Klasifikacija teksta

Seminarski rad u okviru kursa
Istraživanje Podataka 1
Matematički Fakultet

Petar Košanin

Avgust 2019.

Sadržaj

1 Uvod

"IMDB-sentiments"(Maas et al., 2011) je skup podataka napravljen za binarnu klasifikaciju teksta. Sastoji se od trening skupa i test skupa gde test skup čine nelabelirani podaci. Trening skup sadrži 25000 filmskih kritika. Svaka instanca se sastoji od kritike i labele, gde labela uzima vrednost iz skupa $\{0, 1\}$ i to 0 za pozitivnu kritiku, 1 za negativnu. Podaci se mogu pronaći na https://www.kaggle.com/jcblaise/imdb-sentiments.

2 Vizuelizacija podataka

Tabela 1 predstavlja kako izgleda skup podataka. Radi lakšeg prikaza, izabrane su najkraće kritike.

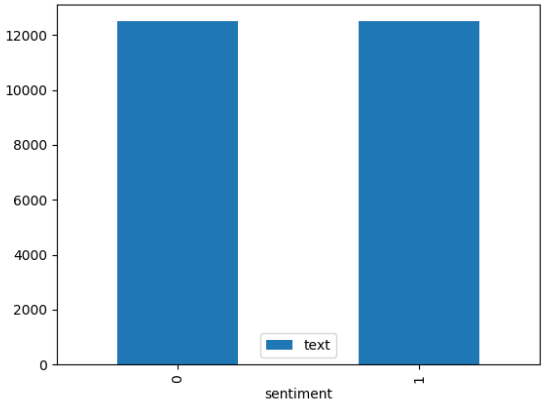
text	sentiment
This movie is terrible but it has some good effects.	1
I wouldn't rent this one even on dollar rental night.	1
Ming The Merciless does a little Bardwork and a movie most foul!	1
You'd better choose Paul Verhoeven's even if you have watched it.	1
Adrian Pasdar is excellent is this film. He makes a fascinating woman.	0

Tabela 1: Isečak trening skupa podataka

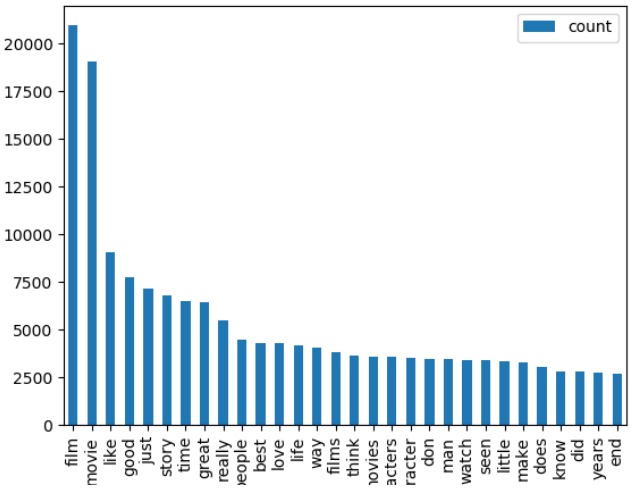
Na slici 1 prikazan je trakasti grafikon(eng. Bar Plot) koji predstavlja raspodelu po klasama i zaključujemo da je skup podataka balansiran. Slike 2 i 3 predstavljaju 30 najčešćih reči u pozitivnim i negativnim kritikama, redom.

Stop reči su reči koje se često javljaju u tekstu, ali ne nose posebno značenje. U engleskom jeziku to su the, is, a... Te reči se uglavnom eleminišu iz skupa podataka kako ne bi negativno uticale na klasifikatore.

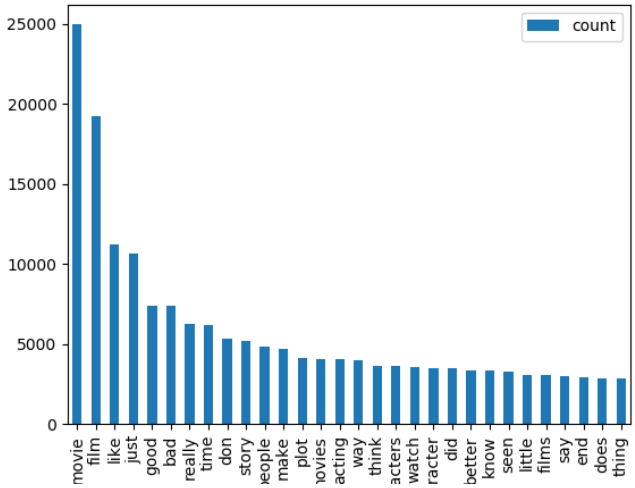
Prethodna dva grafikona(slike 2 i 3) predstavljaju najfrekventnije reči bez stop reči. Može se primetiti da su reči movie i film najfrekventnije u obe klase pa ih možemo smatrati kao stop reči.



Slika 1: Raspodela po klasama



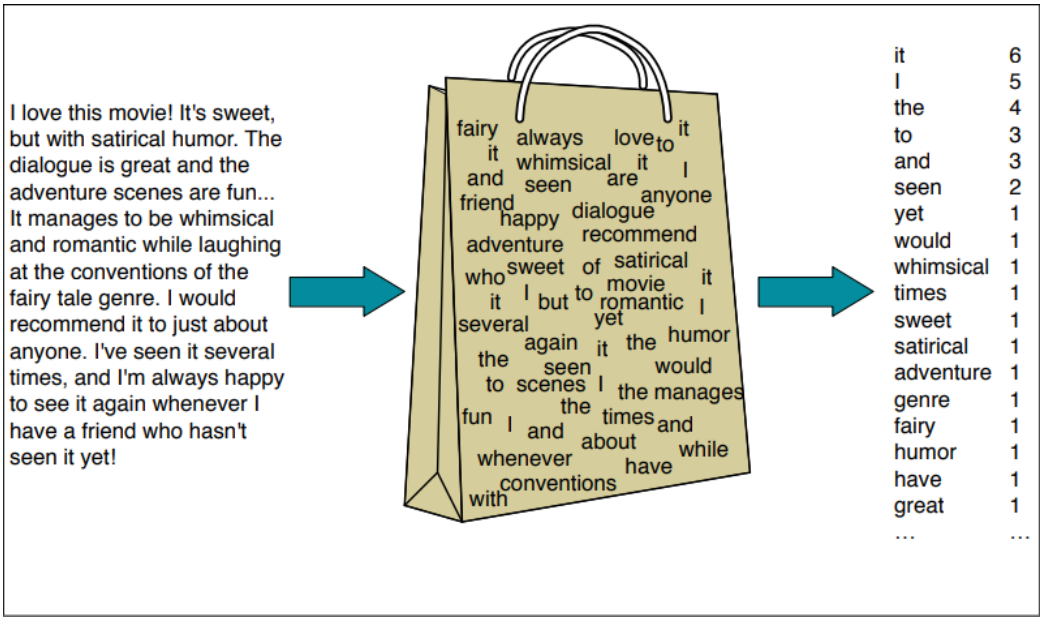
Slika 2: 30 najčešćih reči u pozitivnim kritikama



Slika 3: 30 najčešćih reči u negativnim kritikama

3 Preprocesiranje podataka

Pre same klasifikacije, potrebno je transformisati podatke u format koji odgovara ulazu za modele.



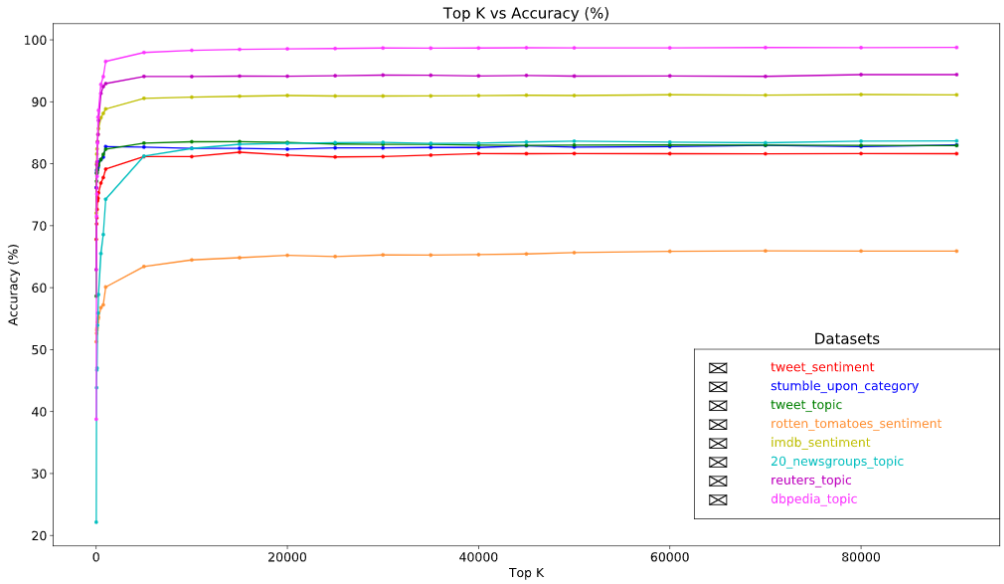
Slika 4: Bag-of-Words Izvor: web.stanford.edu/ jurafsky/slp3/

Dokument predstavljamo kao **vreća reči**(eng. Bag-of-Words, BoW) tj. kao skup reči, ignorišući njihov poredak i uzimajući u obzir samo njihov broj pojavljivanja u datom dokumentu. Skup podataka se predstavlja kao **term matrica**, gde svaku kolonu predstavlja jedna reč(term) iz rečnika, a vrste odgovaraju dokumentima (u ovom slučaju, filmskim kritikama). Rečnik čini uniju svih reči iz svih dokumenata. Na poziciji (i, j) u term matrici nalazi se broj pojavljivanja reči j u dokumentu i.

Jedan problem ove reprezenacije teksta su negacije reči. Razlika izmedju rečenica *I really like this movie* i *I don't like this movie* je očigledna, ali u navedenoj reprezentaciji dobijamo da se reč like javlja u oba teksta, bez informacije da je u jednom negirana. Ovo rešavamo dodavanjem prefiksa NOT_ rečima neposredno praćenim sa {n't, not, never, no} (Dan Jurafsky, 2019). Jos jedan pristup za čuvanje semantike je pomoću n-grama za $n > 1$. Uglavnom se koristi kombinacija unigrama($n = 1$) i bigrama($n = 2$), pa se

za rečenicu *I really like this movie* dobija rezultat *(I, really, like, this, movie, I really, really like, like this, this movie)*.

Mana term matrica je veliki broj atributa. Na IMDB-sentiment skupu, broj jednistvenih reči, bez dodatnog procesiranja je 74849, dok korišćenjem bigrama dobijamo 1520266 atributa.



Slika 5: Izvor: developers.google.com/machine-learning/guides/text-classification

Na osnovu slike 5 vidimo da se maksimalna preciznost dostiže sa približno 20000 atributa. Sve preko toga dovodi do preprilagođavanja i nepotrebnog izračunavanja. Izbor najboljih atributa je izvršen pomoću χ^2 testa. Sledi prikaz 50 najboljih atributa:

```
acting, amazing, annoying, avoid, awful,
bad, badly, beautiful, best, boring,
brilliant, crap, dull, excellent, fantastic,
favorite, great, highly, horrible, just,
lame, life, love, loved, mess,
minutes, money, n't, n't not_even, not_even,
oh, perfect, performance, plot, pointless,
poor, poorly, ridiculous, script, stupid,
superb, supposed, terrible, thing, today,
```

```
waste, waste time, wonderful, worse, worst
```

Reči bad, badly (poor, poorly) imaju istu osnovu, samim tim i srodno značenje. Izvedene reči bi bilo korisno ukloniti kako ne bismo imali više atributa sa sličnim značenjem. Steming[1] predstavlja postupak uklanjana odgovarajućih sufiksa rečima u pokušaju svođenja reči na osnovni oblik. 50 najboljih atributa, nakon steminga (korišćen je Porterov stemer):

```
act, amaz, annoy, aw, bad,
beauti, best, bore, brilliant, crap,
dull, enjoy, excel, fail, fantast,
favorit, great, high recommend, horribl, just,
lame, life, look like, love, mess,
minut, money, n't, n't not_even, not_even,
not_wast, oh, perfect, perform, plot,
pointless, poor, recommend, ridicul, script,
stupid, superb, suppos, terribl, today,
wast, wast time, wonder, wors, worst
```

Kao rezultat mogu nastati nepravilne reči, ali su zato uklonjene različite forme istih.

3.1 tf-idf

U sekciji 2 spomenute su stop reči i kako ih je potrebno ukloniti. Za njihovo uklanjanje, korišćen je predefinisan skup reči, pa reč smataramo da je stop reč ako se nalazi u tom skupu. Na osnovu slika 2 i 3, tom skupu reči smo dodali i *film, movie*. Naravno, postoji još mnogo drugih koje se često javljaju, podjednako u pozitivnim i negativnim kritikama ne noseći poseban značaj pa bi ručno uklanjanje bilo mukotrpno. Još jedan faktor koji je potrebno uzeti u obzir je dužina kritika. Zbog ovoga, umesto brojanja pojavljivanja reči u kritici, koristimo **frekvenciju terma** (eng. Term Frequency-tf), tj broj pojavljivanja reči podeljen sa dužinom dokumenta[2].

$$tf(t,d) = \frac{f_{t,d}}{N} \tag{1}$$

[1] Transkripcija engleske reči stemming

[2] Postoji više pristupa za računanje frekvencije terma

Inverzna frekvencija terma(eng. Inverse document frequency-idf) je mera koliko informacija nosi reč, tj da li se data reč često javlja u dokumentima ili je retka. idf definišemo kao:

$$idf(t, D) = log\frac{N}{1+M} \tag{2}$$

gde je N ukupan broj dokumenata, M je broj dokumenata u kojima se javlja reč t. Na osnovu ovoga, računamo tf-idf reči:

$$tfidf(t, d, D) = tf \cdot idf \tag{3}$$

4 Klasifikacija

U ovoj sekciji, testiraćemo više različitih algoritama za klasifikaciju nad podacima dobijenim preprocesiranjem opisanim u sekciji 3. Algoritmi koji će biti korišćeni su:

- Metod potornih vektora (eng. Support Vector Machine-SVM)
- Logistička regresija
- Stablo odlučivanja
- Naivni Bajes

Uzimajući u obzir da se radi o balansiranom i binarnom skupu podataka, za mere kvaliteta su izabrane preciznost, površina ispod ROC krive i f1 mera (Tan, 2019). Preciznost definišemo kao:

$$acc = \frac{\sum TP + \sum TN}{broj\_instanci} \tag{4}$$

gde je TP broj pozitivnih instanci klasifikovanih kao pozitivne. TN se definiše analogno.

F1 meru definišemo kao:

$$f1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{5}$$

gde Precision i Recall definišemo kao:

$$Precision = \frac{tp}{tp + fp} \tag{6}$$

$$Recall = \frac{tp}{tp + fn} \tag{7}$$

Skup podataka je podeljen na dva dela, na trening skup i validacioni skup[3]. Za algoritme SVM i logistička regresija korišćena je unakrsna validacija kako bi se odredili najbolji hiperparametri. Tabele 2 i 3 prikazuju postignute rezultate.

algoritam	acc	auc	f1
MNB	0.8576	0.8576	0.8606
LogReg	0.8694	0.8694	0.8666
SVM	**0.9068**	0.9068	0.9064
DT	0.6882	0.6882	0.6893

Tabela 2: rezultati nad tf-idf podacima

algoritam	acc	auc	f1
MNB	0.887	0.8867	0.8652
LogReg	0.8852	0.8852	0.8884
SVM	0.8862	0.8862	0.8855
DT	0.7096	0.7096	0.7125

Tabela 3: rezultati nad BoW podacima

Najbolje rezultate daje SVM algoritam u kombinaciji sa tf-idf zapisom, pa će u nastavku biti prikazane performanse SVM algoritma nad novim, nelabeliranim podacima iz test skupa. Radi preglednosti, izabrano je 6 najkraćih filmskih kritika:

1. I hope this group of film-makers never re-unites.

2. Brilliant and moving performances by Tom Courtenay and Peter Finch.

[3]Dostavljen test skup sadrži nelabelirane podatke

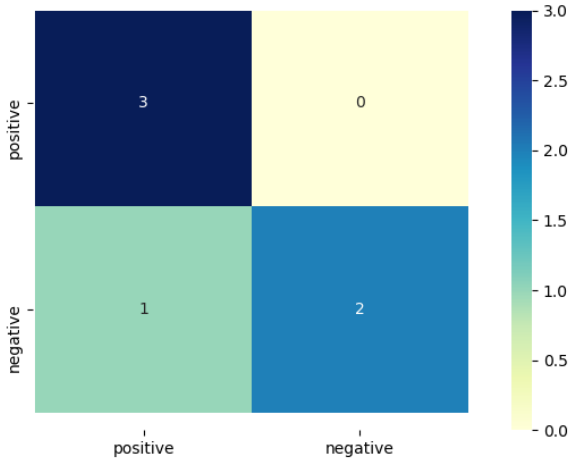
3. Add this little gem to your list of holiday regulars. It is sweet, funny, and endearing

4. A touching movie. It is full of emotions and wonderful acting. I could have sat through it a second time.

5. This is a terrible movie, don't waste your money on it. Don't even watch it for free. That's all I have to say.

6. Don't waste your time and money on it. It's not quite as bad as "Adrenalin", by the same director but that's not saying much.

Ručnim labeliranjem navedenih kritika, dobijamo niz labela $[1, 0, 0, 0, 1, 1]$. Dobijeni rezultat je $[0, 0, 0, 0, 1, 1]$, tj. SVM uspešno klasifikuje 5 od 6 kritika[4].



Slika 6: Matrica konfuzije za 6 test podataka koristeći SVM i tfidf

[4]na konkretnim primerima, SVM u kombinaciji sa brojanjem pojvaljivanja reči uspešno klasifikuje sve test instance

5 Zaključak

U ovom radu prikazan je klasičan pristup preprocesiranja tekstualnih podatak nad IMDB-sentiment skupu podataka. Korišćene su dve reprezentacije teksta pogodne za klasifikacione algoritme. Testirano je više različitih algoritama i najbolji rezultati(90% preciznost) su postignuti sa SVM algoritmom u kombinaciji sa tfidf reprezentacijom teksta. Za dalji rad potrebno je uporediti korišćene reprezentacije teksta sa novijim pristupima kao što je word2vec.

Literatura

Dan Jurafsky, J. H. M. (2019). *Speedh and language processing, 3rd. ed. draft*.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011, June). Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies* (pp. 142–150). Portland, Oregon, USA: Association for Computational Linguistics. Retrieved from `http://www.aclweb.org/anthology/P11-1015`

Tan, P.-N. (2019). *Introduction to data mining, 2nd edition*. Pearson.