i TweetLen, pozivanjem funkcije describe() u okviru biblioteke psych. R je u analizu uključio i varijablu UserName, ali pošto je ova varijabla ID i nije statistički značajna, rezultate analize ćemo u Tabela 1 prikazati samo za SentimentNum i TweetLen.

Naziv varijable	Aritmetička sredina	Standardna devijacija	Minimum	Maksimum	Raspon
SentimentNum	-0.02	0.91	-1	1	2
TweetLen	210.75	65.23	19	339	320

Tabela 1 - *Deskriprivna statistika numeričkih varijabli*

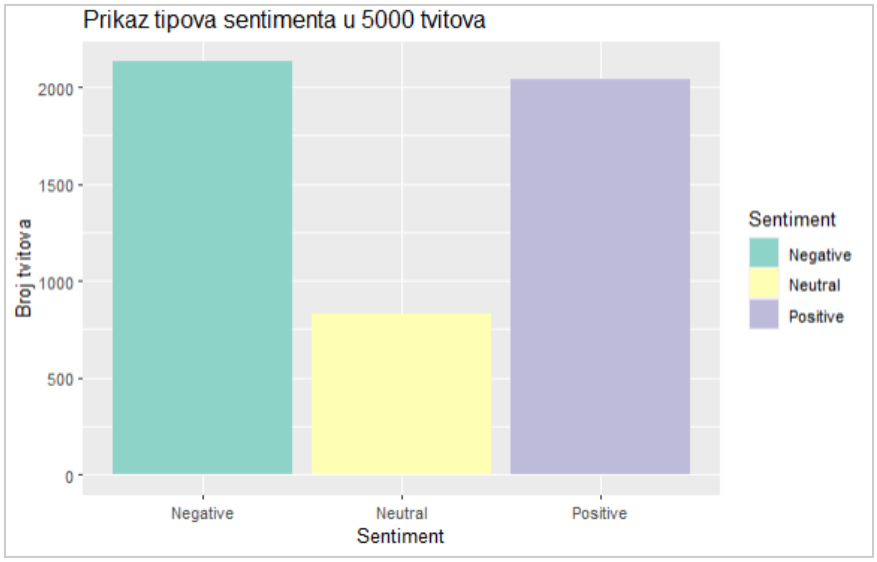
Takođe smo hteli da proverimo da li dužina tvitova zavisi od sentimenta. Za proveru normalnosti distribucije varijable TweetLen korišćen je Shapiro-Wilk test, pozivanjem funckije shapiro.test(). Na osnovu dobijenih rezultata $W = 0.950$, $p\text{-value} < 0.000$, zaključujemo da varijabla TweetLen nema normalnu raspodelu, pa za proveru zavisnosti dužine tvita od sentimenta koristimo neparametarski Kruskal-Wallis test ((kruskal.test)). Na osnovu rezultata Kruskal-Wallis testa

chi-squared = 364.57, df = 2, p-value < 0.000, zaključujemo da dužina tvita ne zavisi od grupe sentimenta, i da su uzorci iz iste populacije. Rezultate smo prikazali boxlot-om (Slika 2), na osnovu kojeg možemo videti da su tvitovi u grupi Neutral nešto kraći od tvitova u grupama Positive i Negative. Ipak, kao što ćemo videti, kako Neutral kategorije u bazi ima znatno manje nego Positive i Negative, dužina ovih tvitova nije dovela do značajnih razlika.



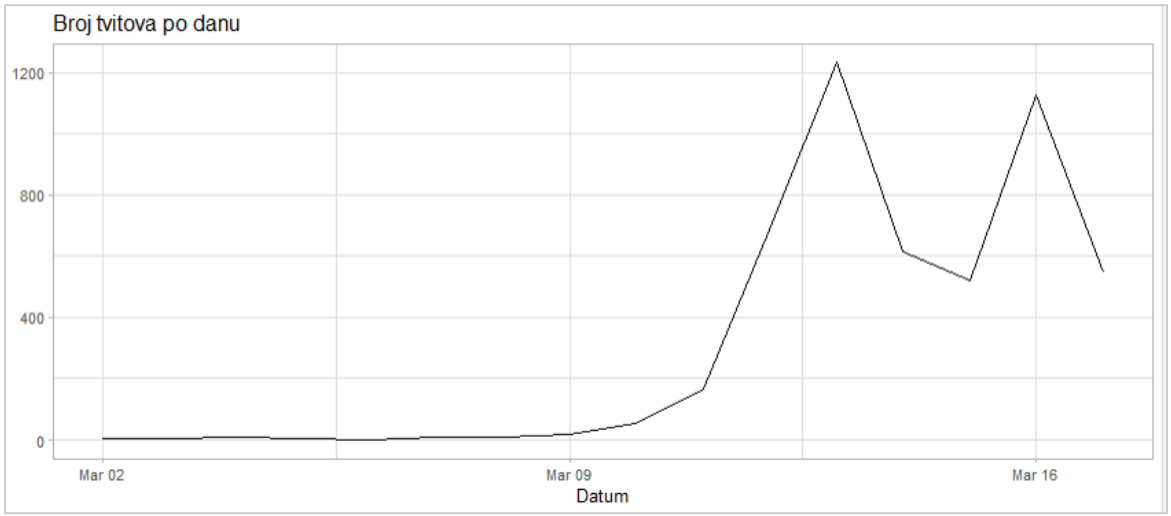
Slika 2 - *Dužina tvitova u odnosu na sentiment*

Broj tvitova u kateogirijama Positive (2040) i Negative (2131) gotovo je jednak, dok je kateogorija Neutral (829) nešto manje zastupljena. S obzirom na to da je 5000 nasumično uzeto iz celokupne baze od oko 49 000 tvitova, možemo reći da je baza podataka relativno dobro izbalansirana.



Slika 3 - *Odnos grupa sentimenta u bazi podataka*

Osim ovoga, prikazali smo i broj tvitova po danu na osnovu varijable TweetAt. Varijablu smo prvo pretvorili u datum sa formatom dan-mesec-godina, a zatim u data frame, kako bismo je vizualizovali. Kao što možemo videti na Slika 4, najviše je tvitovano sredinom marta. Ovo možemo objasniti time što je sredinom marta 2020. godine pandemija korona virkusa zahvatila većinu zemalja.



Slika 4 - *Broj tvitova po danu*

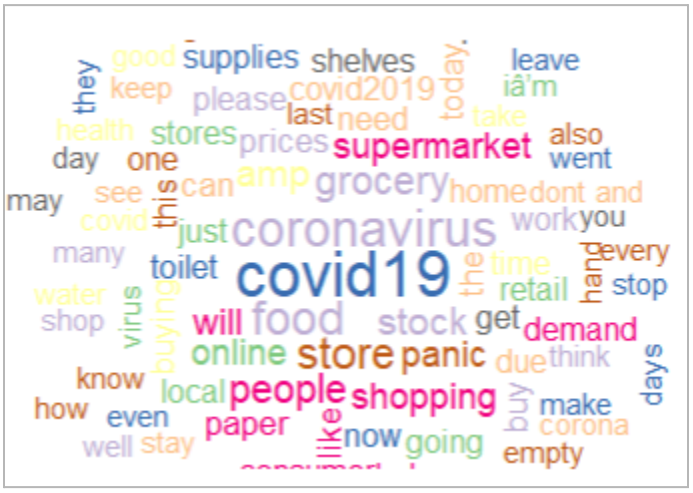
Korišćenjem bibloteke wordcloud, vizualizovali smo varijablu Location (Slika 5), koja sadrži lokacije sa kojih je tvitovano. Najzastupljenije lokacije su Engleska, i različite države u okviru Sjedinjenih Američkih Država. S obzirom na to da lokacije na tviteru imaju slobodan unos i nisu zasnovane na GPS lokaciji, vrednosti u ovoj varijabli treba uzeti sa rezervom.

Slika 5 - *Lokacije sa kojih je tvitovano*

Statistika varijable OriginalTweet biće predstavljena zasebno, u okviru dela koji opisuje pripremu korpusa i sam korpus (tvitove) na kojima su modeli obučavani.

Priprema korpusa

Kako bismo pripremili tvitove za klasifikaciju, razdvojili smo bazu podataka na trening i test set podataka, sa odnosom 80-20, respektivno. Zatim smo kreirali korpus trening seta tvitova, i korpus test seta tvitova, korišćenjem funckije corpus(). Pomoću wordcloud-a vizualizovane su najfrekventnije reči u terning korpusu, kako bismo dobili opštu predstavu o temama u korpusu, ali i o tome koje dodatne stop reči treba ukloniti. Prilikom kreiranja wordcloud-a iskoristili smo funckiju set.seed kako se reči ne bi ponavljale pri svakom kreiranju. Minimalna postavljena frekvencija reči je 1. Kao što možemo videti iz oblaka reči (*word cloud*), najćešće teme o kojima je govoreno su hrana i namirnice (*food, groceries*), kupovina (*store, online, market, price*), ali pominju se i reči kao što su *govern, spread, demand, sick, protect, mask*, *outbreak, panic*, što je leksika koja se često javlja kada se govori o korona virusu.



Slika 6 - *Wordcloud train korpusa*

Nakon kreiranja korpusa, napravili smo funckiju za tokenizaciju i normalizaciju teksta, korišćenjem pipeline operatora %>%. Pri normalizaciji tvitova, uklonili smo velika slova, zatim brojeve, znakove interpunkcije, simbole, linkove, stop reči engleskog jezika korišćenjem liste stop reči "SMART", kao i višak razmaka. Pri određivanju dodatnih stop reči pomogao nam je prethodno kreirani oblak reči (*world cloud*). Kako se radi o tvitovima o korona virusu, neke od dodatnih stopreči bile su *covid, coronavirus, covid_19, covid-19*, a zatim i *https*, kako bismo otklonili ostatike linkova, i *amp*, skraćenica za *Amplifier* koja se koristi u okviru platforme Twitter. Normalizacija je urađena na isti način i za test i za trening skup. Za tokenizaciju i normalizaciju teksta korišćena je bibloteka quanteda.

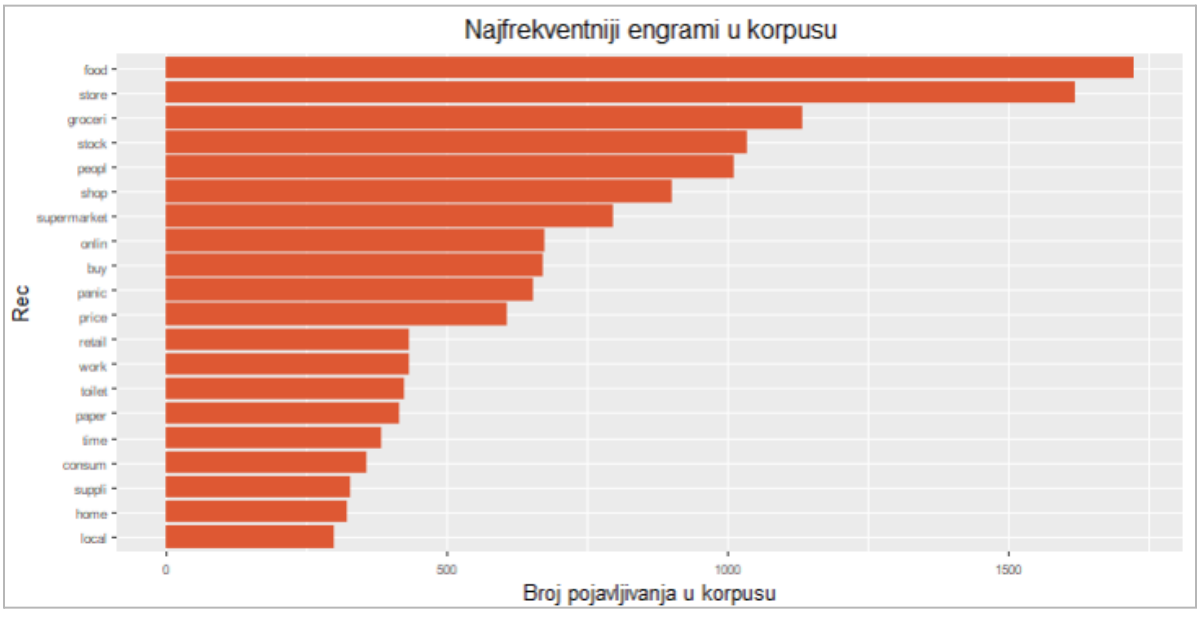
Od dobijenih trening (tokens.train) i test (tokens.test) varijabli napravili smo matrice osobina dokumenta (*Document-Feature Matrix, DFM*), korišćenjem funckije dfm() iz bibloteke quanteda . Takođe smo kreirali varijable koje sadrže relativne frekvencije tokena u trening, odnosno test korpusu (*tfidf*). Oba načina predstavljanja teksta biće korišćena za obučavanje modela.

Pregled engrama i birama u korpusu

Analiza engrama i bigrama u korpusu rađena je na trening korpusu, jer on čini 80% ukupnog korpusa. Prvo su iz teksta izvučeni engrami, a zatim bigrami.

Engrami su iz korpusa izdvojeni na osnovu matrice termina i dokumenata (*Term-Document Matrix, TRM*), kreirane pozivanjem funckije TermDocumentMatrix(). Iz matrice su zatim uklonjeni termini koji se ne pojavljuju u makar 1% tekstova, čime je matrica smanjena (removeSparseTerms()). Sva dokumenta su zadržana, dok je broj termina smanjen na 220. Nakon ovoga sabrali smo vrednosti u redovima matrice, kreirali data frame sa rečima i frekvencijama. Dvadeset najčešćih engrama u korpusu prikazali smo grafički stubičastim dijagramom, pri čemu smo rotirali mesta x i y ose, korišćenjem funckije coord_flip().

Dijagram sa prikazom najčešćih engrama nešto nam bliže govori o tome o čemu se najviše govorilo na Twitter-u u martu 2020. godine. Među najšećih 20 reči našle su se *food, store, groceri, stock* itd, pa s obzorim na to možemo zaključiti da se o potrošačkim navikama na samom početku pandemije, u 5000 tvitova izdvojenih iz originalne baze podataka, govorilo nešto više nego u virusu.



Slika 7 - *Najfrekventiji engrami u korpusu*

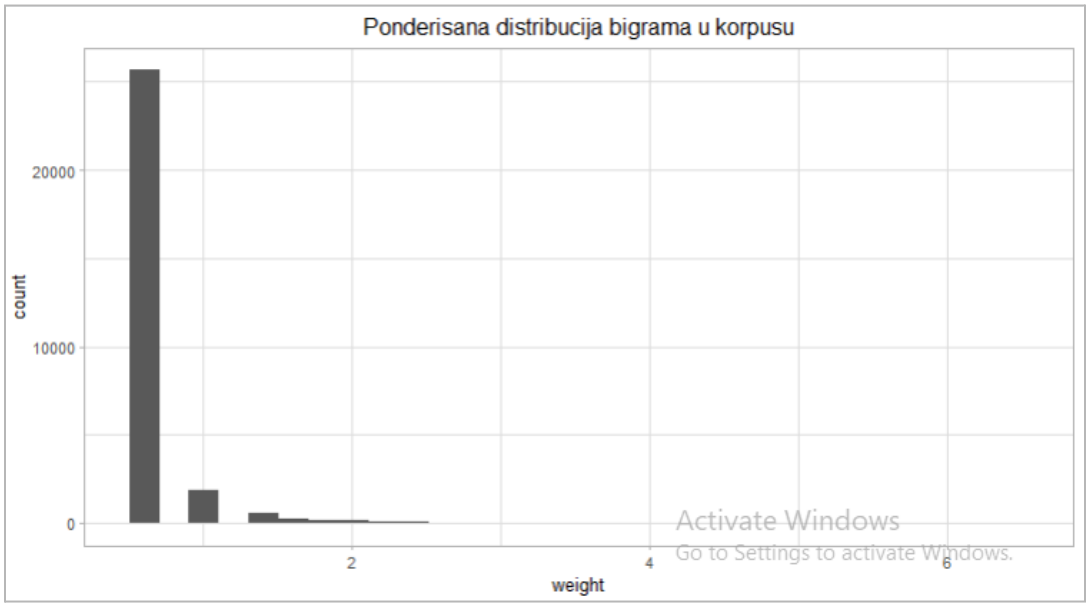
Nakon egrama, iz korpusa smo izvukli i bigrame. Bigrami su izvučeni iz celog korpusa, iz varijable OriginalTweet, a pri samoj ekstrakciji bigrama urađena je i normalizacija teksta. Za ekstrakciju bigrama korišćena je funckija unnest_tokens() iz bibloteke tidytext. Bigrami su razdvojeni na prvu (rec1) i drugu (red 2), pri čemu je svaka reč smeštena u jednoj kolonu.. Korišćenjem funckije count() iz bibloteke prebrojali smo bigrame i u treću kolonu tabele smestili njihove relativne frekvencije. U Tabela 2 prikazali smo 20 najfrekventnijih bigrama.

num.	rec1	rec2	weigh (tezina)
1	grocery	store	844
2	toilet	paper	353
3	online	shopping	313
4	panic	buying	260
5	retail	store	105
6	right	now	90
7	shopping	online	73
8	hand	sanitizer	72
9	grocery	stores	67
10	local	grocery	67
11	oil	prices	58
12	local	supermarket	55
13	empty	shelves	54
14	food	bank	52
15	sick	leave	50
16	stock	market	50
17	food	banks	49

18	supermarket	shelves	49
19	grocery	shopping	48
20	panic	buy	48

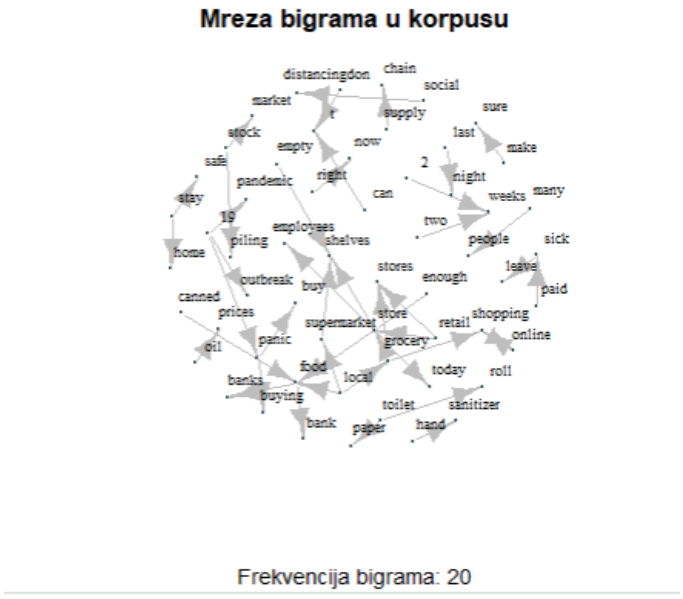
Tabela 2 - *Relativne frekvencije bigrama u korpusu*

Nakon ovoga, prikazali smo distribuciju birama u tekstu koristeći ponderisane relativne frekvencije (težine) bigrama (*log-weigh*) (Slika 8).



Slika 8 - *Ponderisana distribucija bigrama u korpusu*

Kako bismo prikazali odnose između svih bigrama u korpusu kreirali smo dijagram mreže bigrama - Slika 9. Prikazani su bigrami čija je minimalna frekvencija 15. Svaka reč prikazana je jednom tačkom u dijagramu. Reči koje se pojavljuju zajedno u korpusu (bigrami) povezani su linijama. Debljina linije predstavlja jačinu veze između dve reči u korpusu.



Slika 9 - *Mreža bigrama u korpusu*

Modeli

U okviru projekta izrađena su četiri modela za klasifikaciju. Prvi odabrani metod je naivni Bayes (Naive Bayes), a drugi je metoda potpornih vektora (*Support Vector Machine, SVM*). Modeli su izrađeni korišćenjem biblioteke quanteda.textmodels. U okviru oba metoda, prvo smo obučavali model korišćenjem matrice osobina dokumenta (*Document-Feature Matrix, DFM*), a zatim na relativnim frekvencijama termina (*Term Frequency - Inverse Document Frequency, tfidf*).

Pre početka kreiranja modela usklasili smo broj osobina (*features)* test matrice sa osobinama (*features*) trening matrice pozivanjem dfm_match() funkcije. Nova varijabla match.dfm u kojoj je smeštena matrica korišćena je za testiranje modela. Nezavisna varijabla (*DFM*, odnosno *tfidf*) smeštena je u x, a zavisna varijabla (Sentiment) smeštena je u y.

Naive Bayes

Prvi naivni Bayes model (*Naive Bayes, nb*) napravili smo pomoću textmodel_nb() funckije. Prvi model obučavan je pomoću DFM. Model prvo na osnovu konteksta svakoj reči dodeljuje procenat pozitivnog, negativnog i neutralnog sentimenta, a zatim na osnovu toga vrši predviđanje (Slika 10).

```
Class Priors:
(showing first 3 elements)
Negative  Neutral Positive
  0.3333   0.3333   0.3333

Estimated Feature Scores:
             trend     yorker    encount     empti supermarket    shelv    pictur     wegman
Negative 0.0001911 0.00006542 0.00014749 0.0023456    0.003179 0.002392 0.0002264 0.00009951
Neutral  0.0001211 0.00003471 0.00003471 0.0001832    0.003645 0.001388 0.0002075 0.00024299
Positive 0.0002709 0.00001471 0.00001471 0.0007699    0.002943 0.001199 0.0001977 0.00001471
            brooklyn   sold-out    onlin     grocer   foodkick maxdeliveri coronavirus-fear
Negative 0.00005549 0.00010807 0.002390 0.00012519 0.00006542  0.00006542       0.00006542
Neutral  0.00023626 0.00003471 0.002368 0.00048634 0.00003471  0.00003471       0.00003471
Positive 0.00010012 0.00001471 0.003823 0.00009127 0.00001471  0.00001471       0.00001471
            shopper    stock gr76pcrlwh ivmkmsqdt1      find      hand     sanit       fred
Negative 0.0006208 0.003649 0.00006542 0.00006542 0.0007858 0.0011480 0.0008844 0.00005864
Neutral  0.0007079 0.003360 0.00003471 0.00003471 0.0005817 0.0006916 0.0005128 0.00003471
Positive 0.0004426 0.003606 0.00001471 0.00001471 0.0008647 0.0025798 0.0018381 0.00010665
              meyer      turn    amazon      pack      purel     check   concern
Negative 0.00005864 0.0004463 0.0008635 0.0007753 0.00001422 0.0007533 0.0004975
Neutral  0.00003471 0.0003361 0.0003397 0.0004329 0.00003471 0.0007565 0.0008421
Positive 0.00010665 0.0003021 0.0010486 0.0005210 0.00011183 0.0009579 0.0009096
```

Slika 10 - *Sentiment termina u korpusu*

Matrica konfuzije za prvi model prikazana je u Tabela 3. Preciznost, odziv i F skor svih modela biće poređen na kraju opisa projekta. Tačnost modela (*Accuracy*) iznosi 0.6376, a Koenova Kappa 0.3847.

predict.sentnb	Negative	Neutral	Positive
Negative	334	76	115
Neutral	4	17	7
Positive	88	72	286

Tabela 3 - Matrica konfuzije prvog modela

Naive Bayes 1

Drugi naivni Bayes model (*Naive Bayes, nb*) napravljen je na isti način kao i prvi, a obučavan je pomoću tfidf. Matrica konfuzije modela prikazana je u Tabela 4. Tačnost modela (*Accuracy*) iznosi 0.6026, a Koenova Kappa 0.3386.

predict.sentnb	Negative	Neutral	Positive
Negative	298	77	110
Neutral	26	27	21
Positive	102	61	277

Tabela 4 - Matrica konfuzije drugog modela

SVM

Treći model je model zasnovan na metodu potpornih vektora (*Support Vector Machine, SVM*), a obučavan je na *DFM*. Matrica konfuzije modela prikazana je u Tabela 5. Tačnost modela (*Accuracy*) iznosi 0.6587, a Koenova Kappa 0.4514.

predict.sentnb	Negative	Neutral	Positive
Negative	303	40	90
Neutral	41	77	40
Positive	82	48	278

Tabela 5 - Matrica konfuzije trećeg modela

SVM1

Treći model je model je takođe zasnovan na SVM, a obučavan je na *tfidf*. Matrica konfuzije modela prikazana je u Tabela 6. Tačnost modela (*Accuracy*) iznosi 0.5986 , a Koenova Kappa 0.382.

predict.sentnb	Negative	Neutral	Positive
Negative	274	36	94
Neutral	90	94	84
Positive	62	35	230

Tabela 5 - Matrica konfuzije četvrtog modela

Poređenje modela

Preciznost, odziv i F1 mera računati su u okviru svakog modela za svaku klasu zasebno. A predstavljeni su u Tabela 6.

Sentiment:	**Negative**	**Neutral**	**Positive**
Model 1			
Preciznost	0.6362	0.60714	0.6413
Odziv	0.7840	0.10303	0.7010
F1	0.7024	0.17617	0.6698
Model 2			
Preciznost	0.6144	0.36486	0.6295
Odziv	0.6995	0.16364	0.6789
F1	0.6542	0.22594	0.6533
Model 3			
Preciznost	0.6998	0.48734	0.6814
Odziv	0.7113	0.46667	0.6814
F1	0.7055	0.47678	0.6814
Model 4			
Preciznost	0.6782	0.35075	0.7034
Odziv	0.6432	0.56970	0.5637
F1	0.6602	0.43418	0.6259

Tabela 6 - *Metrika modela prema kategorijama*

Kao što možemo videti, svi modeli imaju relativno sličan skor kada su u pitanju negativan i pozitivan sentiment, dok su preciznost, odziv i F1 mera najslabiji kada je u pitanju neutralan sentiment. Razlog ovome može biti znatno manji broj tvitova u korpusu kojima je dodeljena kategorija Neutral, kao što smo videli na grafikonu prikazanom na Slika 3.

Zaključak

Poređenjem četiri modela došli smo do zaključka da svi modeli daju relativno slične rezultate. Nešto boljim od ostalih pokazao se treći model (SVM), čija je tačnost (*Accuracy*) 0.6587, a model sa najmanjom tačnošću (*Accuracy*) bio je četvrti model (SVM1) - 0.5986.

Kao što je pomenuto svi modeli imali su poteškoća sa klasifikacijom neutralnog sentimenta. Treba uzeti u obzir da su na samom početku preuzet samo deo originalne baze podataka od oko 49, 000 tvitova, kao i da je kateogrija sentimenta sa skora od 5 smanjena na 3.