Sentiment Analysis with comparison of BERT and Naive Bayes

Anja Colic
Igor Zolotarev

Sadrzaj

1. Naive Bayes
2. Bert
3. Results

Sentiment analiza

My experience so far has been fantastic!

POSITIVE

The product is okay I guess.

NEUTRAL

Your support team is useless.

NEGATIVE

Sentiment analiza

- Prikupljanje podataka
- Predobrada podataka
- Klasifikacija
- Evaluacija

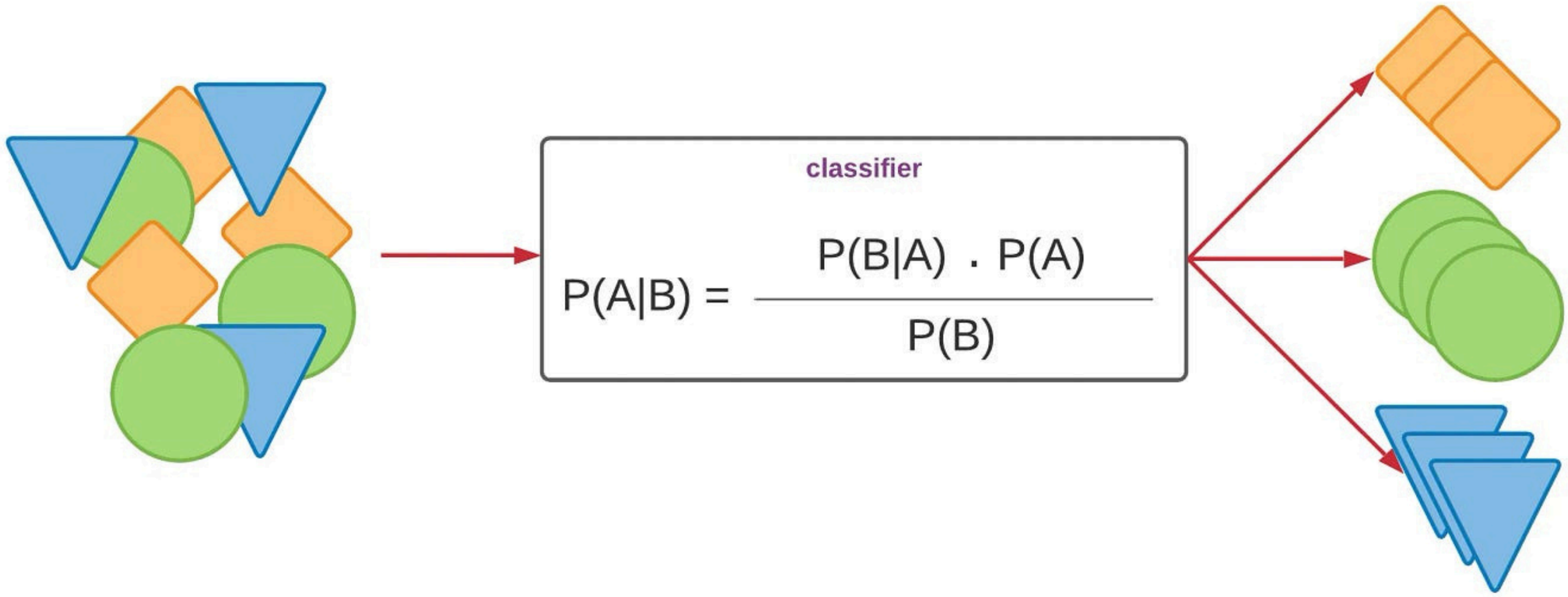
Primena

- Analiza drustevnih mreza
- Finansijsko trziste
- Politicka analiza

Izazovi

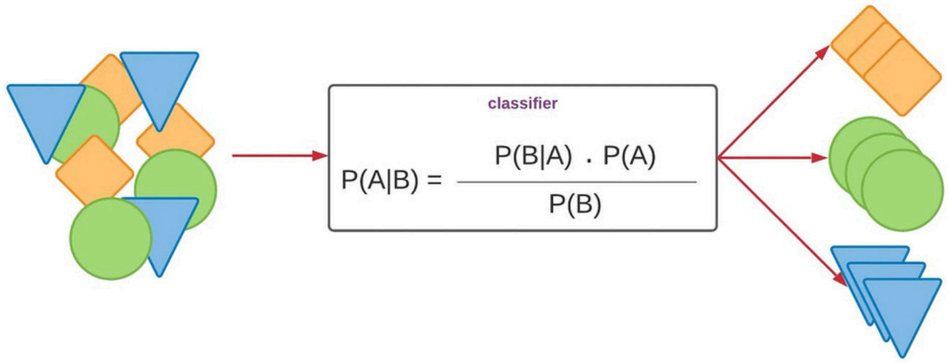
- Sarkazam i ironija
- Kontekstualno razumevanje
- Visejezicnost

Naive Bayes



Naivni Bajes (Naive Bayes) je jednostavan, ali vrlo efikasan algoritam za klasifikaciju zasnovan na Bajesovoj teoremi, koja se koristi za predvianje verovatno´ce klase na osnovu skupa podataka.

Naive Bayes



Skupljanje podataka

Obučavanje na skupu podataka gde je svaka instanca obeležena klasom. Na primer, za zadatak klasifikacije e-mail poruka kao „spam“ ili „nije spam“

Izracunavanje verovatnoca

Nakon što je model obučen, za novu instancu podataka X (na primer, novi e-mail), model koristi Bayesovu teoremu kako bi izracunao verovatnocu da X pripada svakoj mogucoj klasi (npr. spam ili nije spam).

Predikcija

Ako verovatnoca da je e-mail spam P(spam|X) veca od verovatnoce da nije spam P(nije spam|X), model ce ga označiti kao spam.

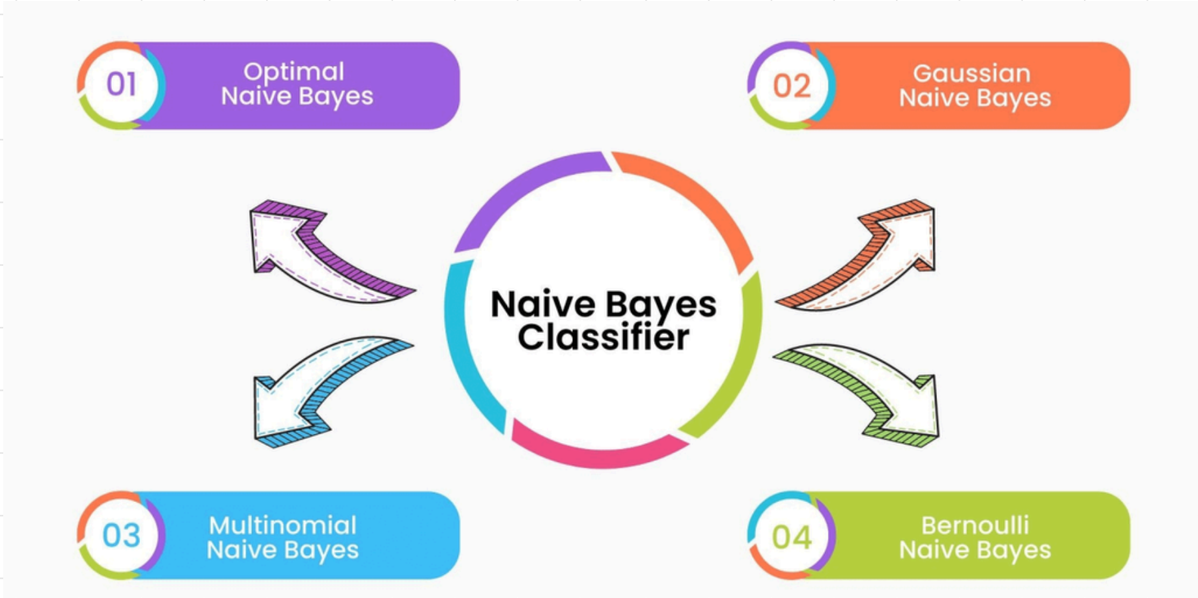
Naive Bayes

Prednosti

- Brzina
- Jednostavnost
- Efikasan na malim skupovima podataka

Mane

- Pretpostavka nezavisnosti (cesto nije tako na stvarnim podcima)
- Osetljivost na podatke sa nultim vrednostima



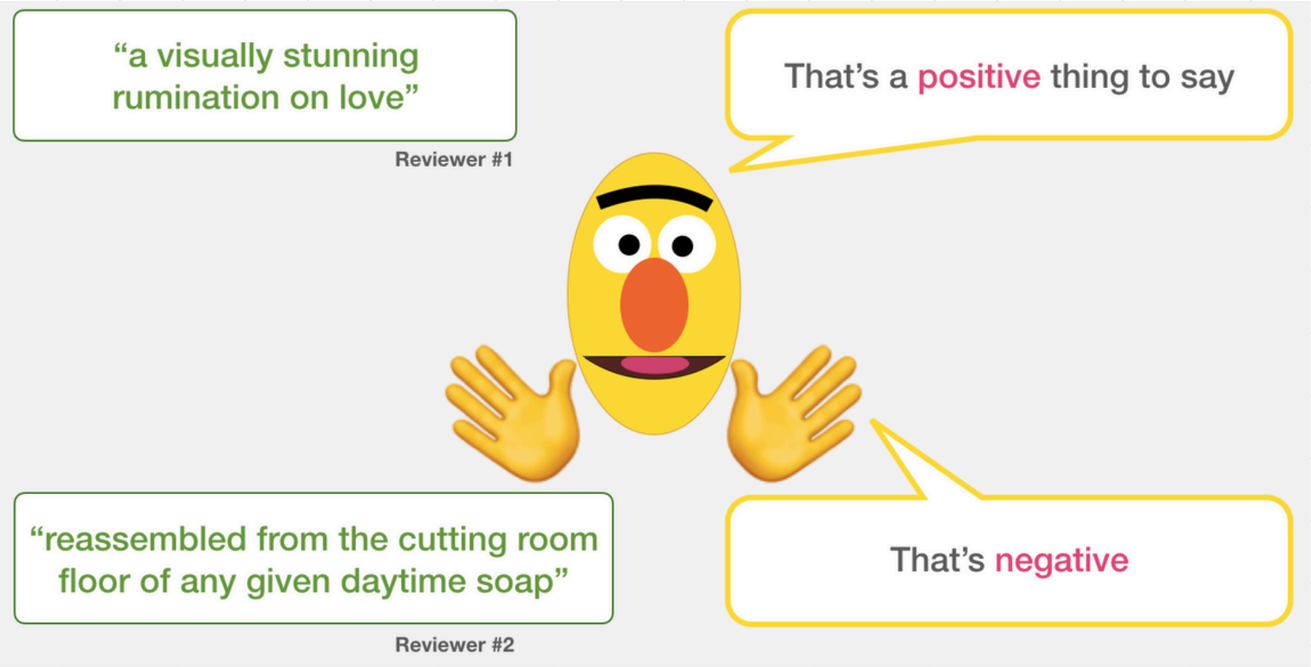
BERT

Google

BERT

Napredni model za obradu prirodnog jezika (NLP) koji je predstavljen od strane Google-a u oktobru 2018. godine

BERT



Self-attention

Transformer arhitektura koristi mehanizam self-attention koji omogućava modelu da usmeri pažnju na relevantne reči u rečenici u odnosu na ostale.

Obrada i tokenizacija

Tekst koji se unosi u BERT se najpre razdvaja na tokene. BERT koristi WordPiece tokenizaciju koja deli reči na osnovne delove. Na primer, reč „playing” može biti podeljena na „play” i „ing”.

Maskirani jezički model (MLM)

Tokom pre-treninga, BERT nasumično maskira određeni procenat tokena u rečenici (oko 15\%) i zatim pokušava da predvidi te maskirane tokene koristeći njihov kontekst.

Predikcija sledeće rečenice (NSP)

Pored MLM zadatka, BERT se trenira da predvidi da li je druga rečenica logički sledeća nakon prve. Ovaj zadatak omogućava BERT-u da razume odnose između rečenica.

BERT

Prednosti

- Razumevanje celokupnog konteksta
- Prilagodljivost
- Tacnost

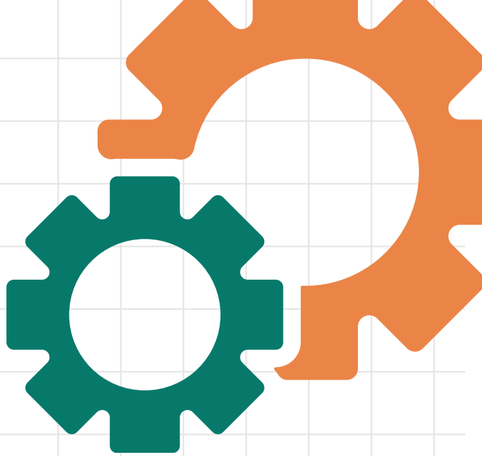
Osobine

- Biderekcionalnost
- Transformer arhitektura
- Pre-trening

Primena

- Sentiment analiza
- Question Answering
- Prevodjenje..

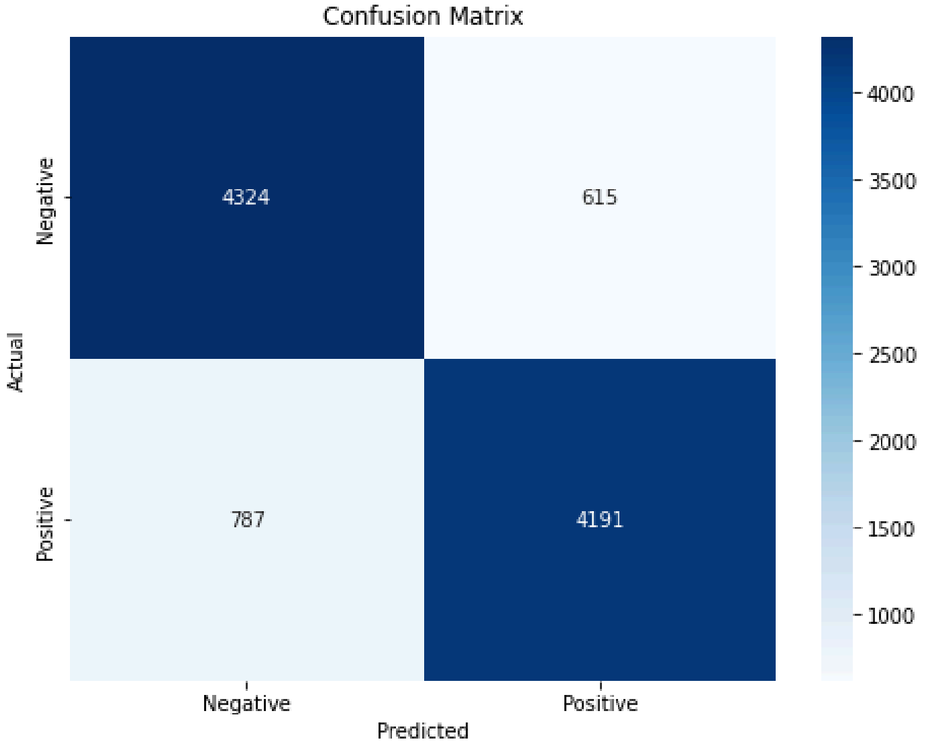
Dataset

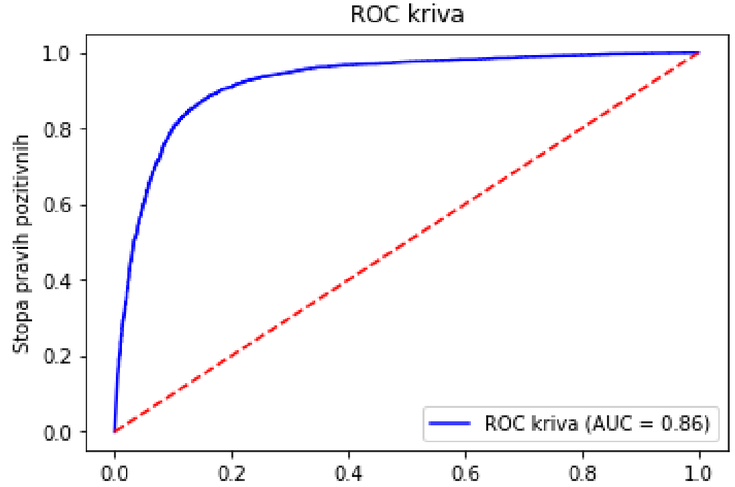


	review	sentiment
0	One of the other reviewers has mentioned that ...	1
1	A wonderful little production. The filming tec...	1
2	I thought this was a wonderful way to spend ti...	1
3	Basically there's a family where a little boy ...	0
4	Petter Mattei's "Love in the Time of Money" is...	1
...	...	...
49995	I thought this movie did a down right good job...	1
49996	Bad plot, bad dialogue, bad acting, idiotic di...	0
49997	I am a Catholic taught in parochial elementary...	0
49998	I'm going to have to disagree with the previou...	0
49999	No one expects the Star Trek movies to be high...	0

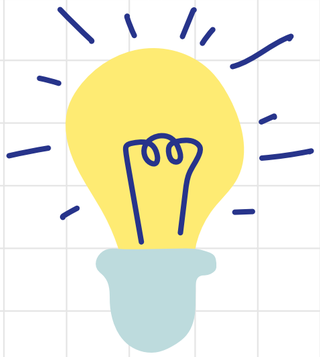
Naive Bayes

ComplementNB model accuracy is 85.86%





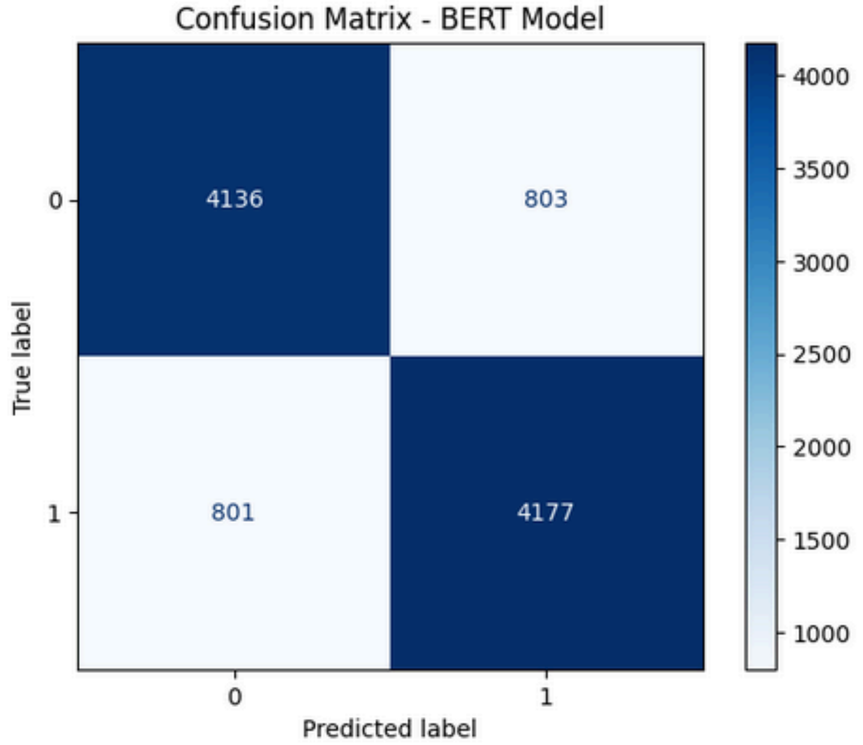
vs

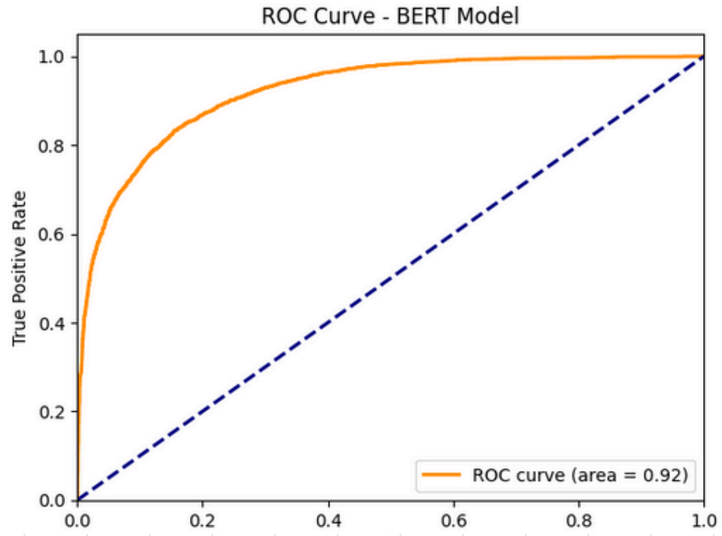


Results

BERT

Test Loss: 0.4006, Test Accuracy: 0.8383





Thank you!

We hope you learn something new today.