Support Vector Machines

Tim mašinski_učenjaci:

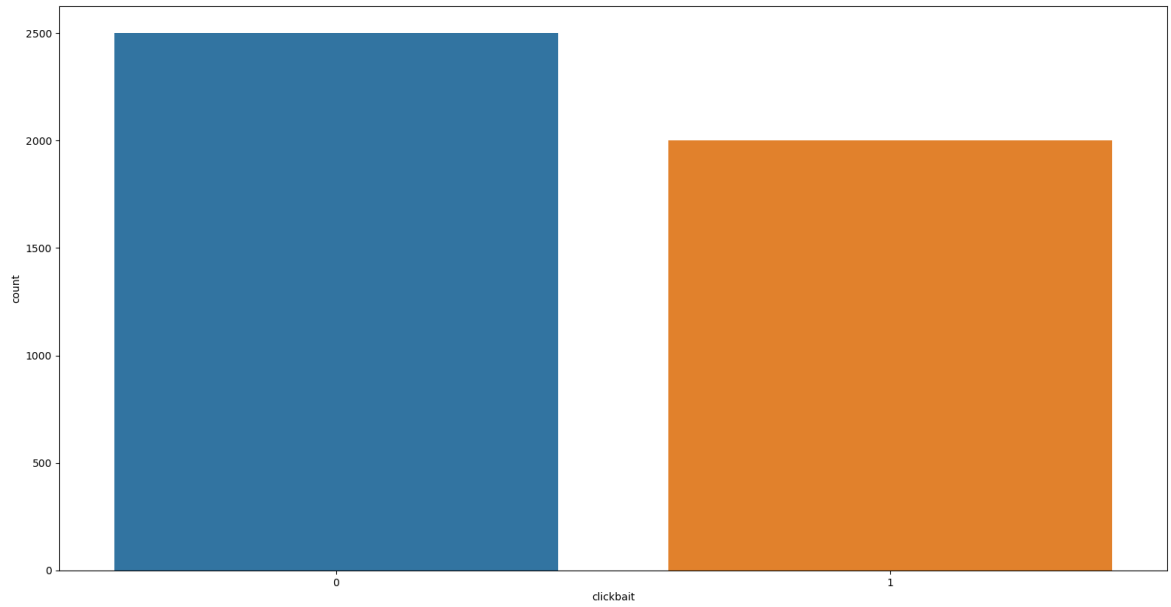
Nikola Zubić (SW-03/2016)
Filip Mladenović (SW-11/2016)
Mihajlo Kušljić (SW-53/2016)

Opis problema:

Klasifikacija naslova onlajn medijskih članaka na engleskom jeziku (ulazna vrijednost: *text*) u dvije klase (izlazna vrijednost: *clickbait*): 0 – naslov nije klikbejt, 1 – naslov jeste klikbejt. Zadatak se smatra uspješno urađenim ukoliko se na kompletnom test skupu podataka dobije tačnost (*accuracy*) veća od 0.89 (89%). Zadatak je potrebno rješiti isključivo upotrebom SVM klasifikatora (*LinearSVC* ili *SVC*).

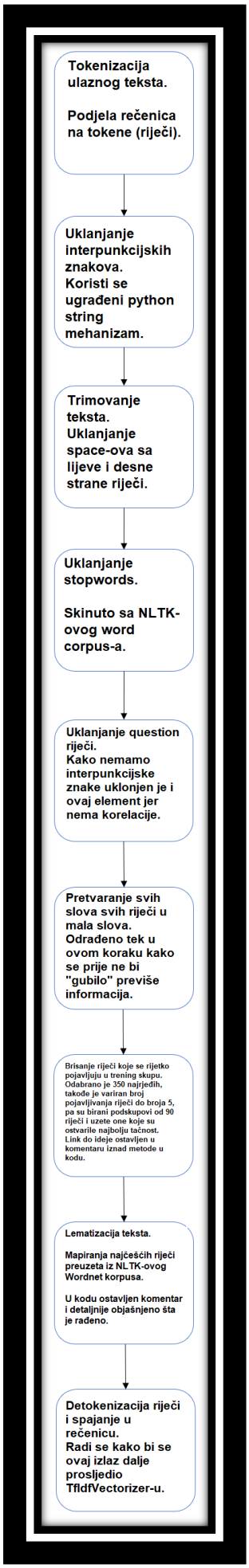
Naš pristup problemu:

Prvo smo koristeći *seaborn* biblioteku (*countplot*) plotovali odnos između *clickbait* i *non-clickbait* naslova na trening skupu kako bi vidjeli koliko kojih ima, jer je u predavanju *"Praktični saveti za primenu ML"* izloženo da može biti problem ako imamo neravnomjernost u izlaznim klasama kod validacije, tj. morali bi dodatno da vodimo računa kako ćemo izvršiti podjelu na *train* i *test* skupove.



Slika 1: Veličina trening skupa je 4500 i odnos* non_clickbait/clickbait *je 0.55

Korišćena je *Stratified K Fold* kros validacija zato što nam je cilj da generišemo test skupove koji imaju istu/sličnu distribuciju klasa, te specijalno jer je problem binaran (K = 10).



Slika 2: Zamišljena Text Transformer klasa za pretprocesiranje teksta

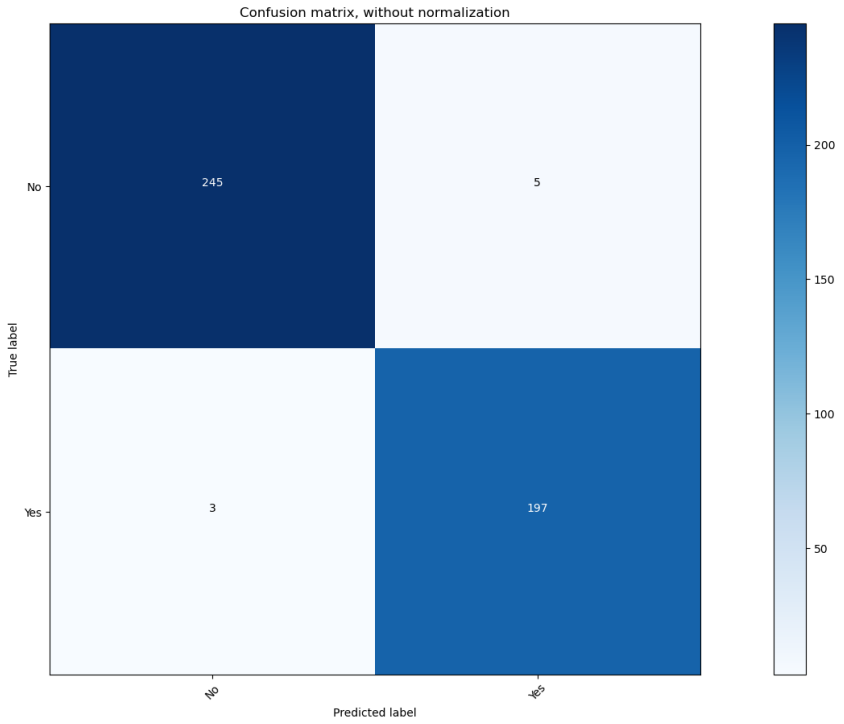
Nakon što se izvrši *Text Transformer pipeline* naveden na [***Slika 2***], njegov izlaz se prosljedi *TfIdfVectorizer*-u koji konvertuje date rečenice u numeričke podatke i na kraju se ti podaci prosljede linearnom *SVM* klasifikatoru.

Isprobani algoritmi:

- *Text Transformer* (klasa zadužena za *text preprocessing* i *feature engineering*)
- *TfIdfVectorizer* i *Bag of Words*
- *LinearSVC* i *SVC* (kerneli: *poly* sa stepenima od 3 do 8, *rbf*, *sigmoid*)

Konačno (odabrano) rješenje:

Najbolje rješenje ostvareno je koristeći *Text Transformer pipeline* naveden u sekciji iznad, potom uz *TfIdfVectorizer* i *LinearSVC*. Za K = 10 **accuracy** je **0.9822222**. Nad test_preview skupom dobijen je **accuracy**: **1.0**.



Slika 3: Matrica konfuzije nad odabranim rješenjem koristeći Stratified K Fold Cross Validation za K = 10