Support Vector Machine

Tim:

- **Matija Aleksić**

Zadatak:

Klasifikovati naslove onlajn medijskih članaka na engleskom jeziku (text) u dve klase (clickbait): 0 - naslov nije klikbejt, 1 - naslov jeste klikbejt. Zadatak je uspešno urađen ukoliko se na kompletnom testnom skupu podataka dobije mikro **f1_mera** (micro f1 score) > 0.90. Zadatak se rešava upotrebom SVM klasifikatora.

Analiza podataka:

Nakon analize ulaznih podataka došli smo do zaključka da je potrebno izvršiti pretprocesiranje podataka. Napravljene su četiri transformacije ulaznih podataka koje su se odnosile i na trening i na test skup:

- Svođenje teksta na tokene
  Svaki naslov članka se rastavio na niz tokena koji predstavljaju jednu reč. Ovo smo uradili da bi mogli da svaku reč zasebno analiziramo, da mozemo da izbacimo neke reči koje nam nisu bitne za raspoznavanje clickbait-a, sto samim tim dovodi do preciznijeg rezultata.

- Svođenje teksta na mala slova
  Celokupan sadržaj svakog naslova je transformisan u mala slova, kako bi ulaz bio case insensitive. Ukoliko se ne radi ova transformacija, dimenzionalnost problema je znatno veća i tačnost modela manja jer se reči “in”, “In” i “IN” tretiraju kao različite.

- Brisanje svih znakova interpunkcije u tekstu
  U ovome procesu brisemo svaki znak interpunkcije da bi izbegli nepotrebno povećanje dimenzionalnosti problema.

- Uklanjanje reči bez značenja (stopwords)
  Jedan od najbitnijih aspekata pretprocesiranja podataka je uklanjanje reči bez značenja koje mogu znatno uvećati dimenzionalnost problema, a ne smatra se da su bitna obeležja za donošenje odluke.

```
In [9]: processed_text.count_words()
Out[9]: [('in', 673),
         ('to', 651),
         ('the', 529),
         ('of', 507),
         ('you', 428),
         ('a', 404),
         ('for', 300),
         ('and', 273),
         ('on', 258),
         ('your', 215),
         ('are', 198),
         ('this', 185),
         ('is', 181),
         ('at', 160),
         ('with', 156),
         ('that', 151),
         ('new', 121),
         ('from', 113),
         ('will', 107),
```

Metodom **count_words()** možemo da vidimo da ima dosta reči koje je potrebno izbaciti, jel ne doprinose preciznosti rezultata. Broj reči koje smo izbacili je 187. Primer reči koje su izbacivane su: 'i', 'me', 'my', 'myself', 'we', 'our', itd.

Pristup rešavanja problema:

Vektorizacija:

Kako smo do sada radile sa tekstualnim zapisom, a SVM kao ulaz prepoznaje numeričke vrednosti, bitan korak predstavlja vektorizacija ulaza. Obučavanje vektorizatora odrađeno je samo na trening skupu (nakon podele na trening i validacioni), dok je validacioni skup samo transformisan korišćenjem obučenog vektorizatora. Za potrebe vektorizacije ulaznih podataka odabrao sam pristup "**TF-IDF**".

Model:

Nad pretprocesiranim i vektorizacije ulaznih podataka potrebno je iskoristiti SVM klasifikator. Izabrani model klasifikatora koji smo koristili da bi dobili ostvarene rezultate je **LinearSVC** sa vrednošću 0.35 za C.

Rezultati:

Za evaluaciju koristili smo f1 metrika. Za validaciju je korišćeno 20% od datog skupa podatka i dobijena je tačnost od 0.982. Na platformi je izvršeno treniranje nad celim trening skupom, a zatim je nad testnim skupom dobijena tačnost od **0.9455**.