Dražen Šokčević, Antonio Vidaković

Klasifikacija rečenica koristeći analizu sentimenta preko BERT modela

Projekt - Dokumentacija

Uvod

Radili smo analizu sentimenta i prikupljanje podataka u python-u. Ideja je bila naučiti klasifikaciju komentara korisnika preko modela kojeg smo trenirali na komentarima sa web-shopa. Analiza sentimenta komentara je korisna jer možemo automatski odrediti koliko je korisnik zadovoljan kupljenim proizvodom na temelju njegovog komentara koji ne mora biti ocijenjen.

Prikupljanje podataka

U početku smo tražili web-shopove koji imaju komentiranje s ocjenama kako bismo skinuli komentare i na njima trenirali i testirali model. Probali smo više stranica, ali ih većina ima zaštitu od automatiziranih zahtjeva (Amazon) ili nema dovoljan broj negativnih i neutralnih komentara (next.co.uk). Na kraju smo odabrali AliExpress koji ima različite kategorije proizvoda, kako ne bismo imali model naučen na jedinstvenoj kategoriji proizvoda i vokabularu specifičnog za tu kategoriju proizvoda, i mnoštvo komentara. Komentari na stranici su označeni sa 1 do 5 zvjezdica koje smo klasificirali u 3 kategorije:

- sa 4 – 5 zvjezdica su mapirani u pozitivne komentare (oznaka 0),
- sa 3 zvjezdice su mapirani u neutralne komentare (oznaka 1),
- sa 1 – 2 zvjezdice su mapirani u negativne komentare (oznaka 2).

Koristili smo dodatak za Google Chrome Web Scraper kojim smo prikupili linkove proizvoda koji su imali prosječnu ocjenu na kartici proizvoda (jer ti proizvodi imaju komentare) sa prvih pet stranica svake kategorije i spremili smo te linkove u AliExpressLinks.csv.

	A	B	C	D	E	F
1	web-scrap	web-scrap	link	link-href	stars	stars-href
2	162360044	https://do	https://wv	https://wv	5	https://ww
3	162359980	https://do	https://wv	https://wv	4.8	https://ww
4	162359892	https://do	https://wv	https://wv	4.9	https://ww
5	162359888	https://do	https://wv	https://wv	4.8	https://ww
6	162359835	https://do	https://wv	https://wv	4.9	https://ww
7	162360060	https://do	https://wv	https://wv	4.6	https://ww
8	162359984	https://do	https://wv	https://wv	4.9	https://ww
9	162360123	https://do	https://wv	https://wv	4.9	https://ww
10	162359864	https://do	https://wv	https://wv	4.9	https://ww
11	162360091	https://do	https://wv	https://wv	4.3	https://ww
12	162360103	https://do	https://wv	https://wv	4.7	https://ww
13	162359786	https://do	https://wv	https://wv	4.2	https://ww
14	162360132	https://do	https://wv	https://wv	5	https://ww
15	162359823	https://do	https://wv	https://wv	4.8	https://ww

Prvih 15 linija AliExpressLinks.csv-a.

Nakon toga iz tih linkova smo vadili identifikacijske oznake prodavača (ownerMemberId) i proizvoda (productId) koji su potrebni za dohvaćanje komentara te smo ih spremili u productSellerIds.csv. Ovo smo implementirali u getIDS.py prilikom čega smo uzimali samo one likove koji su imali prosječnu ocjenu manju od 5.

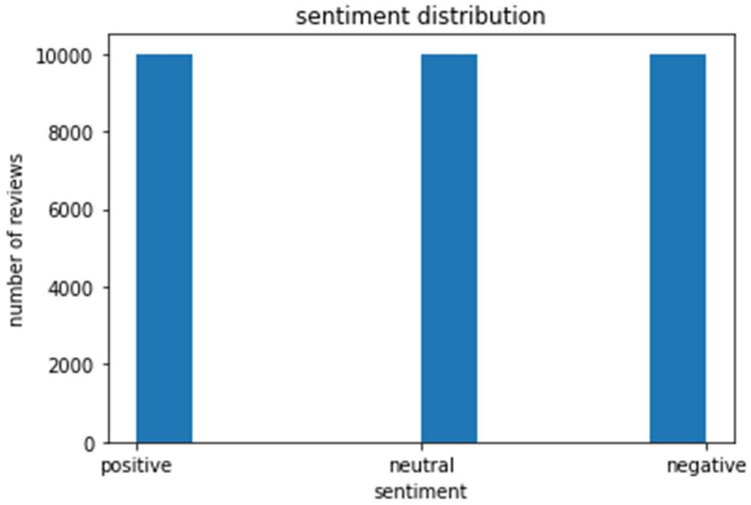
	A	B	C
1		productId	ownerMemberId
2	0	1005002566765520	247392345
3	1	1005001775499530	232198970
4	2	4000734290088	240313962
5	3	1005001656768420	247056311
6	4	32965169159	228391344
7	5	4000176297447	222250037
8	6	33044850111	201453932
9	7	1005001320065980	221041720
10	8	4001066058466	235246733
11	9	33057481981	227765123
12	10	1005002103723300	230385907
13	11	1005001315502400	221381463
14	12	4000829878997	240319983
15	13	1005001994749210	243707164

Prvih 15 linija productSellerIds.csv

Komentare smo dohvaćali preko zahtjeva koje smo filtrirali preko broja zvjezdica do prvih 8 stranica za svaki broj zvjezdica sve dok nismo sakupili 10000 komentara iz svake kategorije i spremili u commentsData.csv te smo ih preveli na engleski koristeći Google Translate i spremili u translatedData.csv. Prikupljanje podataka smo implementirali u getReviews.ipynb

	A	B	C	D
1		positive	neutral	negative
2	0	All the bes	They're un	It's just pla
3	1	Thanks.	Ordinary c	Rosemir
4	2	Come quic	To see in t	Sent a diff
5	3	Good cute	Could be b	Parcel not
6	4	Excellent s	,	Some of th
7	5	EU europe	.	The sticker
8	6	Good qual	I don't arri	cheap qual
9	7	Sneakers a	Not neces:	Terrible qu
10	8	Shipping ve	<2en> thar	Because th
11	9	Quality is e	Its too lon	Poor qualit
12	10	As describe	Good	I don't kno
13	11	comfort ar	People	Shipment c
14	12	Malbe on t	The produ	Shipment c
15	13	Lost with t	I regretted	I never get

Prvih 15 linija translatedData.csv-a



Prikaz distribucije podataka

Treniranje i testiranje modela

Priprema podataka

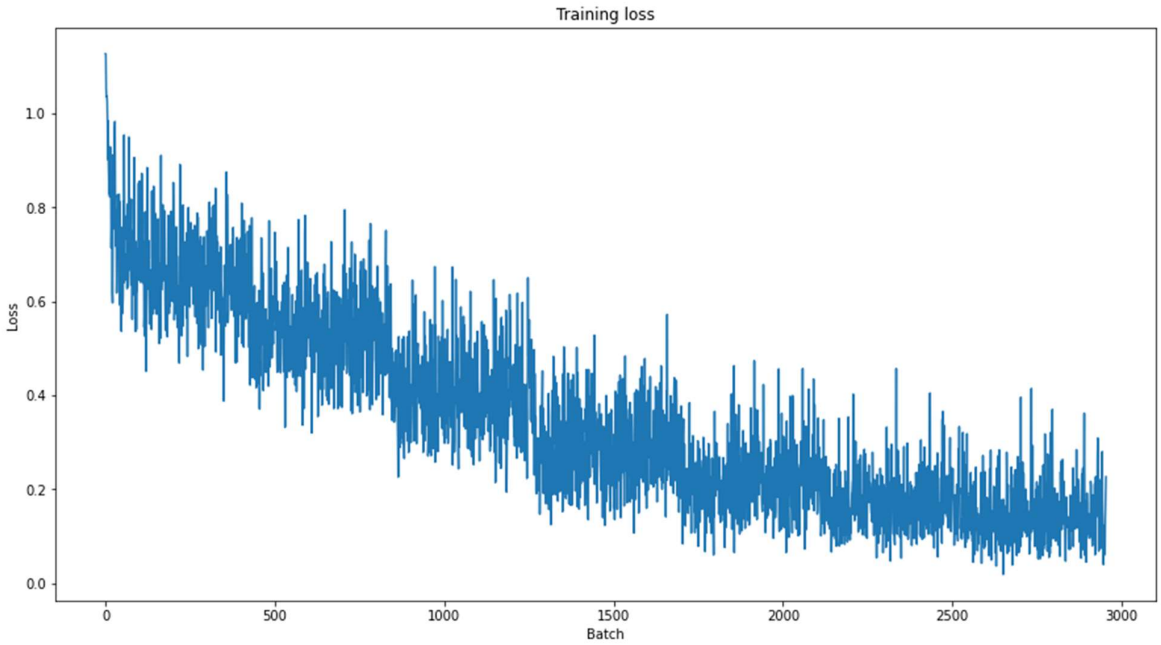
Koristili smo pred trenirani BertForSequenceClassification iz pytorch_pretrained_bert-a. Taj model je modificirani BERT model koji u zadnjem sloju ima linearan sloj za klasifikaciju.
Iz translatedData.csv smo učitali komentare i njihove oznake. Komentare smo tokenizirali automatskom tokenizacijom na komentare na koje smo dodali „[CLS] “ token na početak i „ [SEP]“ token na kraj. Izbacili smo predugačke komentare koje model ne može dobro klasificirati zbog ograničenja modela. Na kraju smo na

tokenizirane komentare dodali padding.
Nakon toga smo za paddirane komentare napravili maske pozornosti.

Dobivene liste smo podijelili na skup za treniranje (90%) i skup za testiranje(10%) te od njih napravili tensore.
Iz dobivenih tensora smo napravili Dataset, RandomSampler i DataLoader koji se koristi za treniranje i testiranje modela.

Treniranje i testiranje

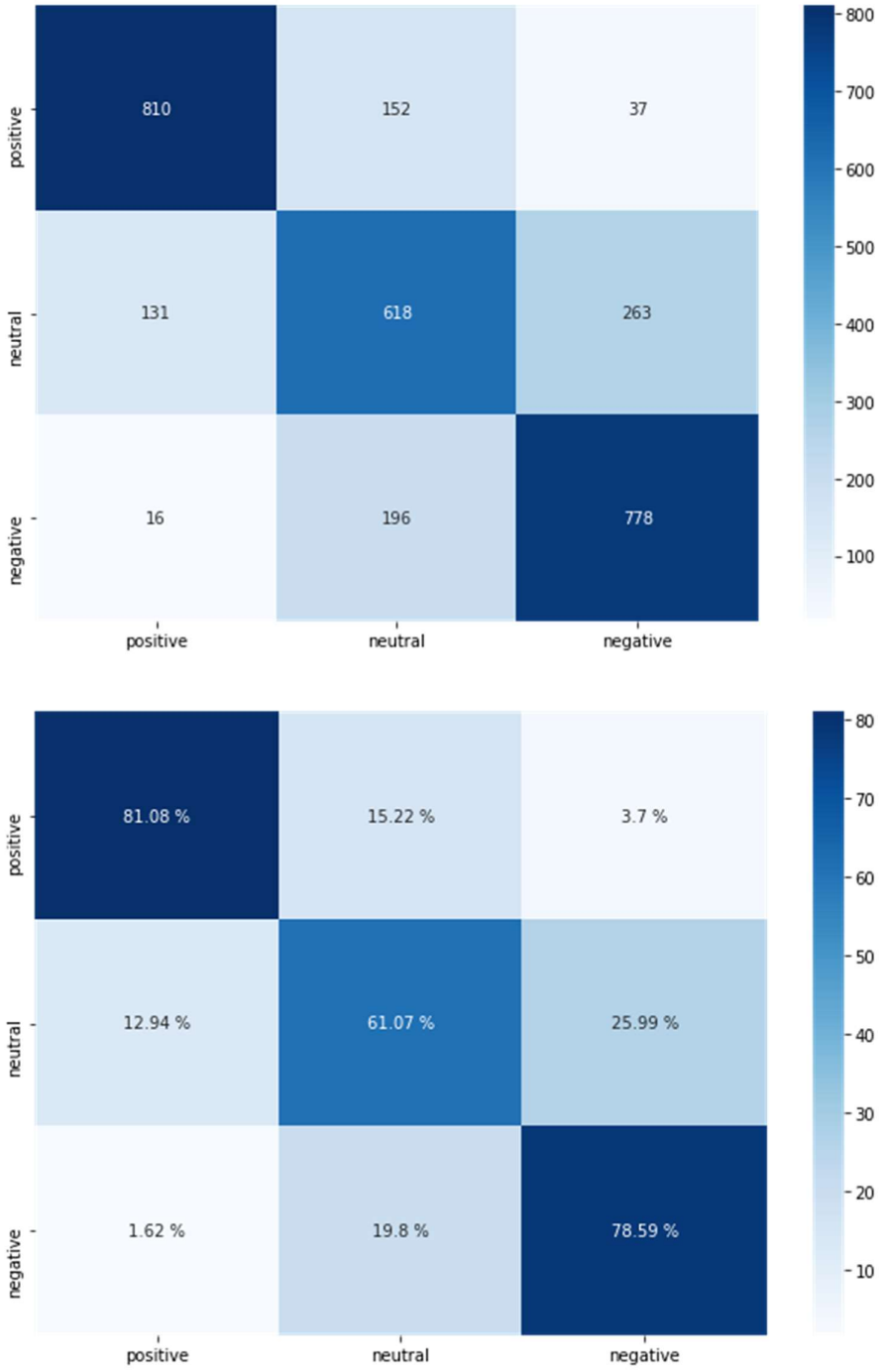
Treniranje modela smo proveli kroz 7 epoha. Koristili smo BertAdam optimizator za treniranje modela koristeći stopu učenja od $2 * 10^{-5}$. Prilikom treniranja smo pratili funkciju gubitka i točnost modela na skupu za testiranje. Postigli smo točnost modela na skupu za testiranje od 73.5% i vrijednost funkcije gubitka od oko 0.149.



Prikaz funkcije gubitka kroz vrijeme

Kroz epohe smo primijetili da vrijednost funkcije gubitka opada, a da se točnost klasifikacije podataka na testnom skupu stabilizira kroz par epoha.

Na rezultatima klasifikacije testnog skupa primijetili smo da model bolje klasificira pozitivne i negativne komentare nego neutralne.



Matrice konfuzije klasificiranosti podataka u testnom skupu

Primijetili smo da među podatcima postoje krivo klasificirani komentari koje su korisnici stranice krivo označili, npr. komentar „Super“ označen sa negativnom ocjenom.

Korištene stranice

https://www.aliexpress.com/, lipanj, 2021.

https://huggingface.co/transformers/model_doc/bert.html#bertforsequenceclassification, lipanj, 2021.

https://ipywidgets.readthedocs.io/en/latest/examples/Widget%20List.html, lipanj, 2021.