**Fakultet tehničkih nauka**

**Milan Rodić RA 13-2022**

**Projekat iz predmeta Softverski alati u sistemima automatskog upravljanja**

***ZADATAK/CILJ PROJEKTA***

Zadat je Dataset u obliku csv fajla koji sadrži 39645 ulaznih parametara i 58 atributa. Podaci opisani u ovom fajlu predstavljaju članke objavljene na platformi Mashable u periodu od dve godine. Dataset ne sadrži originalni tekst već numeričke i binarne podatke dobijene iz sadržaja, naslova, metapodataka i LDA tematskog modeliranja. Izuzetak predstavlja URL o kojem će biti reči u daljem sadržaju.

Svi atributi i njihova značenja dati su u podnožju:





Zadatak projekta je da se na osnovu navedenih parametara odredi očekivan broj deljenja članka na društvenim mrežama. Dakle, target će nam biti poslednji atribut – shares.

***IZRADA***

Automatski vidimo da se radi o predviđanju neke kontinualne vrednosti pa zaključujemo da je u pitanju model regresije. Za pripremu i obradu modela pratićemo uputstvo linkovano na Drive-u

1. *Početno procesiranje podataka*

Već kao prvi korak pri obradi podataka izbačen je atribut URL. Razlog za to je taj što je URL jedinstven za svaki od članaka koji se obrađuju u Datasetu. Kao takav, on se može enkodirati jedinstvenim brojem od 1 do 39645 ali je jasno da nam ti brojevi neće ni na koji način doprineti daljem treniranju i obradi, već samo potencijalnom referenciranju specifičnih članaka.

Enkodiranje podataka nije potrebno jer su nam sva polja numeričkog tipa. Izuzetak je URL ali smo se njega rešili pređašnjim korakom.

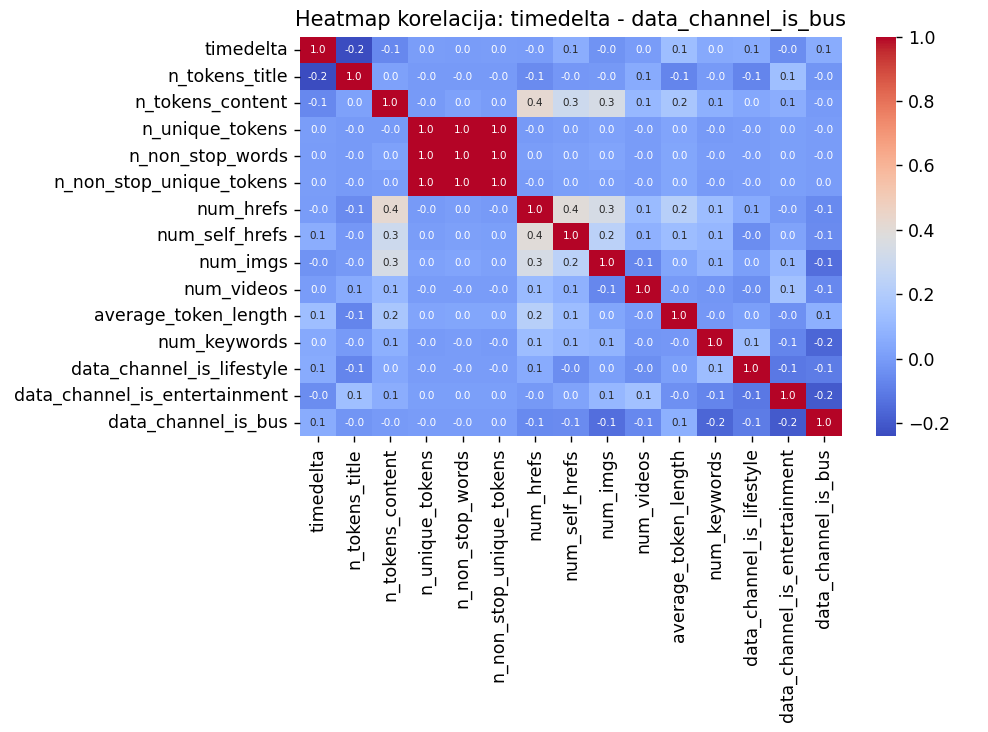
Prolaskom kroz sam dataset, potrebno je da proverimo da li ima rupa u njemu kao i obraditi to šta treba da se desi kada na njih naiđemo. U našem konkretnom primeru nemamo nijednu ali je potrebno da se pripremimo za slučaj da su dodata polja koja ih sadrže. Ovo rešavamo popunjavanjem praznih polja sa medijanom.

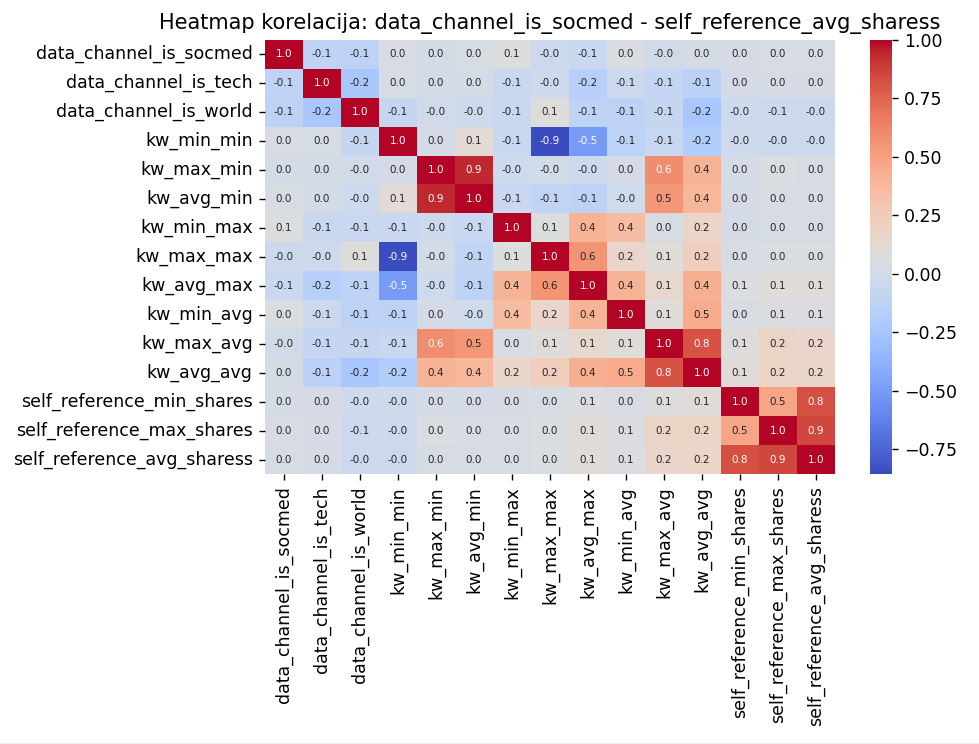
Uradićemo log-transformaciju ulaznih podataka kako bismo otklonili šum i dobili stabilniji model.

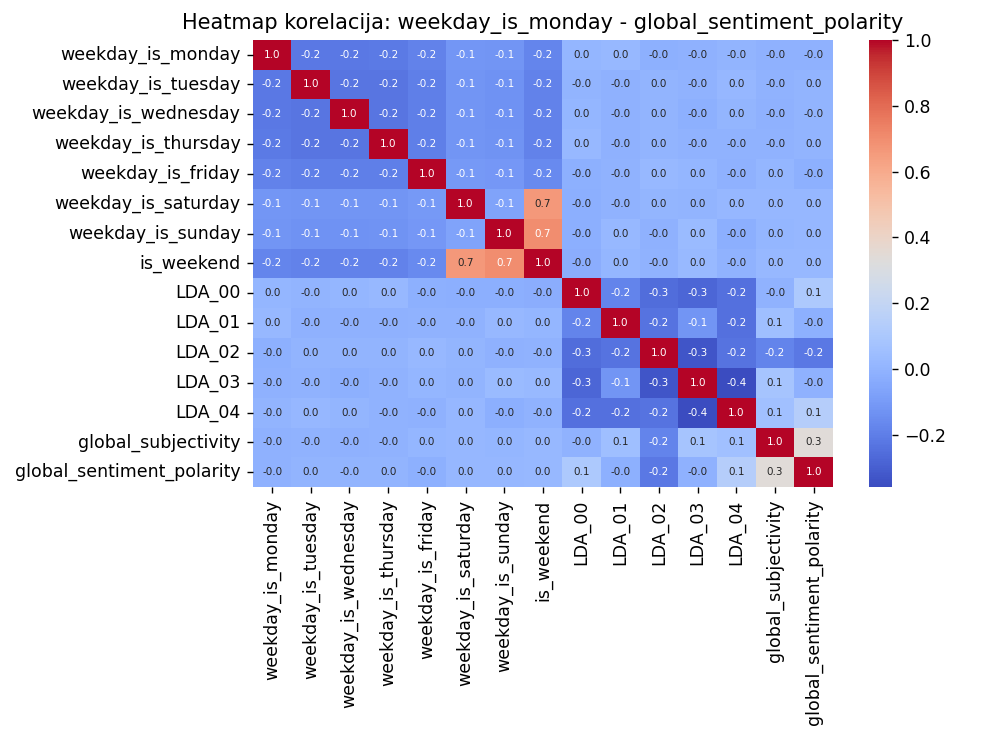
Kao poslednji korak početnog procesiranja, proveravamo fizičku smisao ulaznih podataka. Tipovi podataka dele nam se na brojeve, udele i binarne 0 i 1 vrednosti. Za cele brojeve se proverava da li su deljivi sa brojem 1, udeli su definisani tako da ne smeju biti veći od 1 i manji od 0 i binarne vrednosti ne smeju imati vrednost van diskretnog skupa 0 i 1. Ukoliko se i jedna od vrednosti flaguje kao neispravna, ceo red je potrebno obrisati. U našem datasetu nema neispravnih vrednosti.

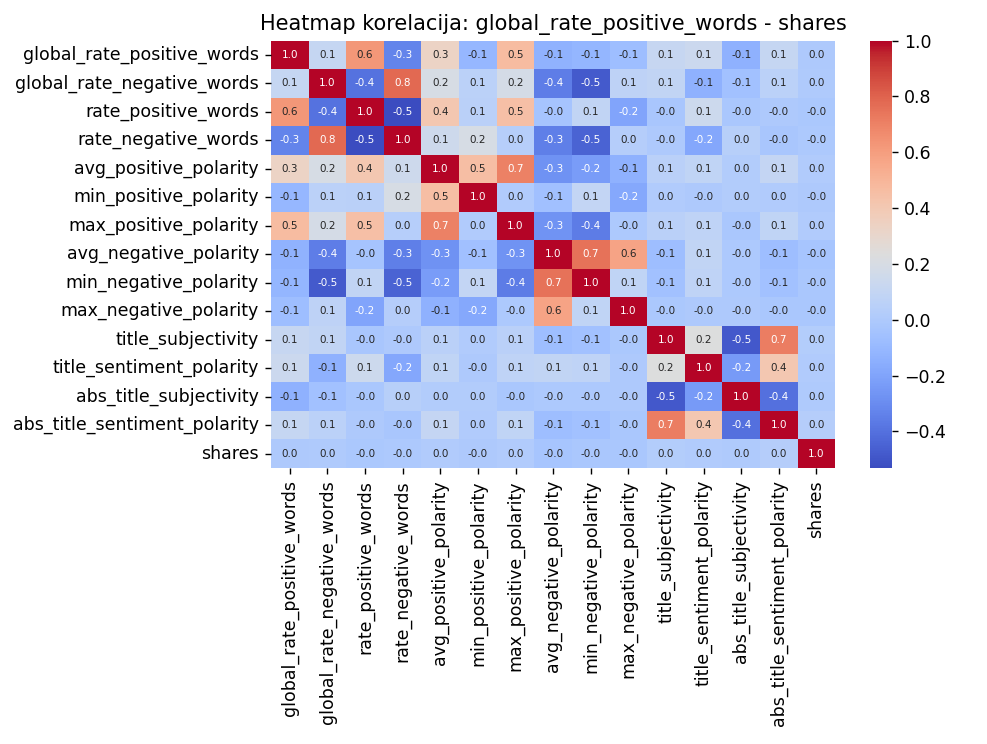
1. *Eksplorativna analiza skupa*

Proveravamo korelaciju svakih od atributa modela. Najpre je moguće izvršiti formiranje korelacione matrice nad celim datasetom, međutim, primetićemo da je izlazni plot veoma nečitljiv. Rešenje za ovu nečitljivost pronalazimo u razdeljivanju dataseta na celine od po 15 atributa po plotu.









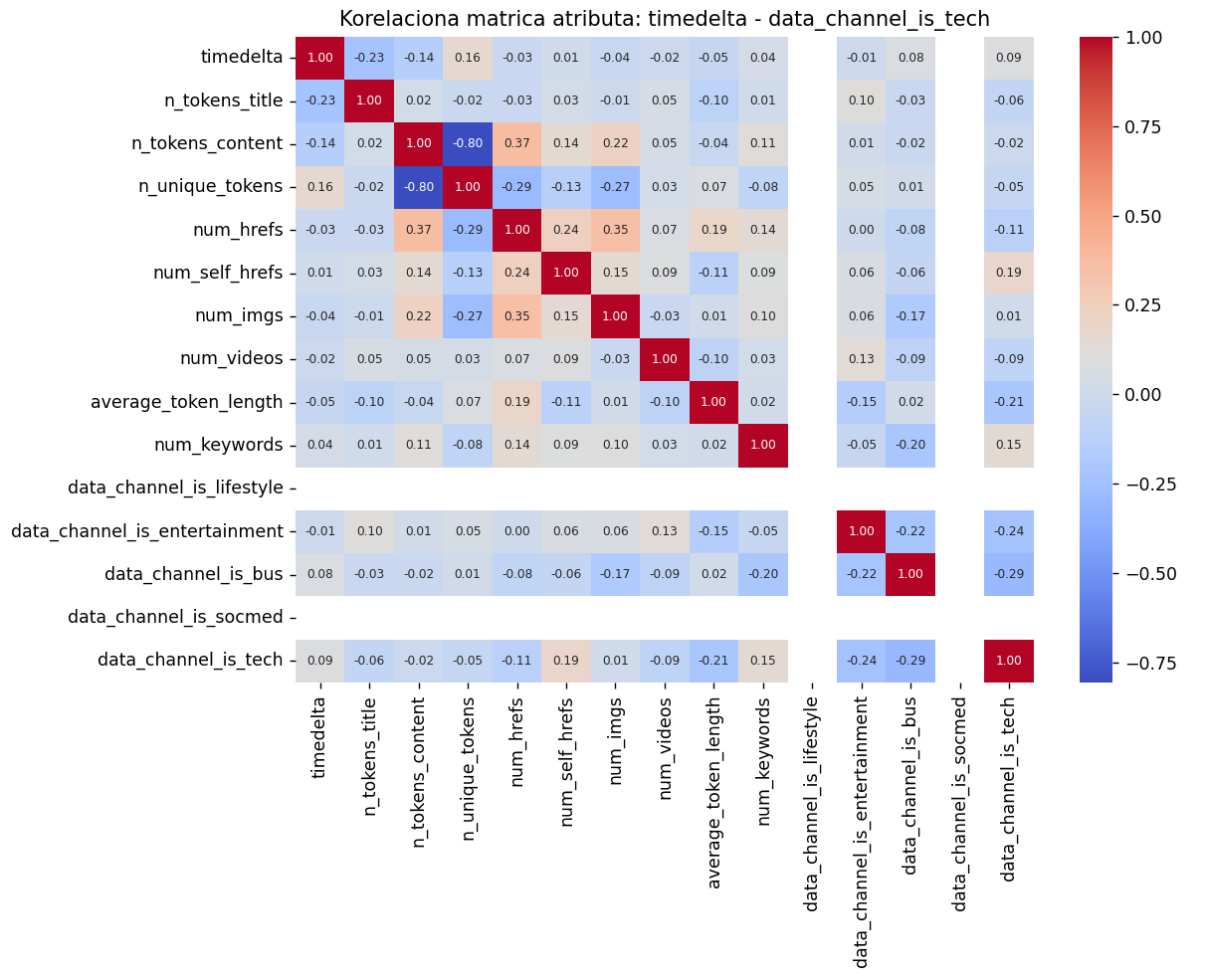
\

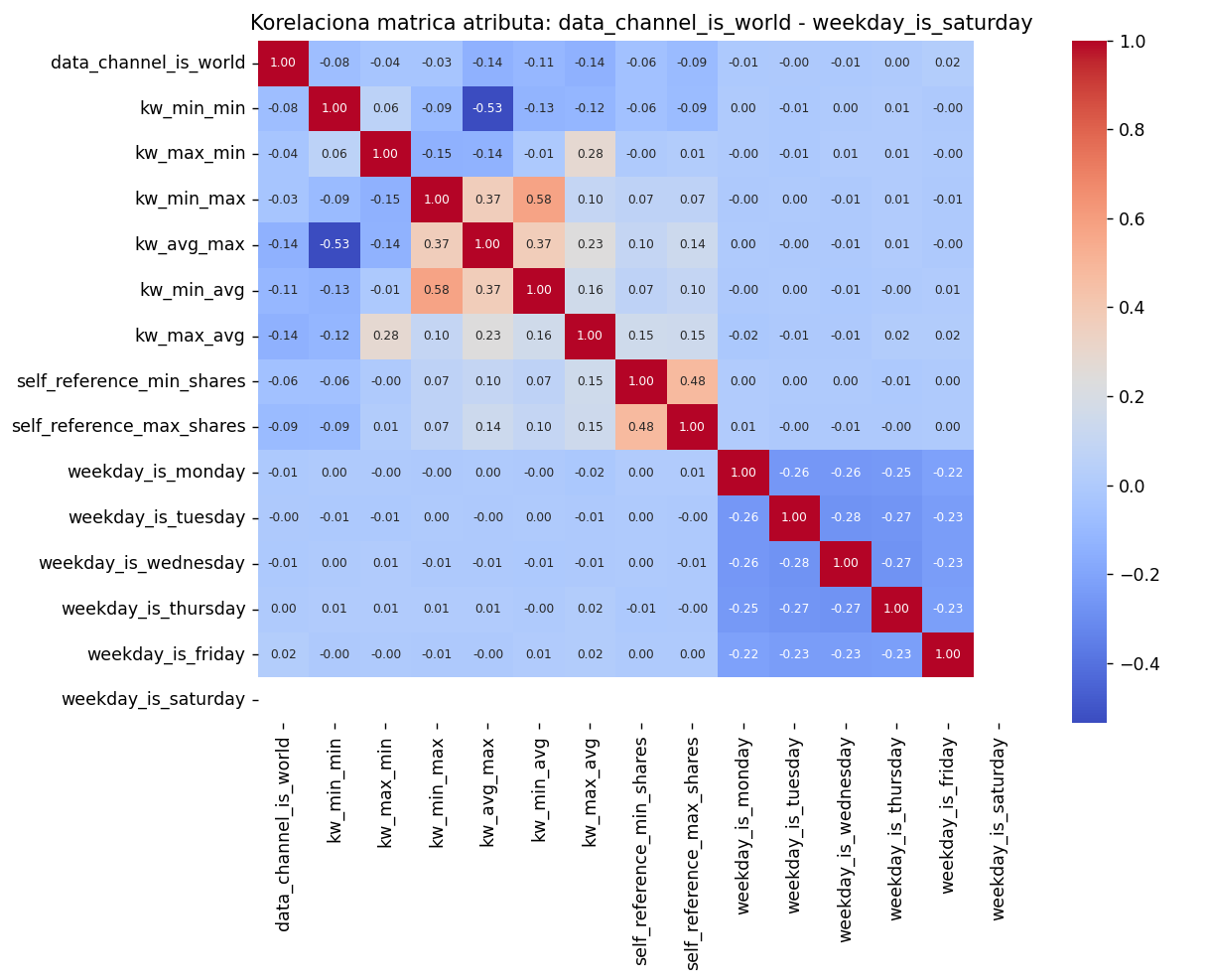
Odmah možemo primetiti da se na njemu nalaze podaci koji su visoko i inverzno korelirajući. Definisaćemo da je prag za izbacivanje ovakvih vrednosti 0.8 po apsolutnoj vrednosti.

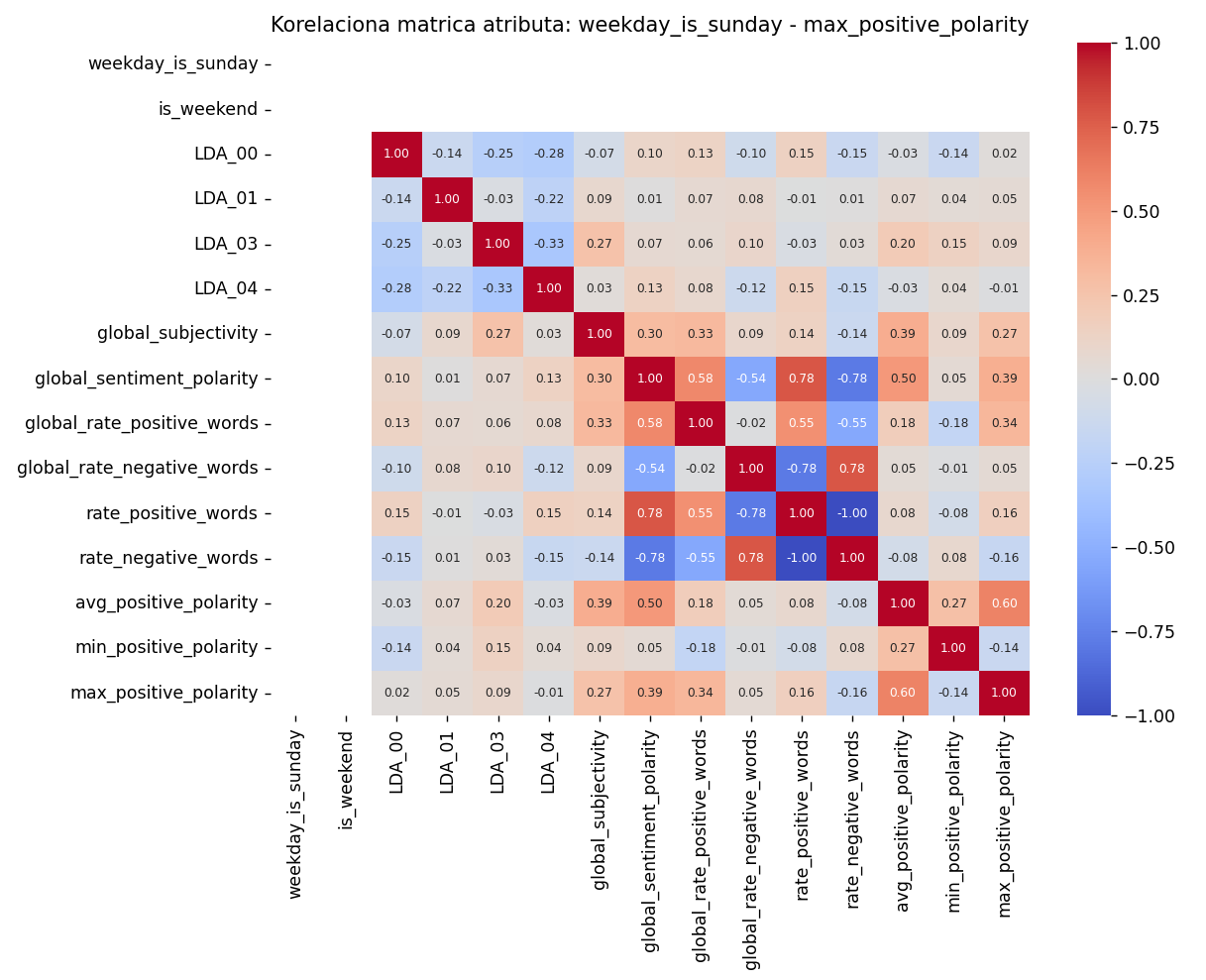
Dalje skraćivanje vršimo pomoću z score-a, odnosno, formule standardne devijacije.

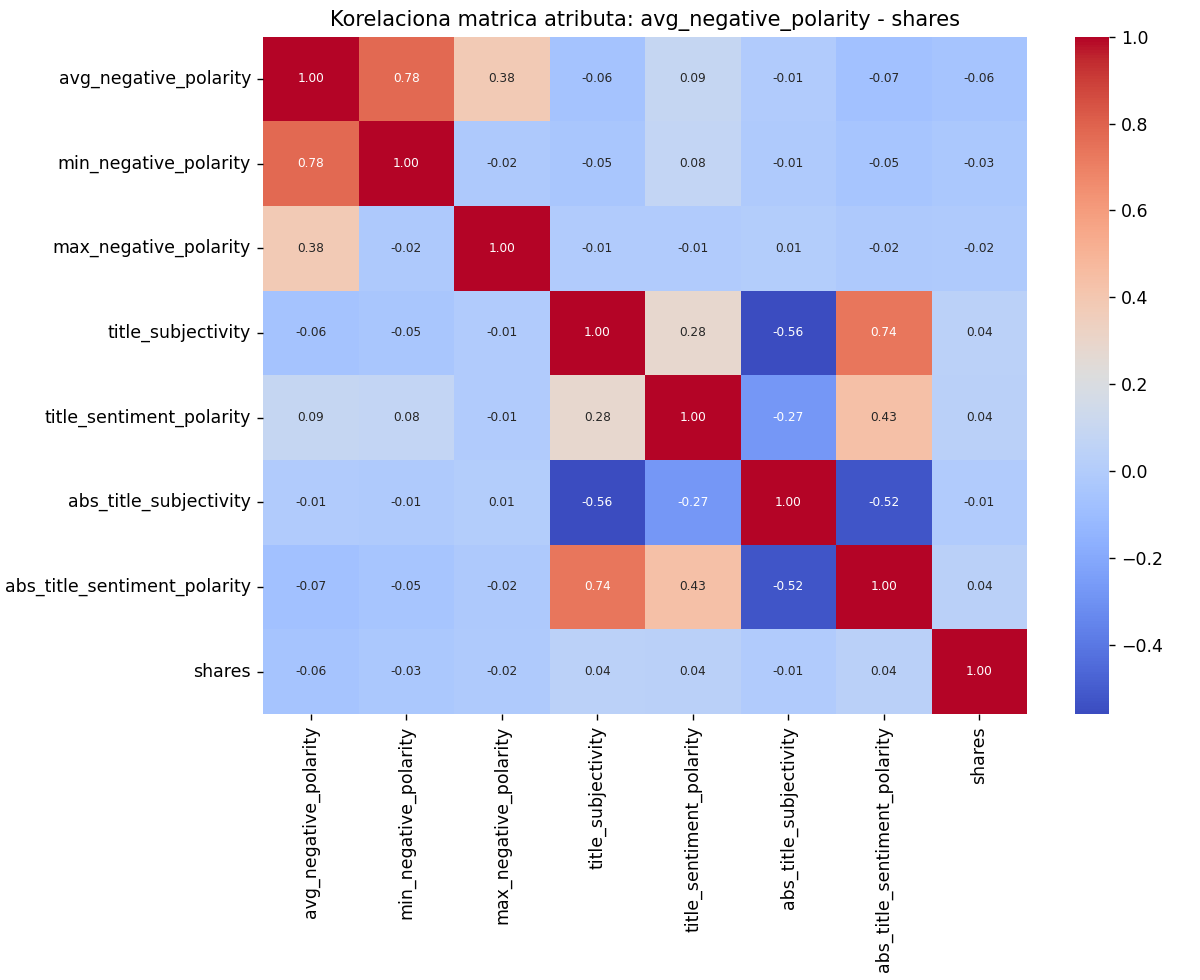
U narednom koraku vršimo skaliranje podataka tako da imaju srazmeran uticaj na predikciju targeta i ne bivaju zanemareni. Za dodatno smanjenje dimenzionalnosti koristimo PCA.

Uporedićemo sada kako izgleda matrica korelacije ulaznog i izlaznog seta









1. *Odabir i treniranje modela, podešavanje hiperparametara i unakrsna validacija*

Postavljamo shares atribut kao target i u odnosu njega vršimo predviđanje. Formiraćemo šablon kako ne bismo morali stalno da splitujemo na trening i test set podataka. Postavljamo trening set da bude na 80%, a testni na 20%. Random state nam je 42. Modeli nad kojima smo trenirali podatke su: Linearni modeli(Linear, Lasso, Ridge), Decision tree regression, Gradient Boost, XGBoost.

Za Ridge i za Lasso koristimo logaritamske skale sa korakom 50 kako bi pronašli optimalno alfa. Ridge prolazi kroz vrednosti od 10^-4 do 10^4, dok Lasso prolazi kroz vrednosti 10^-4 do 10.

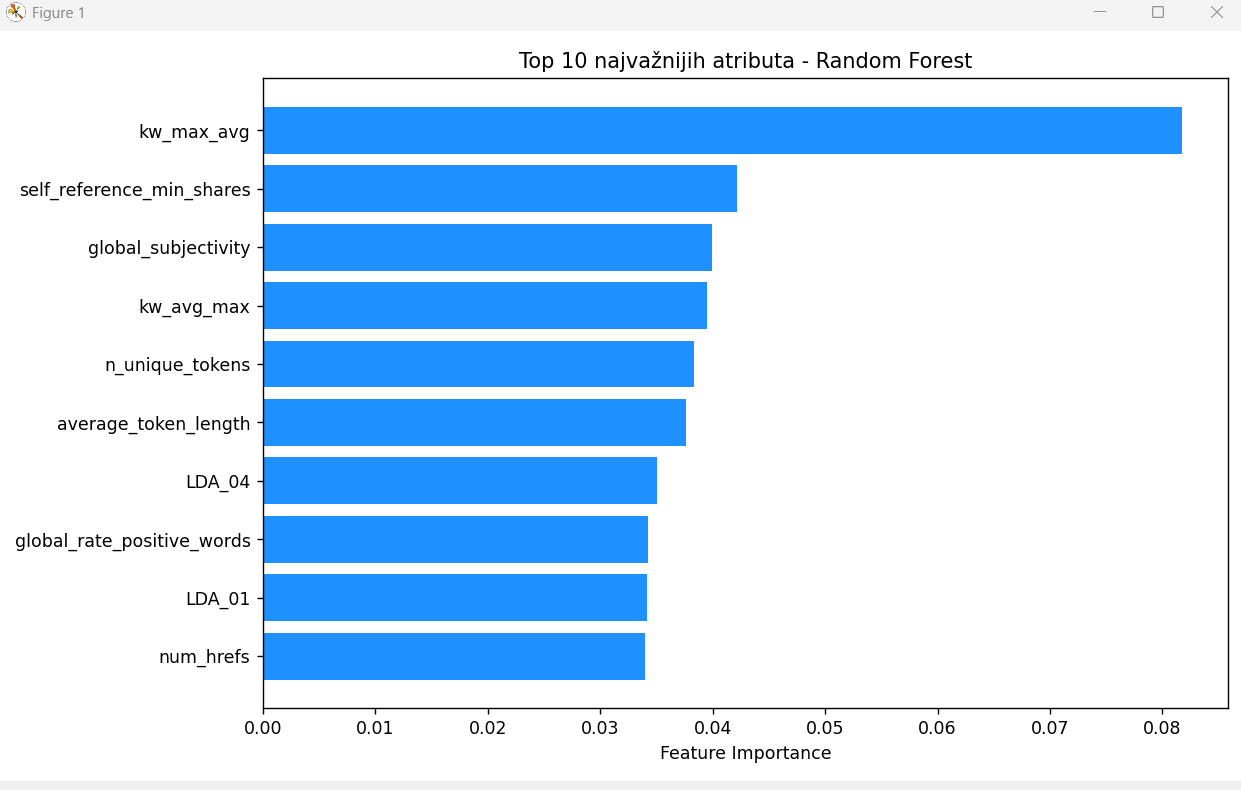
Za Decision tree regresor koristimo random state 42, za max depth uzimamo vrednosti u diskretnom skupu od: 3, 5, 10, 20 i None. Za min\_samples\_split uzimamo vrednosti od: 2, 5, 10, dok za min\_samples\_leaf uzimamo vrednosti: 1, 2 i 4. Korisitmo 5-fold cross validaciju.

Za Random forrest regresor koristimo random state 42, za n\_estimators uzimamo vrednosti 100 i 200. Za max\_depth uzimamo vrednosti od: 10, 20 i None, za min\_samples\_split uzimamo vrednosti: 2 i 5, dok za min\_samples\_leaf uzimamo vrednosti 1 i 2. Korisitmo 5-fold cross validaciju.

Za Gradient Boosting koristimo random state 42, za n\_estimators uzimamo vrednosti 100 i 200, learning rate nam je 0.01, 0.1, 0.2, za max\_depth uzimamo vrednosti od: 3, 5 i 10, za min\_samples\_split uzimamo vrednosti: 2 i 5, dok za min\_samples\_leaf uzimamo vrednosti 1 i 2. Korisitmo 5-fold cross validaciju.

Za XGBoost koristimo random state 42, za n\_estimators uzimamo vrednosti 100 i 200, learning rate nam je 0.01, 0.1, 0.2, za max\_depth uzimamo vrednosti od: 3, 5 i 10, za subsample uzimamo vrednosti 0.8 i 1, dok za colsample\_bytree uzimamo vrednosti 0.8 i 1. Korisitmo 5-fold cross validaciju.

1. *Odabir najbitnijih atributa*

Ispisaćemo 10 najbitnijih atributa upotrebom random forresta. Rezultati biće predstavljeni na grafikonu. Iskoristićemo ovih 10 najboljih atributa za treniranje najboljeg i najgoreg dobijenog nelinearnog modela – Decision Tree i XGBoost. 

***ANALIZA REZULTATA***

Svi pokušani modeli pokazali su izuzetno slabe performanse pri zadatku predviđanja ciljne kolone. Što se tiče treniranja nad obrađenim datasetom poražavajuća je činjenica da su se u proseku najbolje pokazali linearni modeli. Linearna regresija doprinela je R2 skor vrednoti 0,1093. Ridge regresija je nešto bolja sa R2 skorom od 0,1094 dok se na spektru svih modela, uključujući i napredne Lasso pokazao kao najbolji sa R2 skorom od 0,1095.

Što se tiče naprednijih modela, najgore performanse pokazao je decision tree regresor sa R2 skorom od 0,0619, sledeći najgori model bio je model random forrest regresije sa R2 skorom od 0,1073. Nakon njega, pokušan je model Gradient boosting sa R2 skorom od 0,1116. Ovaj model pokazao je i najgore performanse što se tiče vremena izvršavanja. Poslednji i neslavno najbolji model od svih jeste XGBoost sa R2 skorom 0,1219

Uzećemo i deset najboljih parametara kao ulaz i trenirati ih na najboljem i najgorem modelu: Decision tree regresoru i XGBoost-u.

Za ažuriranu Decision tree regresiju dobijamo vrednost -0,8747 što znači da model predviđa ciljnu kolonu gore nego kad bi je pokušali predvideti prosečnom vrednošću. Vrednost R2 za XGBoost sada postaje 0.1193.