Klasifikacija recepata po državi iz koje potiče na temelju liste sastojaka

Stjepan Požgaj
Prirodoslovno-matematički fakultet
Sveučilište u Zagrebu
Zagreb, Hrvatska
stjepan.pozgaj@student.math.hr

Luka Tomić
Prirodoslovno-matematički fakultet
Sveučilište u Zagrebu
Zagreb, Hrvatska
luka.tomic@student.math.hr

***Sažetak*—U ovom radu pokušavamo riješiti problem klasifikacije recepta iz koje države potiče ako znamo popis sastojaka koje recept sadrži koristeći klasične metode strojnog učenja. Rješavanje ovog problema omogućuje stranicama za recepte da automatski klasificiraju novo dodani recept i također dodaju određenu mjeru sličnosti između recepata. Zaključak ovog rada je da metode strojnog učenja mogu klasificirati recept po državi iz koje potiče, ali ne savršeno - dobili smo točnost od 77% koristeći linearni SVM. Isprobali smo i neke druge klasične metode strojnog učenja te dobili sličnu točnost.**

***Index Terms*—strojno učenje, sustavi preporuke, klasifikacija, recepti**

I. Uvod

Cilj ovoga projekta je klasificirati recepte po državi iz koje potiče pomoću liste sastojaka koji se koriste u tome receptu. Automatska klasifikacija kuhinje/države iz koje recept dolazi omogućuje stranicama sa receptima da automatski klasificiraju novo dodani recept i dodaju dodatnu mjeru sličnosti između recepata što se može iskoristiti u sustavu preporuke za tu stranicu. Također aplikacije za dostavu i narudžbu hrane omogućuju korisniku da filtrira restorane po tipu kuhinje (talijanska, meksička, domaća itd.). Restorani mogu sami deklarirati koja njihova jela pripadaju nekoj kuhinji, ali sustav preporuka aplikacije može i sam procjeniti koliko je jelovnik nekog restorana ”meksički” ili ”talijanski” i prema tome ga rangirati među ostalim restoranima.

Za automatsku klasifikaciju koristiti ćemo klasifikacijske metode strojnog učenja. Cilj je napraviti što bolji klasifikator te dobiti uvid koje kuhinje su slične jedna drugoj i koji sastojci su indikativni (u pozitivnom ili negativnom smislu) za pojedine kuhinje.

II. Pregled dosadašnjih istraživanja

Postoji nekoliko istraživanja koje se bave klasifikacijom recepta po državi iz koje potiče koristeći listu sastojaka od kojih se recept sastoji. [1] istražuje povezanost između države i sastojka koristeći klasifikacijske tehnike poput asocijativne klasifikacije i SVM koristeći podatke sa stranica food.com i tako pokušava otkriti koji su esencijalni sastojci recepata neke kuhinje i automatski klasificira kuhinju iz koje recept dolazi. Sličnost između kuhinja je promatrana pomoću matrice konfuzije koju su generirali korišteni klasifikatori. Sličan put prati i [2] koristeći pritom naivni Bayes, Multinomialnu logističku regresiju, Slučajne šume i SVM.

[3] koristi iste podatke kao što ćemo i mi koristiti te ne stavlja naglasak na same algoritme i modele strojnog učenja već na utjecaj preprocesiranaj podataka na točnost klasifikacije. Pokušavaju što više očistiti podatke, pronaći sastojke koji su po samom tekstu različiti, a zapravo jednaki. Tako smanjuju broj sastojaka sa početnih 6714 na 4793. Za klasifikacije algoritme koriste naivni Bayes, neuralnu mrežu i SVM te ostvaruju točnost od 80.9% koristeći SVM algoritam.

Sva prethodno navedena istraživanja podatke o sastojcima za neki recept pretvaraju i format prihvatljiv algoritmima strojnog učenja koristeći bag-of-words pristup ili TF-IDF vektorizaciju teksta sastojka.

III. Opis problema

A. Skup podataka

Skup podataka koji ćemo koristiti objavljen je u sklopu Kaggle natjecanje *What's Cooking?* [4], a podatke za to natjecanje je omogućila web stranica Yummly - baza podataka recepata. *What's Cooking?* natjecanje je održano prije nekoliko godina. Na njemu je sudjelovalo preko 1000 timova stoga će biti zanimljivo vidjeti koliko ćemo biti uspješni u usporedbi s njima.

Skup podataka za treniranje sastoji se od 39 774 recepata, a skup podataka za testiranje od 9 944 recepta. Svaki recept u trening skupu se sastoji od liste sastojaka od kojih se sastoji i oznake iz koje države potiče, dok za recepte u testnom skupu imamo samo listu sastojaka od kojih se sastoje.

Sastojci	kuhinja
sugar, pistachio nuts, white almond bark, flour, vanilla extract, olive oil, almond extract, eggs, baking powder, dried cranberries	italian
roma tomatoes, kosher salt, purple onion, jalapeno chilies, lime, chopped cilantro	mexican

Tablica I: Primjeri redaka iz podataka

Ukupno ima 20 različitih država iz kojih recept može biti. Ukupan broj jedinstvenih sastojaka prije procesiranja njihovih

naziva je oko 6 700. Skup trening podataka nije balansiran kao što možemo vidjeti iz Tablice 2.

Kuhinja	Broj recepata
italian	7838
mexican	6438
southern us	4320
indian	3003
chinese	2673
french	2646
cajun creole	1546
thai	1539
japanese	1423
greek	1175
spanish	989
korean	830
vietnamese	825
moroccan	821
british	804
filipino	755
irish	667
jamaican	526
russian	489
brazilian	467

Tablica II: Distribucija recepata po državi iz koje potiče

Također neki sastojci se pojavljuju puno puta npr. sol, luk, voda, maslinovo ulje, šećer, jaja, a neki samo jedanput. Dio razloga zašto se neki sastojci pojavljuju jako mali broj puta je zbog toga što nazivi sastojaka nisu normalizirani - imaju u sebi ime proizvoda, ime marke proizvoda ili imaju pridjev uz proizvod. Marka proizvoda može donijeti neku prediktivnu moć jer je neka marka specifična za određenu državu, no ako bi htjeli klasificirati u duhu samog problema, da imamo samo ime sastojaka kao takvog, onda bi trebali provesti neki postupak normalizacije.

IV. Metode rješavanja

S obzirom da su naši podaci u obliku teksta, odlučili smo da će nam značajke za pojedini recept biti logička vrijednost 0 ili 1 ovisno o tome nalazi li se sastojak u receptu ili ne. Time dobivamo broj značajki koliki je broj jedinstvenih sastojaka. S obzirom da je broj jedinstvenih sastojaka relativno velik (6 700), biti će poželjno smanjiti taj broj sastojaka preprocesiranjem teksta. Zatim ćemo na tako dobivenim podacima upotrijebiti neke od klasičnih modela strojnog učenja da bismo klasificirali recepte. Sve smo radili u Pythonu primarno koristeći bibilioteke sklearn [5], NLTK [6] i matplotlib [7] na računalu sa procesorom Intel(R) Core(TM) i5-8365U i 16GB RAM-a.

A. Procesiranje teksta

Kod jednog sastojka svaku riječ smo gledali zasebno. Sva slova kod riječi smo pretvorili u mala te smo svako slovo normalizirali - pretvorili u jedno od slova iz standardne engleske abecede. Zatim smo svaku riječ lematizirali koristeci *WordNetLemmatizer*. Ovim postupkom dobili smo 2 912 značajki što je manje od polovice početnog jedinstvenog broja sastojaka.

B. Modeli strojnog učenja

Nakon procesiranja teksta, recepte i sastojke smo numerirali u proizvoljnom poretku te podatke složili u matricu gdje jedinica u i-tom retku i j-tom stupcu označava da j-ti sastojak postoji u i-tom receptu. Matricu smo u memoriji držali kao rijetko popunjenu matricu koristeći Pythonov *scipy* paket jer je popunjenost matrice (postotak elemenata koji nisu 0) jednak 0.6 %. Stoga će implementacije modela strojnog učenja koji podržavaju takvu strukturu podataka biti efikasniji od onih koji ne mogu iskoristiti rijetku popunjenost.

Trening skup podataka smo dodatno podijelili na train (80%) i test (20%) pri tome čuvajući omjer kuhinja. Za svaki model smo tunirali hiperparametre trenirajući na train skupu i testirajući optimalnost hiperparametara na testom skupu. Za kriterij optimalnosti koristili smo funkciju *točnosti*. Nakon pronalaska najboljeg skupa hiperparametara, model je ponovno istreniran na cijelom trening skupu podataka.

Isprobali smo sljedeće modele strojnog učenja:

- Slučajna šuma
- Linearni SVM
- Naivni Bayesov klasifikator
- Gradient Boosting klasifikator

Za sve modele koristili smo implementacije iz Python paketa scikit-learn. Strategija predikcije koju smo koristili za svaki model je *OneVsRest* odnosno za svaku vrstu kuhinje u skupu podataka smo trenirali zaseban primjerak modela koji nam daje vjerojatnost (ili neki oblik bodovanja) da dani recept pripada kuhinji za koju je on odgovoran. Prilikom klasifikacije pojedinog recepta svaki primjerak nam daje svoju vjerojatnost i na temelju najviše vrijednosti (pri tome sama strategija može dodatno balansirati dane vjerojatnosti zbog nebalansiranog skupa podataka) određujemo kojom kuhinjom ćemo klasificirati dani recept.

C. Naivni model

Najosnovniji model - onaj koji jednostavno predviđa kuhinju koja je najzastupljenija (to je talijanska) za bilo koji recept postiže točnost od 19.7%. Taj postotak će nam služiti kao orijentir koliko dobro rade modeli strojnog učenja.

V. Rezultati

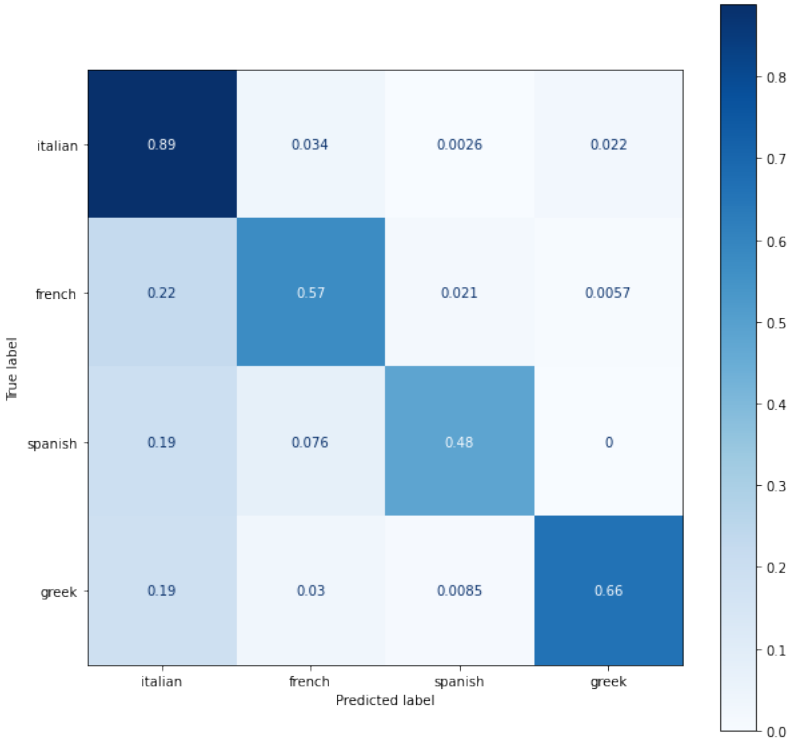
U ovom odjeljku za svaki model kratko ćemo prezentirati rezultate te za klasifikator koji postiže najbolje rezultate detaljnije analizirati njegovu grešku i parametre koje je naučio iz podataka.

A. Slučajna šuma

Za model slučajne šume isprobali smo varirati parametre

- maksimalna dubina stable u šumi (`max_depth`)
- broja stabala u šumi (`n_estimators`),
- veličina skupa značajki prilikom traženja najbolje podjele (`max_features`),
- potreban broj primjeraka u čvoru za dijeljenje istog (`min_samples_split`)

S obzirom da nismo postavljali granicu na dubinu stabala u šumi dobivamo skoro savršeni rezultat na train skupu, dok na



Slika 1: Dio matrice konfuzije koji prikazuje odnose predikcija za talijansku, francusku, španjolsku i grčku kuhinju. Vidimo da klasifikator ima tendenciju prepoznavati druge europske kuhinje kao talijansku te u manjoj mjeri španjolsku kao francusku.

validacijskom skupu podataka dobivamo točnost od 76% za optimalne hiperparametre. Tuniranje parametara nije značajno poboljšalo model, najgori model je imao točnost od 74% na validacijskom skupu.

B. Linearni SVM

Koristili smo SVM sa linearnom jezgrom jer prema Yangu i Liu [4], linearne jezgre su bolje od nelinearnih jezgri na primjenama kod klasifikacije teksta. Za ovaj model varirali smo sljedeće parametre :

- regularizacijski parametar C (`C`),
- normu kod regularizacije (`penalty`),
- minimizacijska funkcija (`loss_function`)
- balansiranje parametre C prema pojedinim klasama (`class_weight`)
- maksimalan broj iteracija algoritma (`max_iter`)

Linearni SVM je samo mrvicu bolji od slučajnih šuma - dobivena točnost je 77%. Kod ovog modela bitno je bilo tunirati hiperparametre jer dobivamo široki raspon točnosti za različite kombinacije hiperparametara.

C. Naivni Bayesov klasifikator

Koristimo Multinomijalni Naivni Bayesov klasifikator koji koristi naivni Bayesov algoritam za multinomijalno distribuirane podatke i standardno se koristi kao naivni Bayesov klasifikator u klasifikaciji teksta (gdje su podaci obično prikazani kao vektori prebrojavanja) [5].

Hiperparametri koje smo tunirali za ovaj model su:

- parametar izgladivanja (`alpha`),
- a priori vjerojatnosti (`fit_priori`)

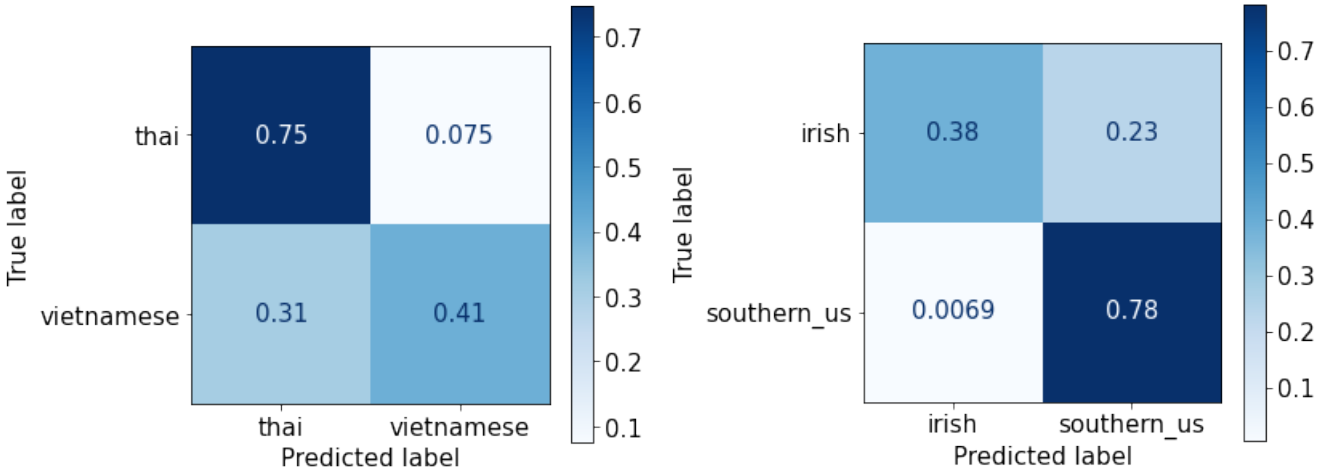
Naivni Bayesov klasifikator je malo lošiji od linearnog SVMa, ali s obzirom na jednostavnost modela pokazuje dobre rezultate - dobivamo točnost od 73%.

Gradient Boosting klasifikator se pokazao lošijim od linearnog SVMa, dobili smo točnost od 75.7%.

D. Analiza greške linearnog SVM

S obzirom da se linearni SVM pokazao najbolji klasifikatorom pogledati ćemo njegovu matricu konfuzije i parametre koje je naučio za klasifikaciju pojedinih kuhinja. Cijela matrica konfuzije nalazi se u dodatku, a ovdje ćemo izdvojiti njezine najbitnije dijelove.

Kao najbitnije dijelove matrica konfuzije smo izdvojili one gdje se događaju najveće greške. Na prvoj slici primjećujemo da se značajan dio europskih kuhinja klasificira kao talijanska te da se 19% španjolske kuhinje krivo predviđa kao talijanska. S obzirom da je talijanska kuhinja najmnogobrojnija te da se greška događa većinom u smjeru talijanske kuhinje, kao jedan od razloga zbog kojeg nam se događa greška u predikciji može biti nebalansiranost skupa podataka. Na drugoj slici imamo dva dijela matrice konfuzije. Na prvom vidimo da model čak 31% vijetnamske kuhinje predviđa kao tajlandsku. A na drugom vidimo da 23% irske kuhinje model predviđa kao kuhinju južnog dijela SADa dok u obrnutom smjeru imamo značajno manju grešku u odnosu na spomenute. Također možemo primjetiti da je za irsku kuhinju općenito



Slika 2: Dijelovi matrice konfuzije koji prikazuju odnose predikcije tajlandske i vijetnamske kuhinje odnosno francuske kuhinje i kuhinje južnog dijela SADa. Možemo primijetiti da se čak 31% vijetnamske kuhinje predviđa kao tajladska te da model predviđa 23% irske kuhinje kao kuhinju južnog dijela SADa.

jako slaba točnost. U dodatku na kraju dokumenta možemo vidjeti grafove najznačajnijih sastojaka prema linearnom SVM klasifikatoru za pojedinu kuhinju te neke od zaključaka koje možemo dobiti iz njih. Jedna od bitnijih stvari koja se može primjetiti iz tih grafova je da kod mnogih kuhinja među najznačajnijim pozitivnim značajkama je samo ime te kuhinje. Također kod nekih kuhinja je među najznačajnijim negativnim značajkama ime neke druge kuhinje.

E. Određivanje sličnih kuhinja i sastojaka

Sličnosti kuhinja i recepata proučavali smo pomoću Nenegativne matrične faktorizacije na analogan način na koji se određuju teme dokumenata, gdje je svaki dokument raprezentiran s bag-of-words reprezentacijom. U ovom slučaju komponente koje daje NMF se interpretiraju kao skupine sastojaka od kojih se neka kuhinja sastoji, od kojih svaka daje određenu vjerojatnost za pojavljivanje određenog sastojka. Pomoću NMF-a smo generirali 10 skupina sastojaka. Slike 3, 4, i 5 prikazuju histograme koji pripadaju određenim skupinama sastojaka. Na lijevom histogramu vidimo najfrekventnijih 10 sastojaka u toj skupini s njihovim udjelima u toj skupini izraženoj u postocima. Desni histogram prikazuje 10 kuhinja čiji recepti su u najvećem postotku koristili navedenu skupinu sastojaka. Na taj način možemo vidjeti koji sastojci se zajedno pojavljuju u receptima, ali i koje kuhinje su međusobno slične jer koriste iste skupine sastojaka. U opisima slika možete vidjeti zaključke dobivenih iz histograma nekih zanimljivih skupina sastojaka.

VI. Zaključak i daljnje istraživanje

Tijekom ovog istraživanja istrenirali smo četiri modela strojnog učenja : Slučajne šume, Linearni SVM, Multinomialni naivni Bayesov klasfikator i Gradient Boosting klasifikator. Najbolji se pokazao linearni SVM sa točnošću od 77% uz marginalnu prednost nasprem Slučajnih šuma - 76% i Gradient Boosting klasifikatora 75.7%. Multinomialni naivni Bayesov klasifikator nije bio daleko iza njih sa točnošću od 73%. Ove razine točnosti su značajno veće od točnosti naivnog klasifikatora stoga možemo reći da metode strojnog učenja smisleno rade na ovom problemu.

Uz to smo koristeći Nenegativnu faktorizaciju matrica uspjeli odrediti skupine sastojaka koji su međusobno povezani i grupe kuhinja koje su slične.

Iz matrice konfuzije primjetili smo da se neke od klasa u značajnom postotku krivo klasificiraju kao određena druga klasa što je problem s kojim bi se trebalo u daljnjem istraživanju pozabaviti. Jedna ideja za istražiti je probati iskoristiti *OneVsOne* strategiju uz dodatni klasifikator koji će klasificirati za klase koje se miješaju.

Također bilo bi zanimljivo vidjeti koliko dobar bi bio klasifikator kada bismo mu maknuli imena država iz sastojaka, vidjeli smo da su takve značajke bile među najznačajnijim u nekoliko kuhinja. Time bismo dobili problem koji je možda malo zahtjevniji za riješiti, ali više u duhu povezivanja kombinacije sastojaka sa kuhinjem nego samih imena sastojaka.

Literatura

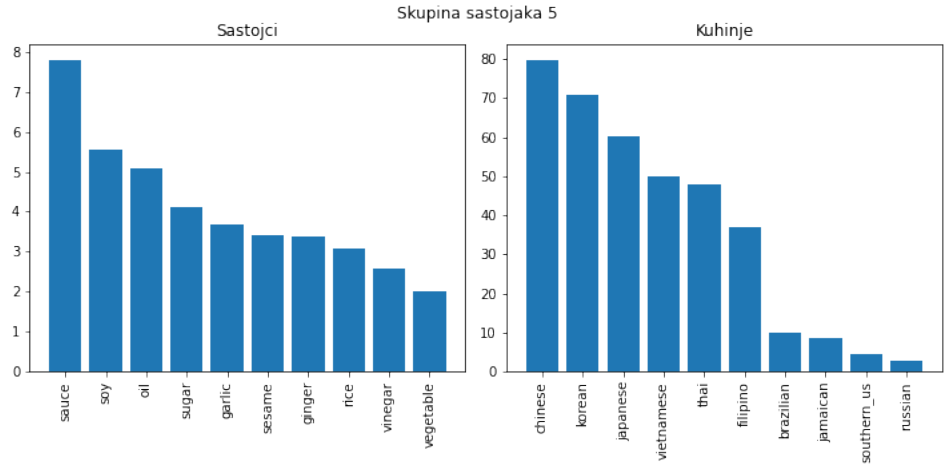
[1] Han Su, Ting-Wei Lin, Cheng-Te Li, Man-Kwan Shan, and Janet Chang. 2014. Automatic recipe cuisine classification by ingredients. In Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication (UbiComp '14 Adjunct). Association for Computing Machinery, New York, NY, USA, 565–570. DOI:https://doi.org/10.1145/2638728.2641335

[2] S. Jayaraman, T. Choudhury and P. Kumar, "Analysis of classification models based on cuisine prediction using machine learning," 2017 International Conference On Smart Technologies For Smart Nation (SmartTechCon), 2017, pp. 1485-1490, doi: 10.1109/SmartTechCon.2017.8358611.

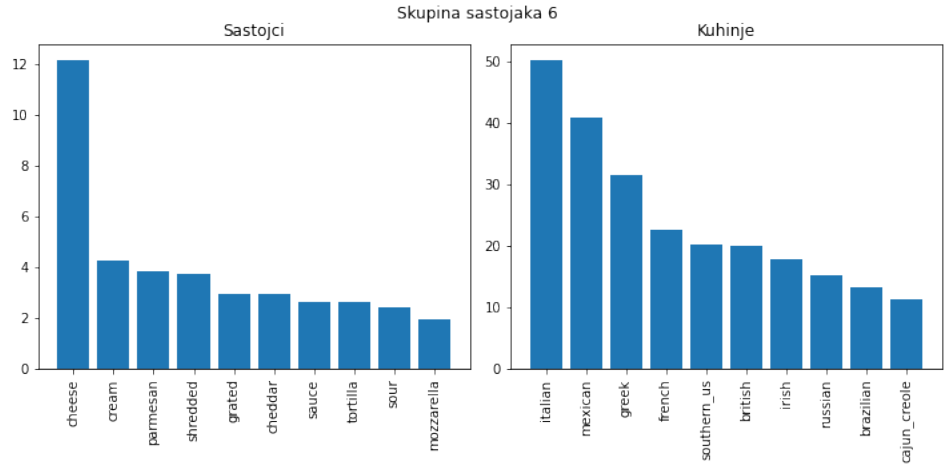
[3] S. Kalajdziski, G. Radevski, I. Ivanoska, K. Trivodaliev and B. R. Stojkoska, "Cuisine classification using recipe's ingredients," 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2018, pp. 1074-1079, doi: 10.23919/MIPRO.2018.8400196.

[4] What's cooking, Kaggle, https://www.kaggle.com/c/whats-cooking/overview

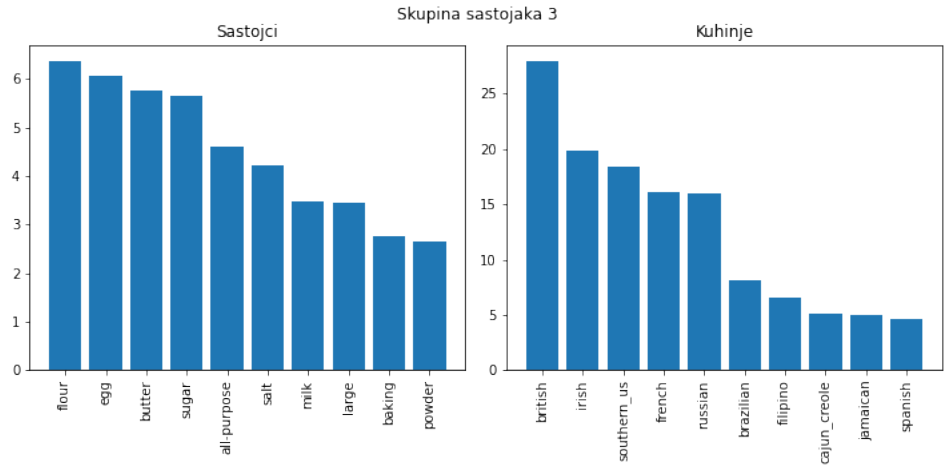
[5] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.
[6] Bird, Steven, Edward Loper and Ewan Klein (2009), Natural Language Processing with Python. O'Reilly Media Inc.
[7] J. D. Hunter, "Matplotlib: A 2D Graphics Environment", Computing in Science & Engineering, vol. 9, no. 3, pp. 90-95, 2007.



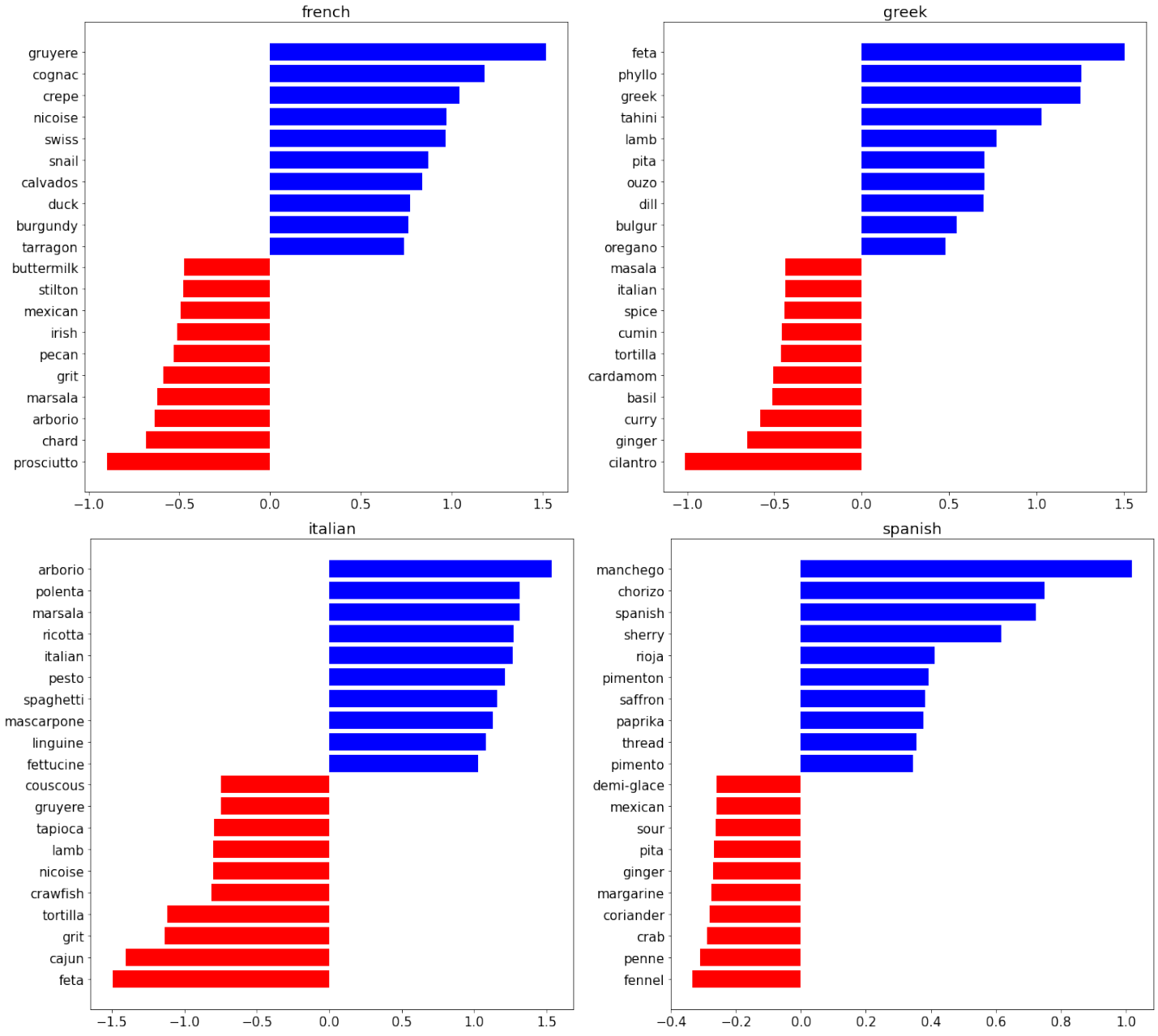
Slika 3: U ovoj skupini su sastojci koji su najdominantniji u azijskim kuhinjama.



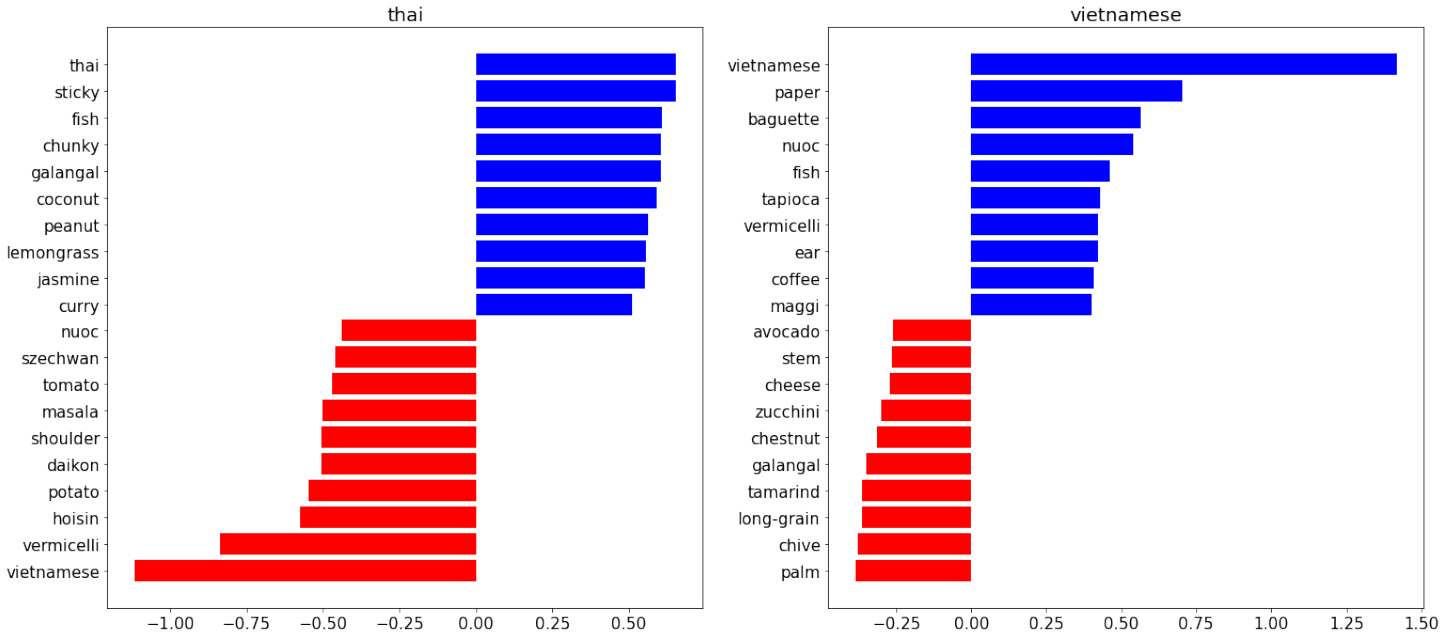
Slika 4: U ovoj skupini su sastojci koji su najdominantniji u europskim kuhinjama, s naglaskom na mediteranske kuhinje i meksičkoj kuhinji koja ima puno zajedničkog s mediteranskim.



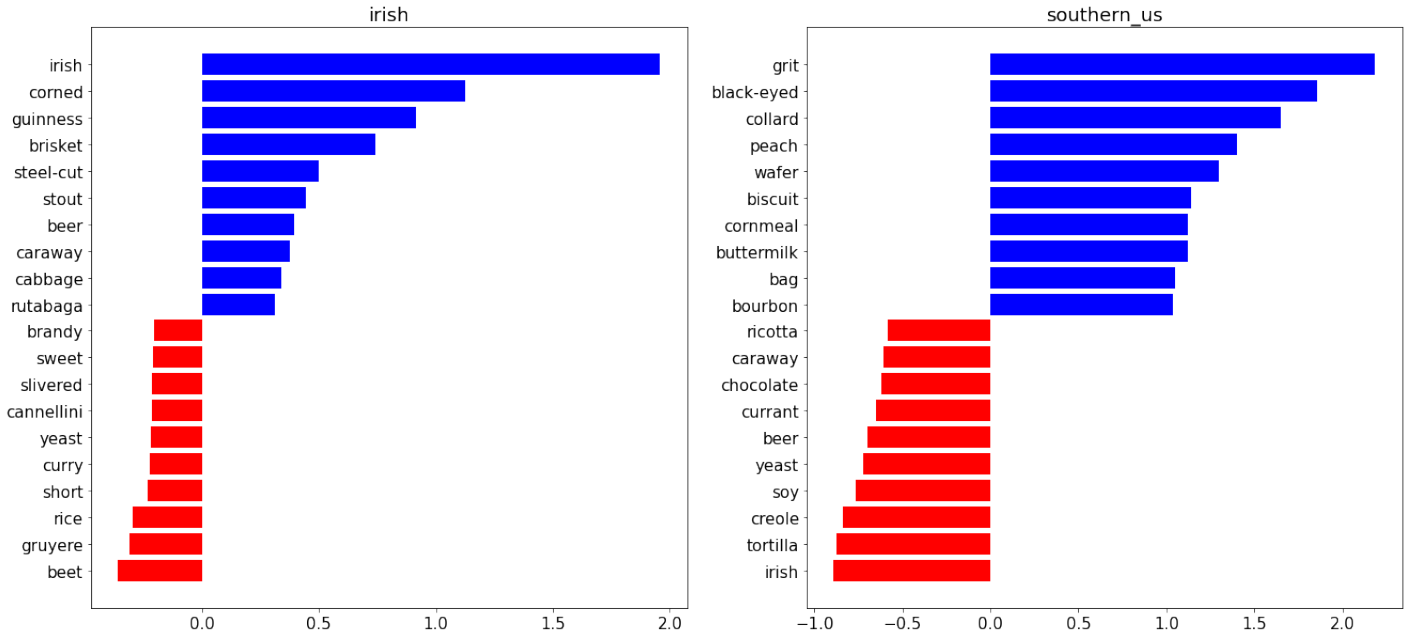
Slika 5: U ovoj skupini su najdominantniji sastojci od kojih se radi tijesto, a vidimo da su takvi proizvodi najčešći u zapadnoj Europi i Americi.



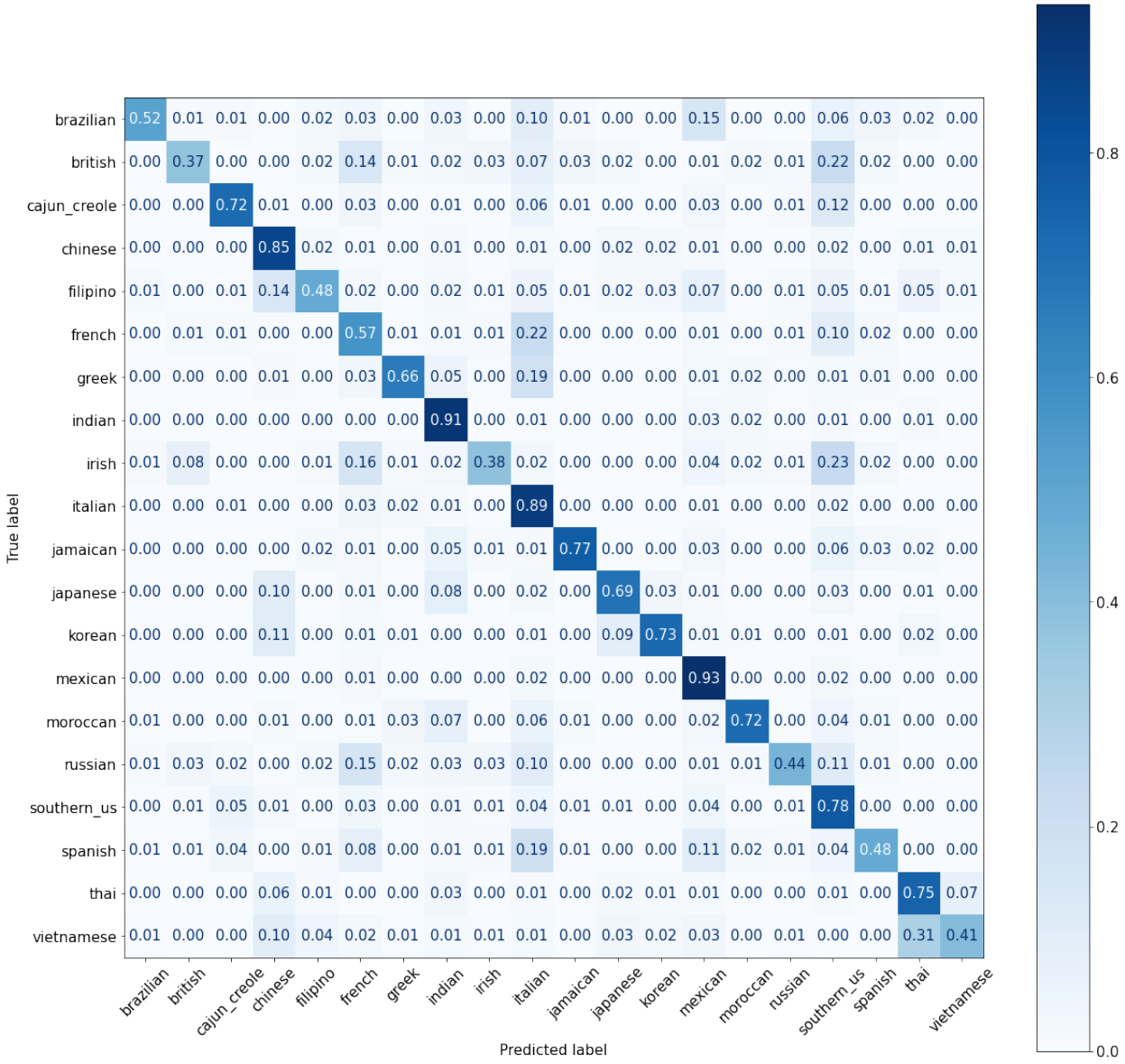
Slika 6: Najznačajnijih 10 značajki (u pozitivnom i negativnom smislu) koje uzima u obzir linearni SVM klasificirajući francusku, grčku, talijansku i španjolsku kuhinju. Neke od stvari koje možemo primjetiti je da francuska, grčka i španjolska kuhinja imaju kao negativnu značajku ime druge kuhinje, a također grčka, španjolska i talijanska kuhinja imaju ime svoje države kao pozitivnu značajku. Nadalje značajka *arborio* je kod talijanske kuhinje najpozitivnija dok je kod francuske kuhinje treća najnegativnija.



Slika 7: Najznačajnijih 10 značajki (u pozitivnom i negativnom smislu) koje uzima u obzir linearni SVM klasificirajući tajlandsku i vijetnamsku kuhinju. Prvo što uočavamo je da tajlandska i vijetnamska kuhinja za najpozitivniju značajku imaju upravo ime svoje kuhinje, a tajlandska kao najnegativniju značajku ima ime vijetnamske kuhinje. To je vjerojatno rezultate *borbe* klasifikatora da ne svrstava vijetnamsku kuhinju pod tajlandsku. Drugo što možemo primjetiti je da obje kuhinje imaju *fish* kao jedan od najpozitivnijih sastojaka što može biti jedan od razloga zašto se događa miješanje tajlandske i vijetnamske kuhinje.



Slika 8: Najznačajnijih 10 značajki (u pozitivnom i negativnom smislu) koje uzima u obzir linearni SVM klasificirajući irsku kuhinju i kuhinju južnog dijela SADa. Uočavamo da je ime irske kuhinje najznačajnija značajka te da nema jako negativnih značajki (uzimajući u obzir ostale grafove). Također uočavamo da je najnegativnija značajka kod *souther_us* kuhinje upravo ime irske kuhinje.



Slika 9: Matrica konfuzije svih klasa za linearni SVM klasifikator normalizirana po retcima. Možemo primjetiti da se neke klase krivo klasificiraju kao neke druge klase npr. vietnamska krivo klasificirana kao tajlandska, ruska kao francuska, francuska kao talijanska itd. Također možemo primjetiti da je postotak greške po klasi koreliran sa brojem primjeraka u skupu podataka za trening.