Koliko su stanovnici Srbije zaista privrženi Japanu i japanskoj kulturi? Analiza sentimenta bazirana na Naïve Bayes algoritmu

Isidora Gatarić

Računarstvo u društvenim naukama, Univerzitet u Beogradu, Srbija

e-mail: gataric.isidora@gmail.com

Apstrakt

Cilj ovog istraživanja bio je da se kvantitativno ispita koliko su stanovnici Srbije privrženi Japanu i japanskoj kulturi. Za potrebe dobijanja adekvatnog odgovora na ovo pitanje primenjena je *Analiza sentimenta*, koja omogućava sticanje uvida u to da li se jezički termini, vezani za Japan i japansku kulturu, pojavljuje u pozitivnom kontekstu, ili ne. Nadalje, na osnovu analize koja prethodi pripremi podataka za *Analizu sentimenta* stekao se uvid i u to koliko su termini vezani za ovu prelepu zemlju i njenu kulturu frekventno upotrebljivani među građanima Republike Srbije. Društvena mreža koja je izabrana je Twitter, iz razloga što je pokazano da je jedna od najpopularnijih mreža na kojoj građani Srbije iskazuju svoje mišljenje i svoje stavove. Nakon kreiranja odgovarajuće baze sa podacima (prvenstveno na osnovu korpusa sa tvitovima sa Twitter .sr domena (Ljubešić et al., 2017)), izvršena je *Analiza sentimenta* bazirana na *Naïve Bayes* algoritmu (za koji je pokazano da je najbolji za srpski jezik). Način primene *Naïve Bayes* algoritma prilikom *Analize sentimenta* sproveden je po uzoru na prethodne studije u srpskom jeziku (Milošević, 2012; Grljević, 2016). Na osnovu svega navedenog, stekao se uvid u to koja od sfera interesovanja vezanih za Japan (npr. umetnost, književnost, ekonomija, filozofija itd.) je najprisutnija među građanima Srbije, a koje sfere su zapostavljene. Za kraj, upravo ovo saznanje je omogućilo sticanje uvida u to gde je potreban agresivniji marketing, radi približavanja tog domena građanstvu Srbije.

Uvod

Već dugi niz godina Ambasada Japana u Republici Srbiji i velike japanske kompanije (npr. Japan Tobacco International) ulažu napore u promociju japanske kulture i umetnosti u našoj zemlji. Osim velike popularnosti koju već decenijama unazad zavređuju japanski umetnici, kao što su režiser Akira Kurosava, ili pisac Haruki Murakama, poslednjih godina mnogo se piše i priča i o implementaciji poslovnih filozofija koji datiraju iz japanske kulture

(npr. Kaizen i sl.). Nadalje, Odsek za japanski jezik i književnost na Filozofskom fakultetu u Beogradu dosta radi na promociji književnosti i jezika koji dolazi iz ove prelepe zemlje, dok različite galerije u zemlji izlažu umetničke radove koji prezentuju Japan iz različitih uglova. Međutim, ono o čemu se najmanje priča jeste konkretno beleženje toga koliko su zaista građani Srbije privrženi Japanu i japanskoj kulturi, preciznije rečeno koliko je kvantitativno sam Japan, i teme u vezi sa njim, prisutan u svakodnevnim životima građana Srbije. Shodno tome, primarni cilj ovog istraživanja bio je upravo to *da se kvantitativno ispita zastupljenost tema vezanih za Japana u svakodnevnom životu građana Republike Srbije*. Za potrebe ovog istraživanja korišćene su mere iz korpusa srpskog jezika (srWac (Ljubešić & Klubička, 2016)), kao i baza sa tvitovima preuzetih sa Twitter profila sa .sr domenom (Ljubešić et al., 2017). Nadalje, uz informaciju o zastupljenosti tema vezanih za Japan i japansku kulturu, ovakav set podataka omogućio je i sticanje uvida *u privrženost građana Srbije Japanu, odnosno o tome da li termine vezane za Japan građani Republike Srbije koriste u pozitivnom ili negativnom kontekstu*.

Japan kao večna inspiracija

Japan je ostrvska zemlja, koja se prostire duž istočne pacifičke obale Azije i broji oko 6 800 ostrva na svojoj teritoriji. Ova arhipelagska država prostire se na površini od 2 994 km^2 duž obale, od ostrva Hokaido (blizu Sibira), pa do otoka Okinava (blizu Tajvana) (Lindelauf & Bornoff, 2018), dok je sveukupna površina cele zemlje (i obalski deo i unutrašnjost) 377 923 km^2. Osim glavnog grada Tokia, koji je politički i ekonomski centar moći, među najpopularnijim gradovima su Osaka, centar savremene elektronike i robotike, kao i turistički centri Kjoto i Fušiokaku (Majstorović, 2007). Upravo zbog izrazitog ulaganja u razvoj ekonomije i srodnih disciplina, jedna od stvari koje Japan odvaja od drugih zemalja jeste poslovna filozofija. Poslovna filozofija koja godinama unazad intrigira domaće tržište, ali i inostrane kompanije, jeste **Kaizen**[1]. Osnovna ideja na kojoj počiva ova filozofija jeste to da je najvažniji stalni napredak, preciznije rečeno da nijedan dan ne sme proći bez nekog napretka (Marković & Hrašovec, 2012). Imajući u vidu da je ova filozofija od japanskih kompanija napravila vodeće lidere u različitim oblastima, ne čudi to što je sve veći broj kompanija kod nas zainteresovan za njegovu primenu (Gigić, 2014).

[1] Sve boldovane reči u ovom delu uvodnog dela istraživačkog rada (Japan kao večna inspiracija) referišu na termine koji se najčešće vezuju za sam Japan, a koji su kasnije korišćeni i za izradu baze sa podacima.

Osim uspešne poslovne filozofije, koja je bez premca na celokupnom svetskom tržištu, Japan odlikuje i veličanstvena kultura. Naime, ono po čemu mnogi prepoznaju Japan svakako jesu japanski vrtovi, koji verno oslikavaju japanski reljef (razne vrste drveća, cveća, žbunja i sl.) (Majstorović, 2007). Međutim, već stolećima, jedno drvo zauzima posebno mesto u istoriji Japana, to je drvo japanske trešnje. Ova trešnja nosi naziv **Sakura**, a poznata je kao simbol sreće i prosperiteta (Marković & Hrašovec, 2012). Prema pričama stanovništva ove zemlje, u parkovima u kojima se nalaze drvoredi Sakure ljudi traže mir i odmor od svakodnevnih obaveza, a u skladu s tim se u mnogim gradovima sade vrtovi japanske trešnje (Senta je jedan od nih).[2] Osim toga, u saradnji sa Ambasadom Japana, kompanija Japan Tobacco International je osnovala i Sakura stipendiju, koja podstiče razvoj mladog naučno-istraživačkog kadra u Republici Srbiji. Upravo zahvaljujući nazivu te stipendije, mnogi mladi u Srbiji su imali prilike da se upoznaju sa tradicijom vezanom za cvet prelepe japanske trešnje. Međutim, ovaj cvet nije jedina cvetna lepota koja se vezuje za Japan. Ništa manje atraktivnim smatraju se **ikebana**, aranžiranje živog cveća, i **bonsai**, japanski naziv za saksijsko, minijaturno cveće. Upravo ova dva načina aranžiranja cveća su potekla iz ove prelepe zemlje, a danas su prihvaćeni širom sveta kao nezaobilazni detalj najpoznatijih bašta i vrtova. Sem specifičnog načina aranžiranja biljaka, Japan je svetu dao i fantastičnu kuhinju, u kojoj se specijaliteti baziraju na sirovoj ribi i pirinču. Najpoznatiji specijaliteti iz japanske kuhinje su **sašimi** i **suši**, koji se služe u japanskim restoranima, kojih je sve više u našoj zemlji. Osim sašimija i sušija, posebno mesto u meniju zauzima i **sake**, alkoholno piće od pirinča.

Iako je sve navedeno nesumnjivo doprinelo popularizaciji Japana i japanske kulture među građanima Srbije, čini se da je najveća popularizacija ove prelepe zemlje proizišla iz filmova, književnih i umetničkih dela japanskih umetnika. Kako izveštavaju članci koji su se bavili tom temom (Marković & Hrašovec, 2012), Ivo Andrić i Miloš Crnjanski su prvi počeli spajanje ove dve zemlje kroz svoja književna dela.[3] Nakon toga, osnovan je i Odsek za japanski jezik i književnost, na Univerzitetu u Beogradu, na kojem predaje nekolicina profesora koji prevođenjem najpoznatijih književnih dela Japana i Srbije na oba jezika neguju tu sponu između ove dve zemlje. Osim već spomenutog Haruki **Murakame**, u Srbiji su prevedena i dela Đunićiro **Tanizakia**, Nacume **Sosekia**, Jukio **Mišima** i drugi, ali i **Kođiki**, najstariji zapis u

[2] Kompanija Japan Tobacco International je Senti poklonila vrt japanskih trešnji.
[3] Kako se navodi u specijalnom broju časopisa Vreme, posvećenom Japanu, Ivo Andrić je prvi put u svom delu predstavio Japan (u ExPontu), dok je Miloš Crnjanski preveo tradicionalne lirske pesme iz Japana i organizovao ih u delo Pesme starog Japana.

istoriji Japana (Marković & Hrašovec, 2012). Ništa manje intrigantna nije ni umetnost karakteristična za Japan. Jedna od prvih asocijacija na japansku umetnost svakako je **origami** (kreiranje figurica od papira), kao i **amezaiku** (izrada figurica životnja od šećera ili skrobnog sirupa). Ni japanska filmografija nije zaostajala za ostalim plodnim oblastima koje šire svoje uticaje iz japanske kulture na srpsku. Zlatne godine japanskog filma svakako su one tokom kojih su nastajala dela najpoznatijih režisera, kao što su Akira **Kurosava**, Jasuđira **Ozua**, Kon **Ičikava**, Iširа **Honda** i drugi (Marković & Hrašovec, 2012). Međutim, poslednih par decenija celokupna japanska filmografija se okrenula horor filmovima, a mladi režiseri koji su popularni u tom domenu su Hideo **Nakata**, Kaneto **Šinda** i drugi (Marković & Hrašovec, 2012). Nadalje, posebno mesto u svetskoj umetnosti zauzimaju **mange** i **anime**. Mange su japanski stripovi koje je stovrio Osama **Tezuka** polovinom proteklog veka, a u poslednjih desetak godina sve više intrigiraju i srpsku javnost, koja im posvećuje posebnu pažnju na danima japanizma i sličnim manifestacijama. Od istog čoveka potekle su i anime, animirani filmovi, ali su najpoznatiji predstavnici ove vrste umetničkog izraza Kacuhira **Otoma** i Hajao **Mijazaki**. Imajući u vidu sve navedeno, može se izvesti zaključak da je ova prelepa zemlja, koju mnogi nazivaju i Zemlja izlazećeg sunca (osnovno značenje reci Nipon, od koje je i nastao naziv Japan), neprikosnovne inspiracija ne samo građanima Srbije, nego i celokupnom svetskom stanovništu.

Računarska analiza teksta i sentimenta

Iako se na prvi pogled čini da je više nego jasno koji termini su najzastupljeniji u našem jeziku, a da se usko vezuju sa Japanom i kulturom ove zemlje, kao i to koje je prirode stav građana Republike Srbije prema ovoj zemlji, ipak nije tako. Naime, poslednjih nekoliko decenija sve je više kvantitativnih radova koji nastoje da pruže uvide u već spomenuta dva segmenta: frekvenciju upotrebe nekih konkretnih termina i kvantitativno iskazivanje sentimenta u vezi sa nekim konkretnim temama. Oblast koja se bavi analizom teksta i ekstrahovanjem najvažnijih informacija iz istog zove se *Text Mining*. Konkretna analiza iz domena *Text Mininga* koja će biti korišćena u ovom radu jeste *Analiza sentimenta*, te će u nastavku teksta upravo njoj biti posvećena pažnja, iako je zaista veliki broj analiza koje se u ovom domenu nalaze (npr. *Bag of Words*, *Vector Space Model* itd.).

Analiza sentimenta danas je jedna od najpopularnijih tehnika mašinske obrade teksta, kako u akademskom svetu, tako i u velikim svetskim kompanijama. Prvi rad koji je poznat široj

javnosti, a da je u okviru njega upotrebljena ova analiza, objavljen je pre tridesetak godina, preciznije 1979. godine (Carbonell, 1979). Međutim, svoju naglu popularnost ova analiza stekla je tek početkom dvehiljaditih, kada je došlo do znatnog tehničkog unapređenja računarske obrade jezika, kao i zbog toga što je sve veći broj podataka počeo da biva dostupan javnosti, koja je dobila uvid u široke mogućnosti primene *Analize sentimenta* (Milošević, 2012). Pre sticanja uvida u detaljniji prikaz *Analize sentimenta,* važno je dati objašnjenje i toga šta je tačno to sentiment. *Sentiment* je termin kojim se definiše proces automatske evaluacije teksta i praćenje predviđanja konteksta u kojem se tekst javlja, u smislu da je nekim tekstom izneto pozitivno mišljenje o nekoj situaciji, ili negativno (praktično, u pitanju je recenzija teksta) (Milošević, 2012). Osnovna upotreba *Analize sentimenta* jesu: recenzija nekog sadržaja i kvantitativno iskazivanje mišljenja o nekoj osobi ili nekom proizvodu. Ova analiza najčešće je upotrebljivanja u politici, prilikom političkih kampanji, ali poslednjih godina sve više pažnje zavređuje i u ostalim sferama, pogotovo u marketingu (Shulman, Hevy, & Zevestoski, 2005). Kako istraživanja pokazuju, mnoge velike kompanije (npr. Google) svoj marketing baziraju na podacima koje dobijaju uz pomoć *Analize sentimenta* potrošača, odnosno njihovih komentara sa društvenih mreža i javnih foruma. Najčešći način klasifikovanja komentara jeste binaran (pozitivni vs. negativni komentari), ali naravno, istraživači mogu podesiti algoritam i tako da ima više od dve kategorije. Iako se u prvi mah čini da modeli mašinskog učenja nemaju preciznost uporedivu sa čovekovim sudom, to nije baš tako (Pang, Lee, & Vaithyanathan, 2002). Međutim, važno je ipak napomenuti i to da je ovaj metod analize komentara korisnika nekih društvenih mreža najdelotvorniji ukoliko se osim čistih kodova koristi i ljudsko iskustvo. Preciznije rečeno, ukoliko istraživač fizički pregleda bazu koju model mašinskog učenja izoluje nakon pregleda tvitova sa mreža, odnosno ukoliko evaluira ono što model prepoznaje kao pozitivan, odnosno negativan komentar[4], tada je *Analiza sentimenta* izuzetan metod ekstrahovanja informacija od značaja. U skladu sa preferencijama istraživača, i još važnije kompleksnošću jezika na kojem se istraživanje sprovodi, biraju se algoritmi na kojima se baziraju *Analize sentimenta*. Klasifikacija teksta podvodi se pod modele mašinskog učenja sa superviziranim mašinskim učenjem (Milošević, 2012), te se u skladu sa tim i biraju odgovarajući algoritmi. Podaci u superviziranom učenju su uvek unapred obeleženi (pozitivan/negativan komentar), a postoje dve faze tokom sprovođenja analize: učenje (kako da klasifikuje tekstove na osnovu unapred definisanih odgovora) i predviđanje (na osnovu novog

[4] Razlog zbog kojeg je ovo posebno važno u jezicima kao što je naš jeste veliki broj višeznačnih reči (reč koja ima više različitih značenja (npr. kosa)). Modeli mašinskog učenja nemaju 100% uspešnu mogućnost tumačenja značenja reči iz konteksta (iako postoje neke varijante modela baziranih na vektorima).

seta podataka predviđa kojoj klasi komentari pripadaju). Tri najpoznatija algoritma na kojima se bazira *Analiza sentimenta* su: *Support Vector Machines, Entropy Classification* i *Naïve Bayes*. Za potrebe ovog istraživanja odabran je *Naïve Bayes*, prvenstveno zbog toga što su prethodna istraživanja (Milošević, 2012; Grljević, 2016) sugerisala to da je ovaj algoritam najpodesniji za naš jezik. Neki od razloga, koji su konkretno vezani za ovo istraživanje, su: 1) *Naïve Bayes* algoritam odlično radi na manjim setovima podataka; 2) jednostavan je za primenu; 3) daje najbrže rezultate. Shodno svemu navedenom, u daljem tekstu biće prikazan samo ovaj algoritam.

Naïve Bayes algoritam bazira se na teoriji verovatnoće, preciznije rečeno na teoremi Tomasa Bajesa, koji je bio statističar i filozof. Ovaj algoritam određuje aposteriornu verovatnoću klase na osnovu distribucija reči u tekstu. Nadalje, on ignoriše poziciju reči u tekstu i fokusira se na realizaciju sentimenta reči. Formula po kojoj ovaj algoritam je sledeća:

$$P\ (klasa/atribut) = \frac{P\ (\text{klasa}) * P\ (\text{atribut/klasa})}{P\ (\text{atribut})}$$

Elementi gorenavedene formule su sledeći: *P* je međunarodna oznaka za verovatnoću, *atributi* su elementi teksta (u našem slučaju, reči), dok su *klase* unapred definisane klasima kojima atributi mogu da pripadnu (pozitivan/negativan odgovor). Matematičko izvođenje ove formule znatno je kompleksnije, ali s obzirom da to nije tema ovog istraživanja, kompletan matematički račun neće ovde biti prikazan (za detaljni račun pogledati Milošević, 2012). Na osnovu ove formule razvijene su mnoge varijante *Naïve Bayes* algoritma za različite probleme u kojima se *Analiza sentimenta* koristi, koje su dostupne na raznim forumima na internetu.

Definisanje ciljeva, hipoteza i koraka istraživanja

Iako je ranije spomenuto to koji je glavni cilj ovog istraživanja, u predstojećim pasusima biće ukratko rezimirani isti i konkretni koraci preduzeti zarad ostvarivanja ciljeva. Primarni cilj ovog istraživanja bio je to *da se kvantitativno ispita zastupljenost tema vezanih za Japana u svakodnevnom životu građana Republike Srbije*. Za potrebe ovog istraživanja korišćena je odgovarajući korpus srpskog jezika (srWac). Nadalje, uz informaciju o zastupljenosti tema vezanih za Japan i japansku kulturu, na osnovu seta podataka sa tvitovima sa .sr domena omogućeno je i sticanje uvida *u privrženost građana Srbije Japanu, odnosno o tome da li*

termine vezane za Japan građani Republike Srbije učestalije koriste u pozitivnom ili negativnom kontekstu. Imajući u vidu to da prethodno slično istraživanje nikada do sada nije rađeno, nije moguće postaviti konkretne hipoteze ovog istraživanja. **Prvi korak** ovog istraživanje bio je kreiranje liste sa termina (reči) koje su karakteristične za Japan i japansku kulturu, kao i sticanje uvida u njihove frekvencije. **Drugi korak** u ovom istraživanju bio je formiranje baze sa odgovarajućim tvitovima (onim koji sadrže ključne reči vezane za Japan i japansku kulturu). **Treći korak** predstavljao je dodeljivanje kategorija tvitovima ("POS" ukoliko je u pitanju pozitivan komentar, ili "NEG" ukoliko je komentar negativan). Kao **krajnji korak** sprovedena je *Analiza sentimenta*, čiji rezultati su interpretirani u završnom delu ovog rada.

Metod

Lista termina (i njihovih frekvencija) karakterističnih za Japan i japansku kulturu

Lista termina kreirana je na osnovu specijalnog broja časopisa *Vreme* (Marković & Hrašovec, 2012) u kojem je detaljno predstavljen uticaj Japana i japanske kulture na građane Srbije. Shodno tome, upravo ova literatura se učinila kao najadekvatnijom za kreiranje liste adekvatnih termina za koje je 1) izračunata frekvencija iz on-line korpusa srpskog jezika - *srWac* (Ljubešić & Klubička, 2016); 2) formirana lista upita (reči) koja je dalje korišćena za kreiranje baze za *Analizu sentimenta*. Korpus *srWac* (Ljubešić & Klubička, 2016) napravljen je od mnogobrojnih pisanih sadržaja – romani, novinski članci, istraživački radovi i slično (dostupnih na internetu sa .sr domenom), na osnovu koji su prebrajanje frekvencije pojavljivanja svih reči srpskog jezika. Iako se obično lična imena i prezimena ne koriste prilikom interpretacija frekvencija reči, u ovom slučaju su i ona uvrštena, prvenstveno zbog toga što se neretko različite sfere umetnosti Japana vezuju za konkretne ljude (npr. filmska režija za Akira Kurosavu i slično). Shodno tome, u Tabeli 1 su prikazane reči i njihove frekvencije (bez ličnih imena i prezimena), dok se lična imena i prezimena (kao i imena velikih japanskih kompanija, koje svoja sedišta imaju i u našoj zemlji), vezana za japansku umetnost i kulturu nalaze u Tabeli 2. Interpretacija je data u odeljku rada *Rezultati i Diskusija*.

Tabela 1

Prikaz termina (i njihovih frekvencija) karakterističnih za japansku kulturu

REČ	FREKVENCIJA[5]
Samuraj	751
Haiku	589
Haiga	2
Haibuna	2
Ukijo-e	27
Kaizen	131
Japan[6]	25163
Tokio	1326
Osaka	199
Hokaido	96
Okinava	187
Kjoto	668
Fušiokaku	0
Sakura	59
Ikebana	405
Bonsai	314
Sašimi	19
Suši	1943
Sake	563
Origami	508
Amezaiku	0
Manga	2087
Anima	810
Kođiki	5

Tabela 2

Prikaz ličnih imena (i njihovih frekvencija) karakterističnih za japansku kulturu

REČ	FREKVENCIJA
Japan Tobacco International	103
Ashai Beer co	0
Micui	7
Panasonic	1471
Kobo Abe	0
Rjunosuke Akutagava	0
Tanikava Šuntaro	1
Kakuzo Okakura	6

[5] Brojke predstavljuje broj pogodata (hits) u okviru celog korpusa *srWac*, koji se sastoji od 1 353 238 reči. Na osnovu toga se dobija kvantitativni podatak o frekvenciji određene reči u korpusu.

[6] U ovoj tabeli se nalaze i frekvencije za imena gradova i same zemlje, s obzirom da se one neretko spominju prilikom pisanja tekstova o kulturi u umetnosti Japana.

Haruki Murakami	68
Đunićiro Tanizaki	2
Nacume Soseki	3
Taketori Monogatari	2
Macuo Bašo	5
Jukio Mišima	29
Jasunari Kavabata	9
Kenzaburo Oe	8
Tecuko Kurojanagi	0
Utagava Tojokuni	0
Utagava Hirošige	0
Akira Kurosava	44
Jasuđira Ozua	1
Kenđi Mizoguči	1
Honda	4152
Iširo Honda	2
Nagisa Ošima	3
Hideo Nakata	4
Kaneto Šindo	0
Masaki Kobajaši	1
Osamu Tezuka	5
Hajao Mijazaki	6
Tojota	1662

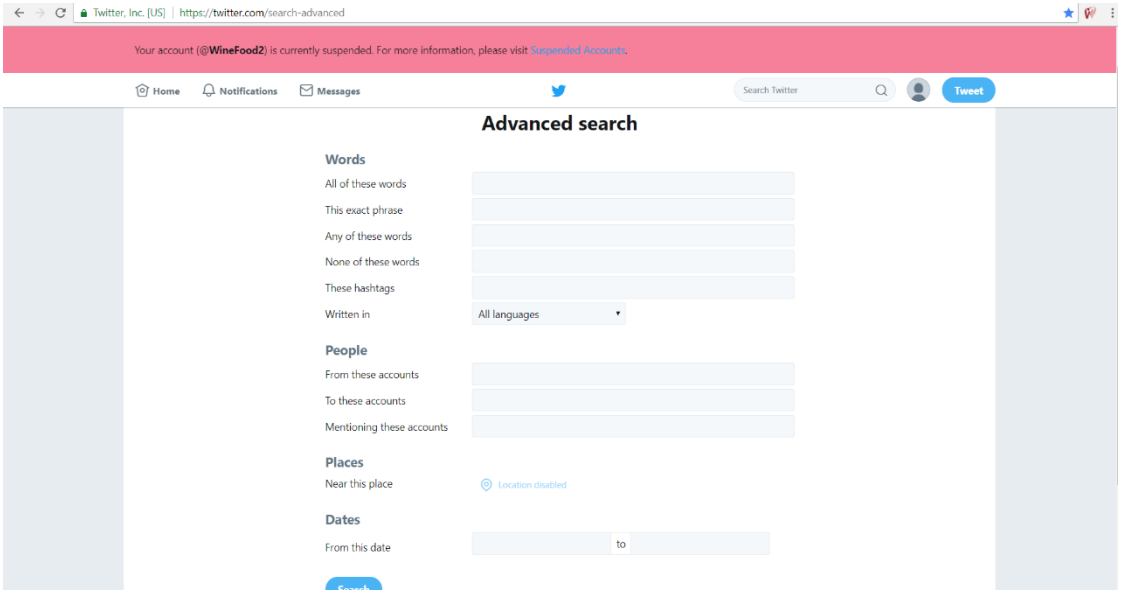
Iz dalje analize su isključeni oni termini kod kojih je zabeležena frekvencija = 0 i oni neće dalje biti komentarisani ni u odeljku *Rezultati i Diskusija*. To su sledeći termini: *Fušiokaku, Amezaiku, Ashai Beer co, Kobo Abe, Rjunosuke Akutagava, Tecuko Kurojanagi, Utagava Tojokuni, Utagava Hirošige, Kaneto Šindo*.

Baza sa tvitovima za Analizu sentimenta

Da bi se preuzeli javno dostupni tvitovi (sa .sr domena) u kojima se spominju termini vezani za Japan i japansku kulturu (navedeni u Tabelama 1 i 2) pristupilo se korpus sa odgovarajućim tvitovima (Ljubešić et al., 2017). Konačan broj selektovanih tvitova (u kojima se nalazi neka od gore pomenutih reči iz Tabela 1 i 2) bio je 8. Imajući u vidu to da je ovo ekstremno mali broj tvitova posvećenih ovoj temi, načinjen je pokušaj pristupa dodatnim tvitovima sa oficijalnog sajta *Twitter API* (Slika 1).[7] Nakon ovog koraka, pronađeno je još 36

[7] Skidanje tvitova iz Twitter baze veoma je zaštićeno (postoje posebni uslovi pod kojima se plaćaju svi podaci), te je stoga jedina mogućnost bila pregled i skidanje (for free) do 30 tvitova koji su selekcionisani od strane kompanije Twitter (Standard Twitter paket).

tvita povezana sa nekom od tema vezanih za Japan i japansku kulturu. Iako je i ovaj broj tvitova premali za jednu analizu mašinskog učenja, ipak se pristupilo sprovođenju *Analize sentimenta*, s obzirom na to da sam algoritam može da radi i na manjim setovim. Ukupna baza sastojala se od 44 dostupna tvita, a kreirana je kao *.csv* fajl. Nakon toga je svakom tviteru dodeljena vrednost, "POS" ukoliko je pozitivan kontekst u kojem se pojavljuje, a "NEG" ako je negativan kontekst.



Slika 1. Prikaz *Twitter API* sajta ne kojem se kucaju upiti koji omogućavaju uvid u tvitove u kojima se nalaže željene reči.

Procedura sprovođenja Analize sentimenta

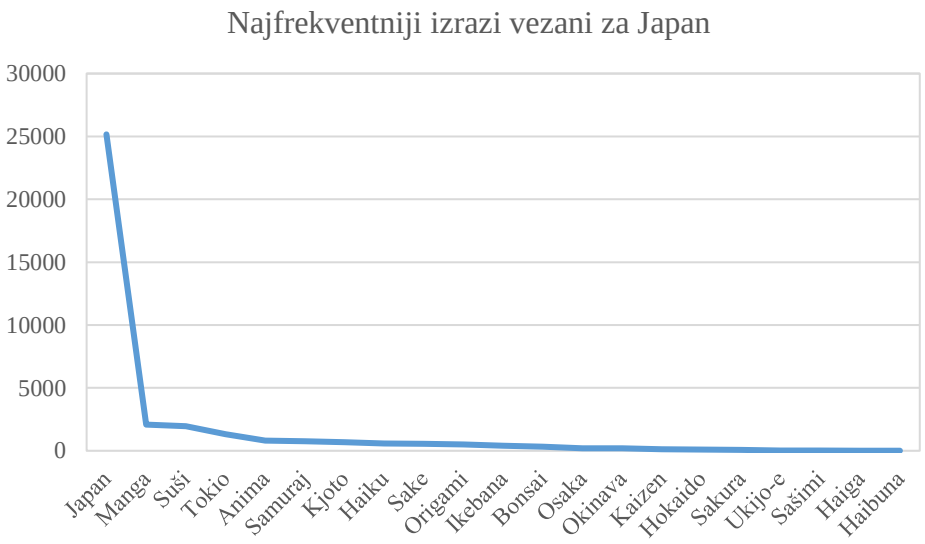
Celokupna *Analiza sentimenta* bila je sprovedena u besplatnom programskom jeziku *R* (R Core Team, 2014), uz pomoć odgovarajućih paketa koji služe sa procesuiranje teksta. Prvi korak bio je prečišćavanje teksta i transformisanje sirovih podataka, tako da oni budu u potpunosti spremni za obradu teksta. Paket koji je korišćen u ovom koraku jeste *tm* (Feinerer & Hornik, 2018), koji omogućava korisnicima da tekst prečiste od mnogobrojnih znakova interpukcije (#,.;-@ itd.), da velika slova pretvori u mala i slično. Drugi korak u ovoj analizi bio je pravljenje takozvanih *Document Term Matrix* od već postojećeg *.csv* fajla sa tvitovima koji su predmet istraživanja ovog rada. Ovaj korak neophodan je zbog pripreme podataka za finalni korak, sprovođenje mašinskog učenja, preciznije rečeno *Analize sentimenta* bazirane na *Naïve Bayes* algoritmu. Za kraj, sprovedena je *Analiza sentimenta*, bazirana na *Naïve Bayes* algoritmu. Za ovaj korak bilo je potrebno instaliranje paketa *caret* (Kuhn et al., 2018), *e1071*

(Meyer et al., 2018) i *pRoc* (Robin et al., 2018). Iako je programski jezik konstantno slao obaveštenje o tome da je set podataka izuzetno mali, o čemu se i ranije pisalo u ovom radu, analiza je sprovedena do kraja. Detaljni rezultati biće prikazani u odeljku *Rezultati i Diskusija*.

Rezultati i Diskusija

U ovom odeljku rada akcenat će biti na interpretaciji rezultata dobijenih primenom metoda opisanih u istoimenom odeljku.

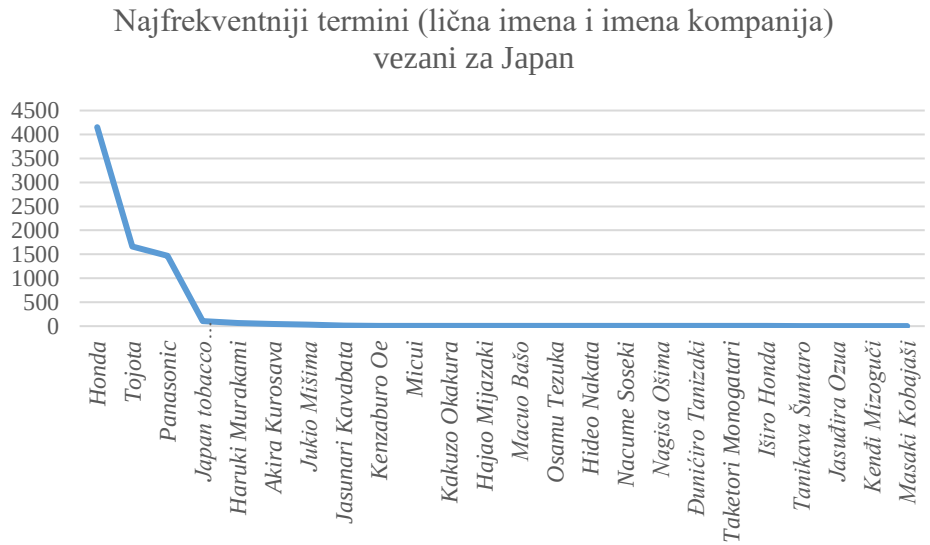
Najfrekventniji termini povezani sa Japanom i japanskom kulturom



Grafik 1. Vizuelni prikaz termina (i njihovih frekvencija) karakterističnih za japansku kulturu; na y-osi se nalaze vrednosti frekvencija (iz korpusa), a na x-osi se nalaze termini.

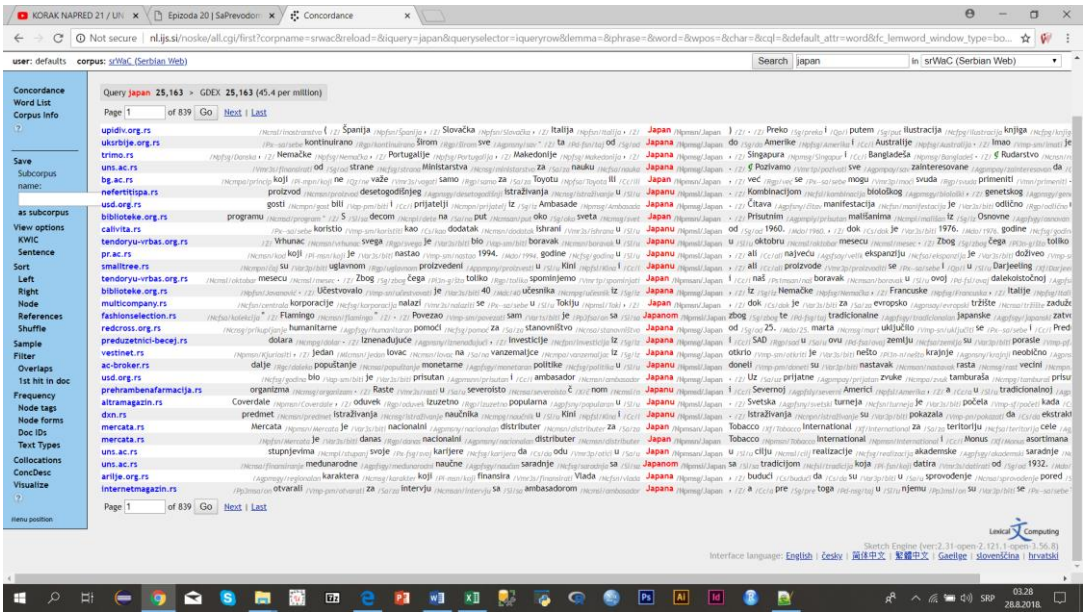
Na osnovu podataka sa Grafika 1 i iz Tabele 1 može se izvesti zaključak da je najfrekventnije upotrebljivana reč Japan, što i nije čudno, imajući u vidu to da je ovo celokupno istraživanje usmereno upravo na ovu temu. Shodno tome, pretpostavlja se da bi **ovaj rezultat mogao sugerisati to da je u svakom tekstu u kojem se pojavljuje bilo koji termin vezan za japansku kulturu spomenuto i ime te države, kako bi se čitaocima dao detaljniji uvid u određene teme koje se u tekstovima obrađuju** (pošto se korpusi prave na osnovu velike količine pisanih sadržaja sa interneta). Nadalje, sledeći po frekvenciji su izrazi vezani za *umetnost i istorije* (manga; anima; samuraj), *hranu* (suši), a nakon toga slede *imena najvećih japanskih gradova* (Tokio, Kjoto itd.). Na osnovu ovih rezultata, **može se izvesti zaključak da**

je među građanima Srbije najzastupljenija tema japanska umetnost (pogotovo mange i anime), što i nije iznenađujuć podatak, ako se uzme u obzir da se neretko organizuju događaji upravo posvećeni ovim sferama umetnosti (npr. Međunarodno takmičenje za najbolji manga strip i slično). Nadalje, među najfrekventnijim terminima je i *Samuraj* koji se vezuje za japansku istoriju, **koja stolećima već intrigira ljude širom sveta, pa ne čudi ni to što kod građana Srbije postoji podjednako interesovanje za tu temu**. Nakon toga, sledi termin *Suši*, koji je vezan za neprikosnovenu japansku kuhinju. Popularnost ovog termina među građanima Srbije nije iznenađujuća, **ukoliko se uzme u obzir to da u Beogradu postoji restoran specijalizovan samo za japansku kuhinju** (restoran Sakura), **kao i to da se često u novinskim člancima pažnje posvećuje japanskoj kuhinji, a pogotovo ovom specijalitetu koji iz nje dolazi**. Ovaj trend slede i japanski gradovi, koji se takođe neretko pominju među građanima Srbije, **a pretpostavlja se da je to pre svega prisutno u novinskim člancima koji pišu reportaže o ovoj prelepoj zemlji i njenim najistaknutijim geografskim tačkama**. Za kraj, kao **najmanje frekventni termini dolaze takođe iz oblasti umetnosti**, što je pomalo kontradiktorno i iznenađujuće. Međutim, imajući u vidu da su u pitanju dva izuzetno specifična termina vezana za slikarstvo i književnost, **pretpostavlja se da većini građana Srbije još uvek nisu poznata ova dva specifična pravca**. Nadalje, **na osnovu ovih rezultata može se izvesti i taj zaključak da su građani Srbije upoznati samo sa najpopularnijim terminima iz svih oblasti japanske kulture** (iz umetnosti mange i anime, iz kulinarstva suši itd.)**, ali da im oni specifični i manje popularni nisu poznati**.

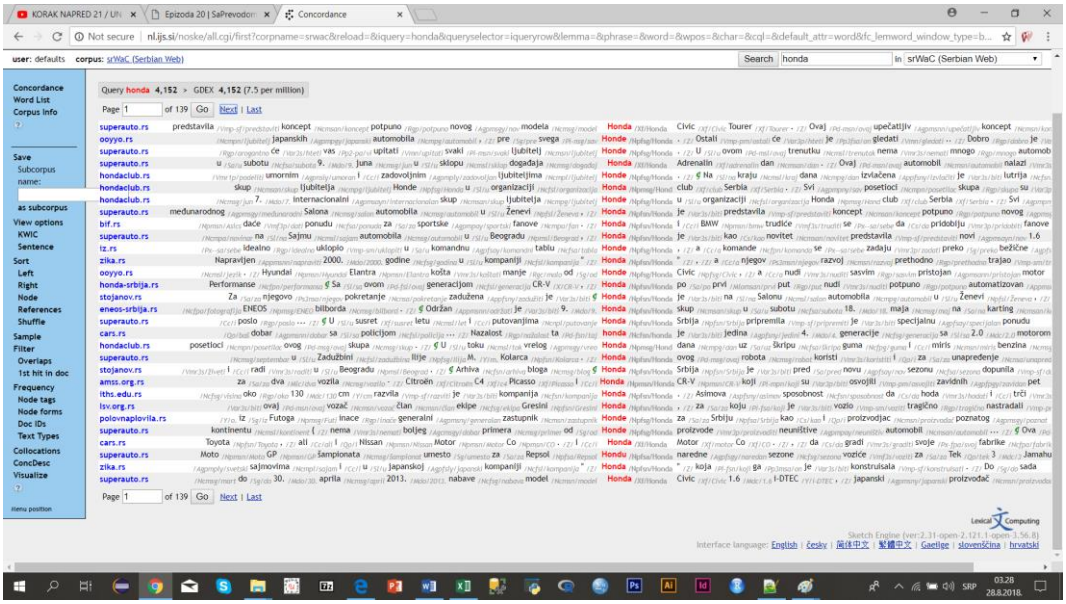


Grafik 2. Vizuelni prikaz termina (i njihovih frekvencija) karakterističnih za japansku kulturu; na y-osi se nalaze vrednosti frekvencija (iz korpusa), a na x-osi se nalaze termini (lična imena i prezimena umetnika, kao i imena velikih kompanija koje posluju u Srbiji).

Na osnovu informacija koje dobijamo sa Grafika 2 i iz Tabele 2 može se izvesti generalan zaključak da su **najfrekventniji termini**, koji se odnose na lična imena i prezimena i imena kompanija, **definitivno iz domena poslovanja**, odnosno da su 4 najfrekventnija termina imena velikih kompanija. Kako rezultati sugerišu, **ubedljivu prednost imaju termini iz japanske auto industrije** (Honda i Tojota), što ne čudi imajući u vidu to da su ova dva giganta prisutna na našem tržištu već dugi niz godina, te da se građani Srbije neretko odlučuju za kupovinu vozila ovih marki. Nakon toga slede *Panasonic*, **elektronska japanska korporacija koja nije jedna od vodećih samo na našem tržištu, nego i u celom svetu**, kao i *Japan Tobacco International*. Razlog toga što je ova kompanija zauzela četvrto mesto među najfrekventnijim imenima koje se vezuju za Japan, ukoliko se po strani stavi činjenica da je u pitanju jedna od najuspešnijih svetskih kompanija u svom domenu poslovanja, treba prvenstveno tražiti u tome što ona već godinama **dodeljuje Sakura stipendiju mladim istraživačima u Republici Srbiji** (te se neretko spominje u štampi i na akademskim sajtovima). Takođe, kompanija **finansijski podržava i događaje kojima se direktno promoviše Japan i japanska kultura** (npr. sadovi japanske trešnje u Senti i slično), te je samim tim veoma zastupljena u javnosti i neretko se spominje u tekstovima u kojima se priča o Japanu u Srbiji. Nadalje, **veoma je interesantno to što su imena umetnika znatno manje zastupljena u korpusu od imena kompanija**. Međutim, očekivano dva najfrekventnija imena umetnika koja se vezuju za Japan, u srpskom korpusu, su *Haruki Murakami* i *Akira Kurosava.* Upravo ovaj rezultat **ide u prilog gore spomenutoj tvrdnji** da rezultati o frekvencijama izraza prvenstveno ukazuju na to **da građani Srbije učestalo koriste samo najpoznatije termine iz svih domena vezanih za Japan, ali da su im manje poznati umetnici, pravci umetnosti i hrana retko zastupljeni u javnim spisima** (od kojih se prave korpusi). Za kraj, dodatno je na Slikama 2 i 3 prikazano je to na kojim sajtovima (i samim tim, u kakvim tekstovima) se najčešće spominju određeni termini vezani za Japan (prikazan je po jedan najfrekventniji iz obe tabele).



Slika 2. Prikaz sajtova u čijim tekstovima se pojavljuje termin *Japan* (prikaz iz korpusa srpskog jezika - srWac).



Slika 3. Prikaz sajtova u čijim tekstovima se pojavljuje termin *Honda* (prikaz iz korpusa srpskog jezika - srWac).

Na osnovu sadržaja Slike 2 i 3 mogu se izvesti sledeći zaključci. Pre svega, termin *Japan* (Slika 2) se **najčešće sreće na sajtovima koji se tiču umetnosti i književnosti** (*upidiv.org.rs* i *uksrbije.org.rs*), **akademskih pitanja** (*uns.ac.rs* i *bg.ac.rs*), ali **poslovanja** (*preduzetnici-becej.rs*). Sve navedeno ide u prilog već spomenutim zaključcima o tome da su među građanima Srbije najdominantnije teme iz oblasti umetnosti, ali da se kao pojedinačna imena najčešće

spominju velike kompanije (povezano sa poslovanjem), koje neretko stipendiraju mlade istraživače (povezano sa akademskim portalima). Nadalje, termin *Honda* (Slika 3) se **uglavnom sreće na sajtovima posvećenim auto industriji** (*superauto.rs*; *hondaclub.rs*; *honda-srbija.rs* i slično). Upravo ovo saznanje ide u prilog već spomenutoj tvrdnji da je **japanska auto industrija jedna od najcenjenijih u Srbiji**, te da ne čudi to što građani Srbije u velikom broju biraju upravo vozila neke od japanskih auto kompanija prisutnih na ovdašnjem tržištu. Međutim, ono što se može primetiti jeste to da **nijedan od sajtova nije povezan sa društvenim mrežama, kao i to da su tekstovi opšteg tipa** (nema iskazvanja sentimenta, nego su prevashodno informativni).

Analiza sentimenta

Podaci na kojima je rađena *Analiza senitmenta* bili su podeljeni na dva dela: *trening* i *test* (faza učenja i faza predviđanja), a nakon toga je implementiran *Naïve Bayes* klasifikator. Pre detaljnije interpretacije podataka, važno je još jednom naglasiti to da je **set podataka izuzetno mali**, te da rezultati nisu toliko interpretabilni na globalnom nivou. U **trening fazi** (fazi učenja) **stekao se uvid u to da model izuzetno dobro predviđa realno stanje stvari**. Postojeći broj tvitova sa "POS" (pozitivan kontekst) je 37, dok je onih sa oznakom "NEG" (negativan kontekst) bilo 7. To je upravo ono što je model i predvideo na training podacima, **na svaki peti "POS" tvit ide po jedan "NEG"** (Tabela 3), dakle predviđeni podaci su ekvivalentni onim koji su opaženi (37 naspram 7).

Tabela 3

Predikcije modela Analize sentimenta

	"POS" (predikcija)	"NEG" (predikcija)
"POS"	1	0
"NEG"	0	5

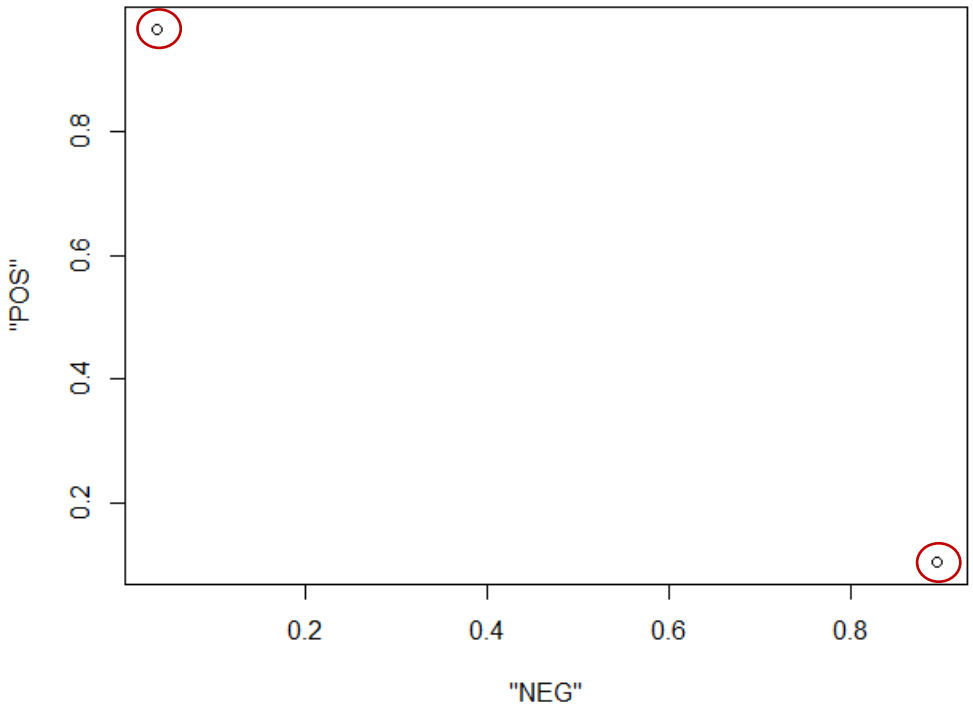
Nakon toga, pristupilo se računanju predikcija za test podatke (faza predviđanja), za šta je takođe upotrebljen *Naïve Bayes* algoritam (formula je opisana u uvodnom delu ovog rada). Na osnovu podata iz Tabele 4 može se steći uvid u to da **model odlično predviđa** (Accuracy = 1 (što je maksimalna vrednost)). Identični skorovi su zabeleženi i na preostale dve mere *Sensitivity* i *Specificity*, dok je nešto niža vrednost (tj. poprilično niska vrednost) zabeležena na meri *Threshold* (Tabela 4). Pretpostavlja se da je **osnovni razlog tome ekstremno mali set podataka, te da je modelu potrebno mnogo više podataka da bi se izvukle realnije**

predikcije. Međutim, ono što može da se zaključi svakako je to da model **odlično predviđa odnos „POS“ i „NEG“ tvitova** (Grafik 3), te se očekuje da bi ovako dobre predikcije i u slučaju većeg seta podataka.

Tabela 4

Predikcije (konkretne vrednosti) modela Analize sentimenta

	Accuracy	Sensitivity	Specificity	Threshold
Vrednost	1.000	1.000	1.000	0.465



Grafik 3. Prikaz predikcija odnosa „POS“ i „NEG“ tvitova. Zaokružene tačkice govore o verovatnoći predviđanja da je neki tvit „POS ili „NEG“, u oba slučaja to je preko 0.8 (što znači preko 80% slučajeva u kojim model ne greši (predvidi tačno ono što je i dobijeno)).

Zaključak

Na osnovu rezultata koji su dobijeni u prvom delu analize, moguće je izvesti nekoliko finalnih zaključaka. Pre svega, **najfrekventniji termini**, korišćeni od strane građana Srbije, a da su povezani sa Japanom i japanskom kulturom su vezani sa **poslovanje**, preciznije rečeno

za gigante japanske auto industrije (Honda i Tojota). Za njima slede *Panasonic* i *Japan Tobacco International*, što nedvosmisleno sugeriše to da su Srbi najviše impresionirani japanskim stilom poslovanja. Nadalje, osim samog imena zemlje o kojoj se ovaj rad piše (koje je jedan od najfrekventnijih izraza u ovoj studiji), primećeno je da su po gradaciji sledeći po frekvenciji izrazi koji se vezuju za **japansku kulturu i umetnost** (manga i anima). Na osnovu toga može se izvesti zaključak da je to druga dominantna tema koja imrepsionira građanstvo Srbije, a da je povezana sa Japanom i japanskom kulturom, te da se o njoj intenzivno čita i piše. Za kraj, primećeno je i to da su **konteksti u kojima se reči prikazuju** (Slike 2 i 3) poprilično **formalni** (akademski sajtovi, časopisi, oficijalni sajtovi i forumi i slično), te da je mali broj sajtova koji su povezani sa nekim društvenim mrežama na kojima se iskazuje mišljenje. Ovo saznanje potkrepljeno je premalim setom podataka skupljenih sa Twittera, o čemu će reč biti u naredom pasusu.

Na osnovu rezultata dobijenih u drugom delu analize, u kojem je primenjena *Analiza sentimenta* bazirana na *Naïve Bayes* algoritmu, takođe je moguće izvesti nekoliko važnih zaključaka. Pre svega, osnovni zaključak jeste taj da je **set podataka u kojima se nalaze tvitovi sa .sr domena** (a u kojima se pojavljuje neka od ključnih reči vezana za Japan i japansku kulturu) **izuzetno mali**. Ali uprkos toj činjenici, na koju nije bilo moguće uticati, ipak su performance modela *Analize sentimenta* veoma dobre. Kako je model sugerisao, većina tvitova je pisana u pozitivnom kontekstu, te je **verovatnoća da svaki novi tvit** (povezan sa Japanom i japanskom kulturom) **bude u pozitivnom kontekstu 5 puta veća**, **nego da bude napisan u negativnom kontekstu** (preciznije rečeno, ukoliko bi nek građanin Srbije sada seo i napisao tvit o Japanu 5 puta je veća verovatnoća da će taj tvit biti pozitivan, nego da će on biti negativan). Na osnovu ovog rezultata može se izvesti finalni zaključak **da građani Srbije veoma pozitivno evaluiraju sve u vezi sa ovom prelepom državom i njenom kulturom**. Međutim, na osnovu rezultata obe analize može se zaključiti i to da **Japan i njegova kulturna znamenja nisu toliko frekventno prisutna među korisnicima društvenih mreža u Srbiji**, te da je možda upravo u ovom segment potrebno unaprediti marketing. Naime, jedan od predloga bi mogao biti taj da se građani Srbije podstaknu da na Twitteru iskazuju svoje mišljenje o nekim manifestacijama vezanim za Japan, njegovim umetnicima, kompanijama i slično (ili možda da se postavljaju neki upitnici na društvene mreže). Na taj način bi se omogućilo širenje baze sa podacima (tvitovima) sa Twittera, odnosno bilo bi moguće sporvesti *Analizu sentimenta* na mnogo većem uzorku, te dobiti i globalno reprezentativnije podatke. Samim tim informacija bi bilo više i zaključci bi mogli biti detaljniji.

Literatura

Gigić, S. (2014). *Kaizen kao poslovna filozofija uspešnih lidera (nepublikovana master teza)*. Fakultet za poslovnu ekonomiju, Univerzitet Singidunum, Beograd.

Grljević, O. (2016). *Sentiment u sadržajima sa društvenih mreža kao instrument unapređenja poslovanja visokoškolskih ustanova (nepublikovana doktorska disertacija)*. Ekonomski fakultet u Subotici, Univerzitet u Novom Sadu.

Feinerer, I., & Hornik, K. (2018). *Package 'tm'*. R Core Team: R Foundation for Statistical Computing, Vienna, Austria.

Lindelauf, P., & Bornoff, N. (2018). *Japan (5th Edition)*. USA: National Geographic Traveler.

Ljubešić, N., Farkaš, D., Klubička, F., Erjavec, T., Miličević, M., & Vuković, T. (2017). *Serbian Twitter training corpus ReLDI-NormTag-sr 1.1*. CLARIN.SI repositry (Open Science), url: https://www.clarin.si/repository/xmlui/handle/11356/1120

Ljubešić, N., & Klubička, F. (2016). *The Serbian web corpus srWaC*. Ljubljana: Jožef Stefan Institute. http://nl.ijs.si/noske/all.cgi/wordlist_form?corpname=srwac

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., Candan, C., & Hunt, T. (2018). *Package 'caret'*. R Core Team: R Foundation for Statistical Computing, Vienna, Austria.

Marković, S., & Hrašovec, I. (2012). Japan u Srbiji: kulturne veze i uticaji. *Vreme, 1146*, 45-54.

Milošević, N. (2012). *Mašinska analiza sentimenta rečenice na srpskom jeziku (nepublikovana master teza)*. Elektrotehnički fakultet, Univerzitet u Beogradu.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C-C., & Lin, C-C. (2018). *Package 'e1071'*. R Core Team: R Foundation for Statistical Computing, Vienna, Austria.

Mladenović, M. (2016). *Informatički modeli u analizi osećanja zasnovani na jezičkim resursima (nepublikovana doktorska disertacija)*. Matematički fakultet, Univerzitet u Beogradu.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 79-86), University of Pennsylvania, Philadelphia, USA.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Robine, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J-C., Muller, M., & Siegert, S. (2018). *Package 'pROC'*. R Core Team: R Foundation for Statistical Computing, Vienna, Austria.

Carbonell, J. (1979). *Subjective Understanding: Computer Models of Belief Systems (Unpublished PhD thesis)*. Yale University, USA.

Shulman, J., Hovy, E., & Zavestoski, S. (2005). Language processing technologies for electronic rulemaking: A project highlight. *Proceedings of Digital Government Research* (pp. 87-88), Atlanta, Georgia, USA.