

**Državni univerzitet u Novom Pazaru**

**Departman za tehničke nauke**

**Smer: Softversko inženjerstvo**

**Seminarski rad**

**Predmet:** Mašinsko učenje

**Tema:**

Prepoznavanje emocija u tekstu

**Mentori: Studenti:**

Ulfeta Marovac Mihajlo Veljković

Aldina Avdić Dženis Aličković

Novi Pazar, 2023.godina

Sadržaj

[1. Uvod 3](#_Toc135923605)

[2. Osnovni pojmovi 3](#_Toc135923606)

2.1 Tekst mining 3

2.2 Prirodna obrada jezika (NLP)………………….………………………………………………3

2.3 Predstavljanje rečenice preko vektora reči.................................................................................3

2.4 TF-IDF........................................................................................................................................4

2.5 Transformeri..............................................................................................................................4

3. Opis skupa podataka..................................................................................................................4

4. Metrike za evaluaciju modela..................................................................6

5. Algoritmi korišćeni u projektu.........................................................................................................................7

6. Prikaz rezultata......................................................................................................................8

7.Analiza rezultata…………………………………………………………………………………………………………………….17

7. Zaključak..................................................................................................................................17

8. Literatura.................................................................................................................................18

# **Uvod**

U cilju efikasnijeg dobijanja informacija iz velike količine javno dostupnih informacija, korisno je izvršiti njihovu klasifikaciju. Klasifikaciju teksta (eng. text classification) predstavlja analitički proces kojim se tekstualni dokumenti svrstavaju u određene kategorije ili klase na temelju njihovog sadržaja.

Prepoznavanje emocija u tekstu ima široku primenu u oblastima kao što su: analiza društvenih medija, korisničko iskustvo, analiza sentimenta, mentalno zdravlje i personalizacija sadržaja. Potreba za prepoznavanjem emocija u tekstu postoji kod različitih vrsta dokumenata. To mogu biti: emailovi, ankete, različite vrste recenzija…

Cilj ovog seminarskog rada jeste prepoznavanje emocija u tekstu primenom metoda mašinskog učenja.

1. Osnovni pojmovi

# **Tekst mining:**

Tekst mining, odnosno rudarenje teksta, predstavlja proces otkrivanja korisnih informacija, obrazaca i znanja iz tekstualnih podataka. Ova tehnika koristi NLP i mašinsko učenje kako bi analizirala, razumela i dobila uvid iz tekstualnih podataka. Tekst mining uključuje aktivnosti poput prepoznavanja entiteta, klasifikacije teksta, analize sentimenta, otkrivanja uzoraka i još mnogo toga. Podaci se mogu analizirati na nivou reci, recenica ili celih dokumenata.

1. **Prirodna obrada jezika (NLP):**

Obrada prirodnih jezika (eng. natural language processing) je disciplina koja se bavi obradom teksta napisanog prirodnim jezikom i veoma je popularna u oblasti računarstva, pre svega zbog činjenice da je najveći deo svih dostupnih informacija zapisan upravo prirodnim jezikom u tekstualnom formatu. Cilj je omogućiti računarima da razumeju, tumače i kominiciraju s ljudskim jezikom na način na koji je smislen za ljude.

1. **Predstavljanje rečenice preko vektora reči:**

Predstavljanje rečenica preko vektora reči (**Word Embedding**) je tehnika koja pretvara reči ili rečenice u vektorske reprezentacije. Ova tehnika pomaže u tome da se tekstualni podaci numerički predstave kako bi se mogli koristiti u mašinskom učenju. Popularne metode vektorske reprezentacije uključuju: Word2Vec, GloVe i FastText.

1. **TF-IDF:**

**TF model** (eng. **term frequency**) meri koliko često se određena reč pojavljuje u dokumentu.

**IDF** meri koliko često se određena reč pojavljuje u skupu dokumenata.

**TF-IDF model** je vrednost za određenu reč u dokumentu. To je proizvod TF i IDF vrednosti za reč, odnosno:

**TF-IDF = TF \* IDF**

Veća TF-IDF vrednost ukazuje na to da je reč važna u dokumentu, ali je istovremeno retka u celom skupu dokumenata.

**TF-IDF metoda** ima značajnu ulogu u prepoznavanju emocija u tekstu jer pomaže u identoifikaciji ključnih reči koje su povezane sa određenim emocijama.

1. **Transformeri:**

**Transformer** je vrsta arhitekture dubokih neuronskih mreža koja je revolucionirala područje NLP-a. Transformer se temelji na mehanizmu pažnje (self-attention) koji omogućuje modelu da nauči kontekstualne veze između reči u tekstu. Ova arhitektura je posebno korisna u zadacima mašinskog prevođenja i generisanje teksta.

**BERT** (**Bidirectional Encoder Representations from Transformers**) je model dubokog učenja temeljen na Transformer arhitekturi. Predstavljen je 2018. godine od strane kompanije Google. BERT je proširenje Transformer modela koje je trenirano na velikom broju tekstualnih podataka kako bi naučio duboko kontekstualno predstavljanje reči. On je sposoban za razumevanje i generisanje teksta te postiže izvanredne rezultate u mnogim NLP zadacima, uključujući klasifikaciju, prepoznavanje entiteta i analizu sentimenta.

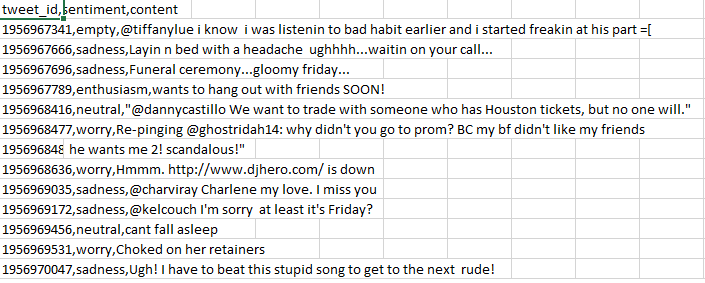
1. Opis skupa podataka

Za naš skup podataka koristićemo „Prepoznavanje emocija u tekstu“ sa stranice Kaggle. Skup podataka se sastoji iz 13 emocija koje smo svrstali u 3 glavne klase sa: pozitivnim, negativnim i neutralnim emocijama.

* **Positive**: relief, happiness, love, enthusiasm, surprise, fun
* **Negative**: boredom, hate, sadness, anger, worry
* **Neutral**: empty, neutral

Važno je da skup podataka bude dovoljno veliki i uravnotežen, sa dovoljnim brojem primera za svaku klasu (pozitivna, negativna i neutralna) kako bi se model pravilno obučio i generalizovao.

Deo skupa podataka:



Ključni koraci u ovom procesu će uključivati:

* **Korak pretprocesiranja** je važan u celokupnom procesu klasifikacije teksta prema sentimentu i može znatno uticati na tačnost klasifikacije. Pretprocesiranje nije obavezno, ali poboljšava performanse celog procesa. Ulaz u proces pretprocesiranja je sirovi tekst, koji najpre treba obraditi. Neke metode koje se koriste u fazi pretprocesiranja su: uklanjanje stop-reči, stemovanje, lematizacija, prevođenje velikih slova u mala, uklanjanje znakova interpukcije i normalizacija teksta.
* **Vektorizacija reči:** Za vektorizaciju reči koristili smo biblioteku sklearn. Koristili smo CountVectorizer gotovu funkciju za vektorizaciju teksta. Ovaj vektorizator pretvara tekstualne podatke u numerički oblik, gde se svaka reč ili n-gram reči predstavljaju kao zaseban atribut, a vrednost atributa je broj pojavljivanja te reči u datoj instanci. Odabir adekvatnog algoritma za mašinsko učenje, kao što su Naive Bayes, Logistic regression, SVM, IBK ili Random Forests, i treniranje modela na trening skupu podataka.
* **Treniranje modela:** Kod treniranja modela koristili smo gotove biblioteke, kao što su sklearn. Pre nego što smo počeli sa treniranjem našeg modela, iskoristili smo gotovu funkciju za podelu celokupnog skupa na dva skupa – skup koji će služiti za treniranje **(***trening skup*) i skup koji će služiti za evaluaciju klasifikacionog modela (*test-skupa*), gde smo uzeli 20% za test skup. Potom smo model istrenirali koristeći gotove funkcije iz biblioteke sklearn. Na kraju smo koristili gotovu funkciju kako bi predvideli izlaz našeg modela.
* **Evaluacija modela:** Poslednji korak je evaluacija modela. Skup koji nismo iskoristili služi za proveru tačnosti klasifikacije nad napravljenim modelom. Nad skupom za testiranjem izvršićemo predviđanje pripadnosti klase na osnovu istreniranog modela i uporediti sa njegovom stvarnom klasom. Postoje razne mere za evaluaciju klasifikacionog modela: tačnost klasifikacije, preciznost klasifikacije, odziv klasifikacije, f-mera.

1. Metrike za evaluaciju modela

* **Tačnost klasifikacije** (eng. **accuracy**) predstavlja udeo tačno klasifikovanih instanci u odnosu na ukupan broj instanci i izražava se:



* **Preciznost klasifikacije** (eng. **precision**) predstavlja udeo pozitivnih instanci u svim instancama koje su proglašene pozitivnim, tj.



* **Odziv** (eng. **recall**) je udeo pronađenih pozitivnih instanci u svim pozitivnim instancama:



* **F1-mera** je kombinacija preciznosti i odziva i predstavlja harmonijsku sredinu preciznosti i odziva:



* **TP** predstavlja broj instanci za koje je klasifikator predvideo da pripadaju pozitivnoj klasi, a koje zaista pripadaju toj klasi. **TF** predstavlja broj instanci za koje je klasifikator predvideo da pripadaju negativnoj klasi, a koje zaista pripadaju toj klasi. **FP** predstavlja broj instanci za koje je klasifikator predvideo da pripadaju pozitivnoj klasi, a koje zapravo pripadaju negativnoj klasi, a **FN** predstavlja broj instanci za koje je klasifikator predvideo da pripadaju negativnoj klasi, a koje zapravo pripadaju pozitivnoj klasi.
* Ovi podaci definišu **matricu konfuzije** (eng. **confusion matrix**):



1. Algoritmi korišćeni u projektu
2. **Logistička regresija (Logistic Regression):**

**Logistička regresija** je metoda binarne klasifikacije koja se često koristi za predviđanje pripadnosti jedne od dve moguće klasifikacijske kategorije. Koristi se linearna kombinacija atributa ulaza uz sigmoidnu funkciju kako bi se predvidela verovatnoća pripadnosti određenoj klasi. Za evaluaciju logističke regresije koriste se metrike kao što su tačnost (accuracy), preciznost (precision), odziv (recall) i F1-mera.

1. **Stabla odlučivanja (Decision Trees):**

**Stabla odlučivanja** je metoda koja koristi hijerarhijsku strukturu stabla kako bi se donela klasifikacijska odluka. U svakom čvoru stabla osim listova nalazi se po jedan test. Svaki test ima više od jednog ishoda. Svakom ishodu odgovara jedna grana stabla koja vodi do sledećeg čvora. Listovi su označeni vrednostima koje predstavljaju predviđanja stabla. Evaluacija stabala odlučivanja uključuje metrike poput tačnosti, preciznosti, odziva i F1-mere.

1. **Random Forest (Slučajne šume):**

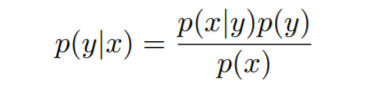
**Random Forest** je metoda koja se temelji na kombinaciji više stabala odlučivanja. Svako stablo se trenira na različitom podskupu skupa podataka, a konačna klasifikacija se izvodi spajanjem rezultata svakog stabla. Random Forest se često koristi zbog svoje sposobnosti da se nosi s visokom varijacijom u podacima. Za evaluaciju Random Forest modela koriste se metrike poput tačnosti, preciznosti, odziva i F1-mere.

1. **Metoda potpornih vektora (Support Vector Machines - SVM):**

**Metod potpornih vektora** (eng. **support vector machine**, **SVM**) jedan je od najčešćih algoritama za klasifikaciju. **SVM klasifikator** razdvaja klase koristeći hiperravni kao granice odlučivanja. SVM pronalazi optimalnu hiperravan koja je podjednako udaljena od najbližih predstavnika obe klase. Ova metoda se oslanja na koncept potpornih vektora, koji su primeri iz skupa podataka koji se nalaze najbliže hiperravni. Za evaluaciju SVM modela koriste se metrike poput tačnosti, preciznosti, odziva i F1-mere.

1. **Naivni Bayes klasifikator (Naive Bayes Classifier):**

**Naivni Bajesov** klasifikator se najčešćeprimenjuje na problem klasifikacije i zasniva se na Bajesovoj teoremi:



Za dato x od svih ishoda y bira se onaj sa maksimalnom verovatnoćom p(y/x). Ovaj algoritam pretpostavlja da je vrednost nekog atributa u klasi potpuno nezavisna od vrednosti ostalih atributa u toj klasi. Za evaluaciju Naivnog Bayes klasifikatora koriste se metrike poput tačnosti, preciznosti, odziva i F1-mere.

1. **KNN:**

Algoritam K-najbližih suseda, često skraćen k-nn, pristup je klasifikaciji podataka koji procenjuje koliko je verovatno da će tačka podataka biti član jedne ili druge grupe u zavisnosti od toga u kojoj su grupi najbliže tačke podataka.

K-najbliži sused je primer algoritma „lenji učenik“, što znači da ne gradi model pomoću skupa treninga dok se ne izvrši upit za skup podataka.

1. Prikaz rezultata

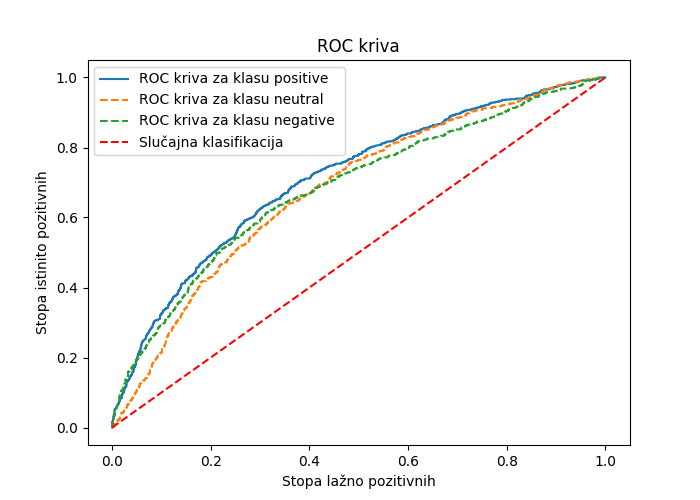
* **Tabelarni prikaz za trening skup:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Metoda** | **Accuracy** | **Recall** | **Precision** | **F-measure** |
| **Logistička** **regresija** | 0.9333 | 0.9333 | 0.9335 | 0.9333 |
| **Random** **Forest** | 0.6121 | 0.6118 | 0.6378 | 0.6107 |
| **SVM** | 0.8618 | 0.8618 | 0.8716 | 0.8628 |
| **Naivni** **Bayes** | 0.8787 | 0.8790 | 0.8834 | 0.8785 |
| **K-najblizih** **suseda** | 0.4955 | 0.4942 | 0.6194 | 0.4440 |

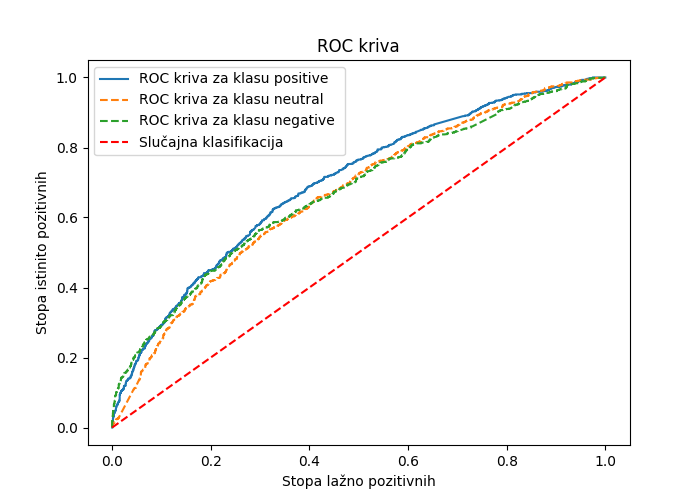
* **Tabelarni prikaz za test skup:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Metoda** | **Accuracy** | **Recall** | **Precision** | **F-measure** |
| **Logistička** **regresija** | 0.5163 | 0.5162 | 0.5193 | 0.5162 |
| **Random** **Forest** | 0.484 | 0.4853 | 0.5005 | 0.4809 |
| **SVM** | 0.5183 | 0.5185 | 0.5207 | 0.5183 |
| **Naivni** **Bayes** | 0.4977 | 0.4959 | 0.5001 | 0.4862 |
| **K-najblizih** **suseda** | 0.383 | 0.3823 | 0.4192 | 0.3341 |

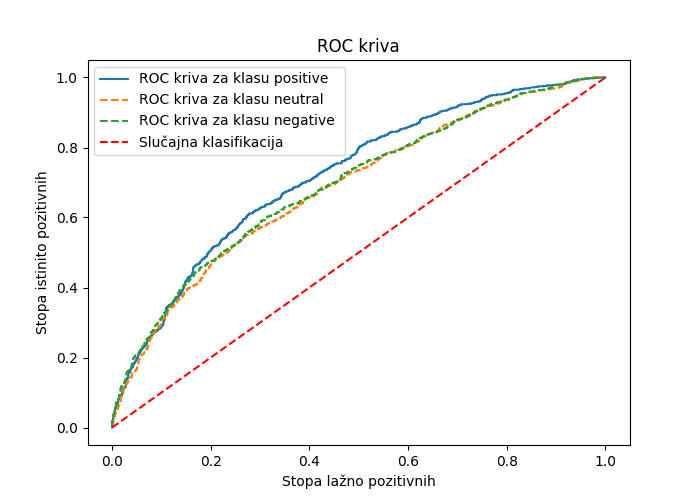
* **Logistička regresija:**



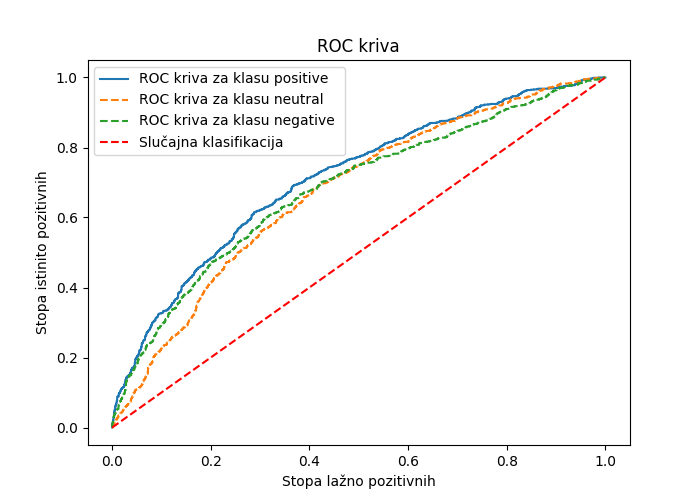
* **Random forest:**



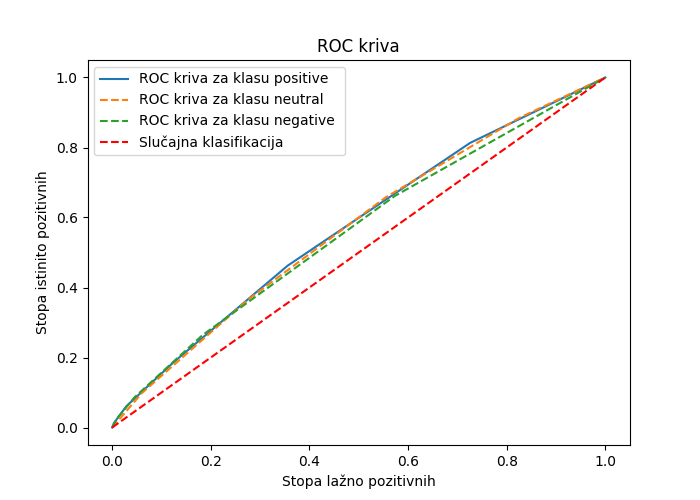
* **SVM:**



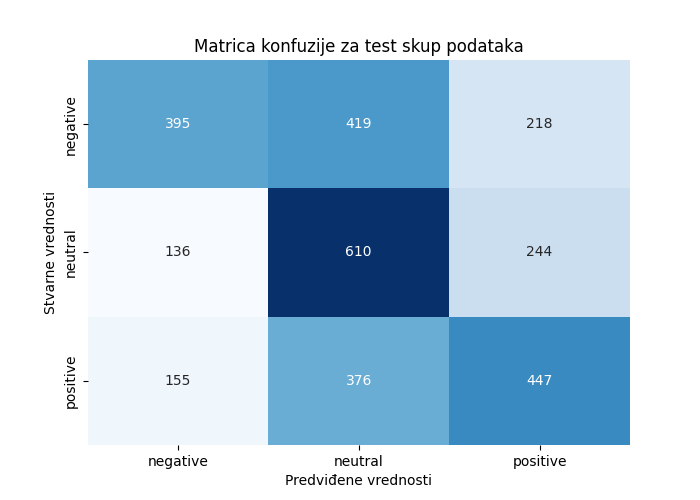
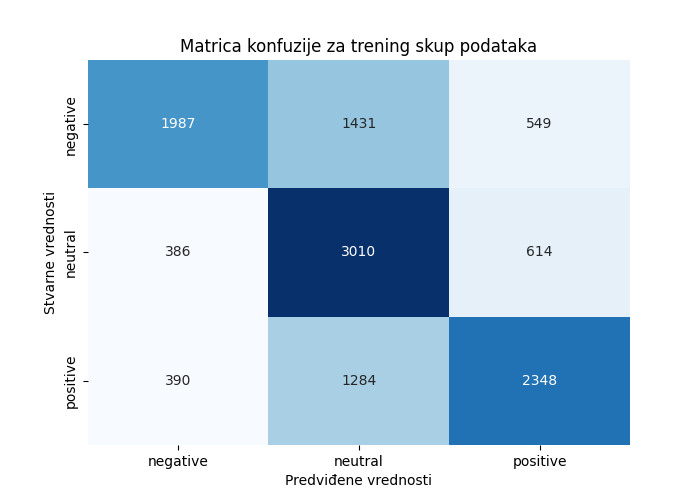
* **Naivni Bajes:**



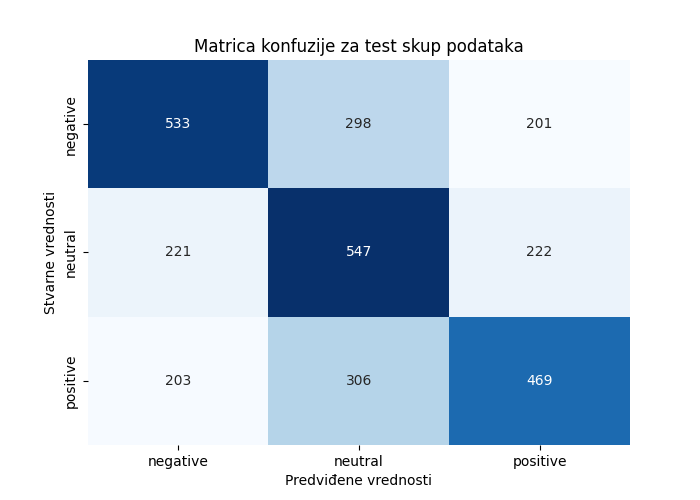
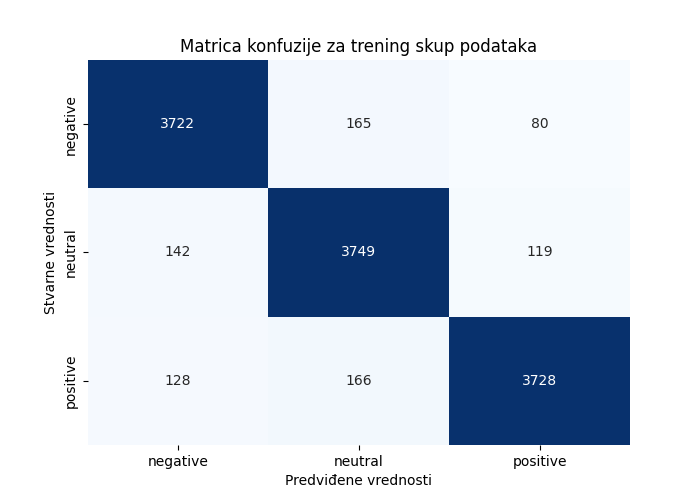
* **KNN:**



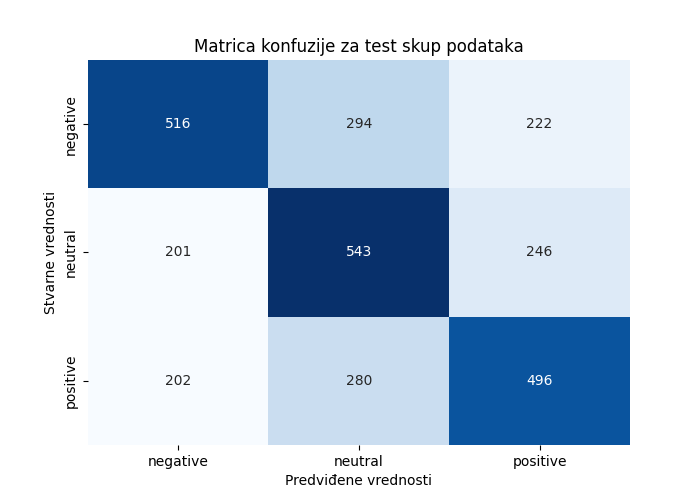
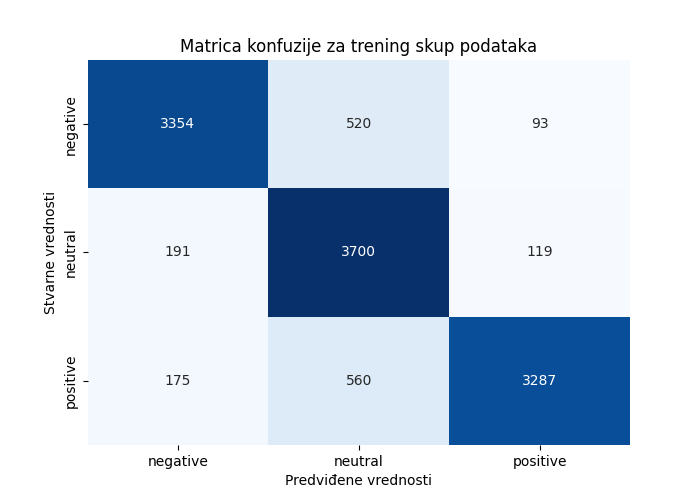
* **Random forest:**



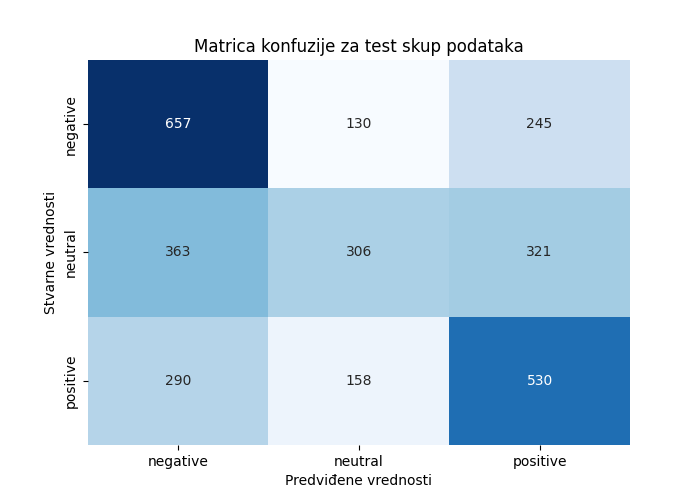
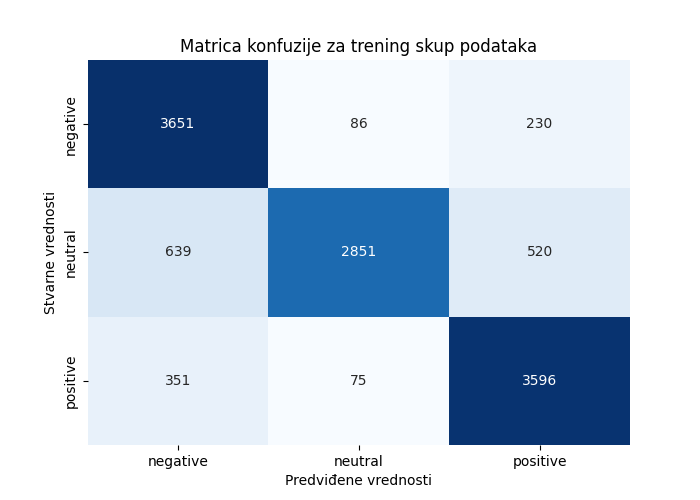
* **Logistička regresija:**



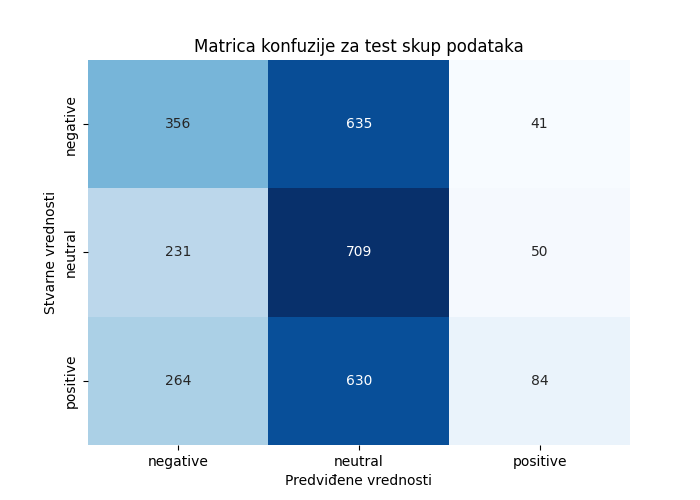
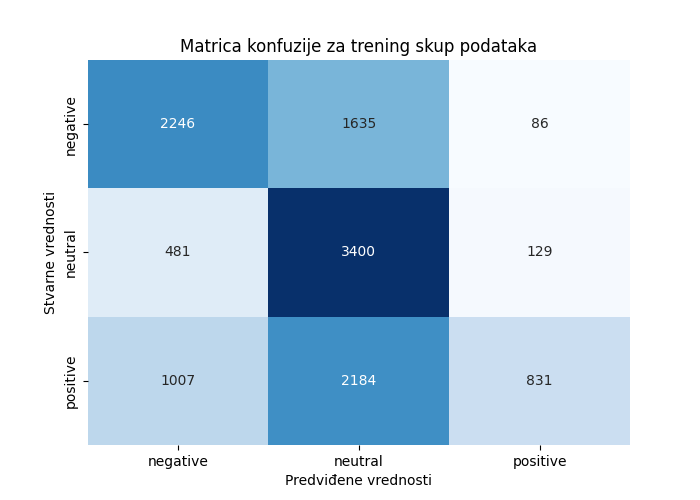
* **SVM:**



* **Naivni Bajes:**



* **KNN:**



## Analiza rezultata

* **Najbolji rezultati:**

Najbolje rezultate su postigao algoritam logističke regresije, zatim slede algoritmi Naivnog Bayesa i SVM sa približnim rezultatima. Logistička regresija je ostvarila visoku tačnost, preciznost, odziv i F1-meru na test skupu podataka, što ukazuje na njenu sposobnost da precizno klasifikuje instance.

* **Najmanje pouzdani rezultati:**

Najmanje pouzdane rezultate je dao algoritam K-najbližih suseda. Niže vrednosti tačnosti, preciznosti, odziva i F1-mere na test skupu sugerišu da ovaj algoritam može imati izazove u pravilnom generalizovanju naučenih obrazaca na nove podatke. Ovi rezultati ukazuju na potrebu za dodatnom optimizacijom ili možda razmatranjem drugih algoritamskih pristupa kako bi se poboljšale njegove performanse.

1. Zaključak

Prepoznavanje emocija je složen proces koji je često komplikovan i za ljude, a posebno za računare na kome se još uvek radi.

Na osnovu analize rezultata različitih algoritama, zaključak je da postoje značajne razlike u njihovim performansama u rešavanju datog problema. Logistička regresija se istakla kao algoritam sa najboljim rezultatima, postižući visoku tačnost, preciznost, odziv i F1-meru. Ovaj algoritam je pokazao sposobnost efikasne klasifikacije i generalizacije na nove podatke. S druge strane, K najbližih suseda je imao najmanje pouzdane rezultate, sa najnižim ocenama u svim merama performansi. Važnost odabira algoritma leži u specifičnim potrebama i karakteristikama problema, a dalja optimizacija ili kombinacija algoritama može doprineti boljim performansama u rešavanju ovog problema.

1. Literatura

* M. Nikolić, A. Zečević : Mašinsko učenje, Matematički fakultet, Beograd 2019
* <http://moodle.np.ac.rs/course/view.php?id=564> – Predavanja, Vežbe
* https://www.kaggle.com/pashupatigupta/emotion-detection-from-text