SENTIMENTALNA ANALIZA

ANA ARSIĆ, TAMARA ŠAPONJIĆ

SENTIMENTALNA ANALIZA

- Prikupljanje podataka
- Pretprocesiranje podataka
- Klasifikacija i evaluacija

DATASET

IMDb Dataset of 50K Movie Reviews

50 000 recenzija (25k pozitivnih / 25k negativnih)

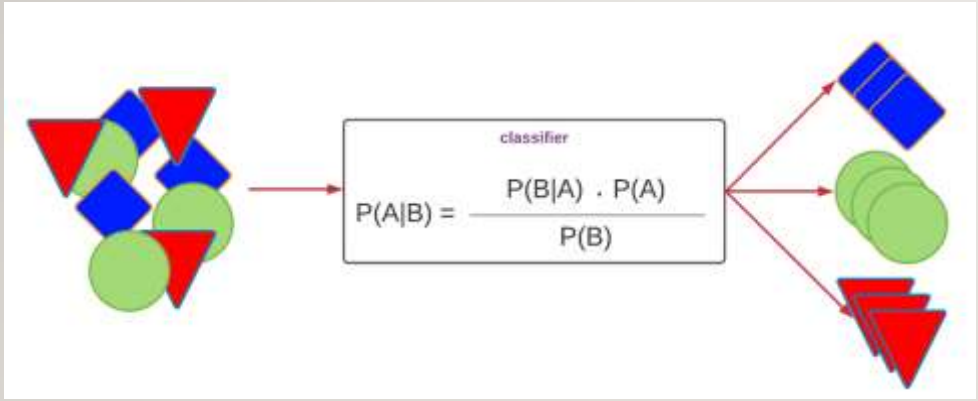
Balansiran skup podataka za binarnu klasifikaciju

	review	sentiment	label
0	One of the other reviewers has mentioned that ...	positive	1
1	A wonderful little production. The...	positive	1
2	I thought this was a wonderful way to spend ti...	positive	1
3	Basically there's a family where a little boy ...	negative	0
4	Petter Mattei's "Love in the Time of Money" is...	positive	1

PRETPROCESIRANJE PODATAKA

- Uklanjanje HTML tagova i šuma
- Tokenizacija i normalizacija teksta
- Uklanjanje stop-reči

NAIVE BAYES



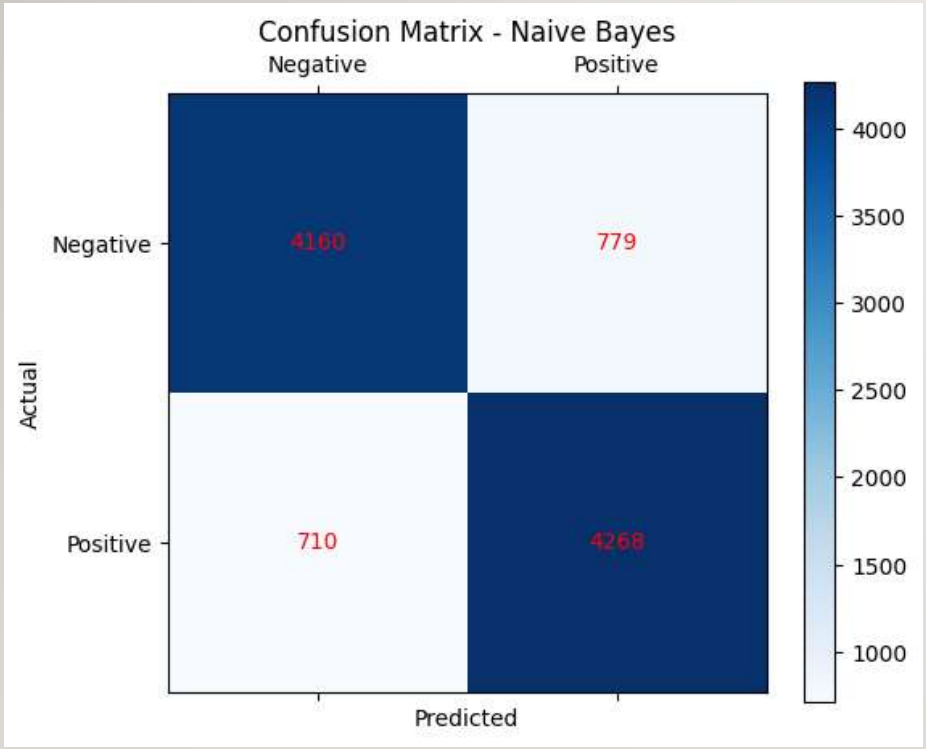
- Probabilistički klasifikator zasnovan na Bajesovoj teoremi
- Koristi TF-IDF reprezentaciju teksta
- Prednosti NB modela su brzina treniranja i mala memorijska zahtevnost, dok je mana ograničena sposobnost hvatanja složenih jezičkih obrazaca

SUPPORT VECTOR MACHINE (SVM)

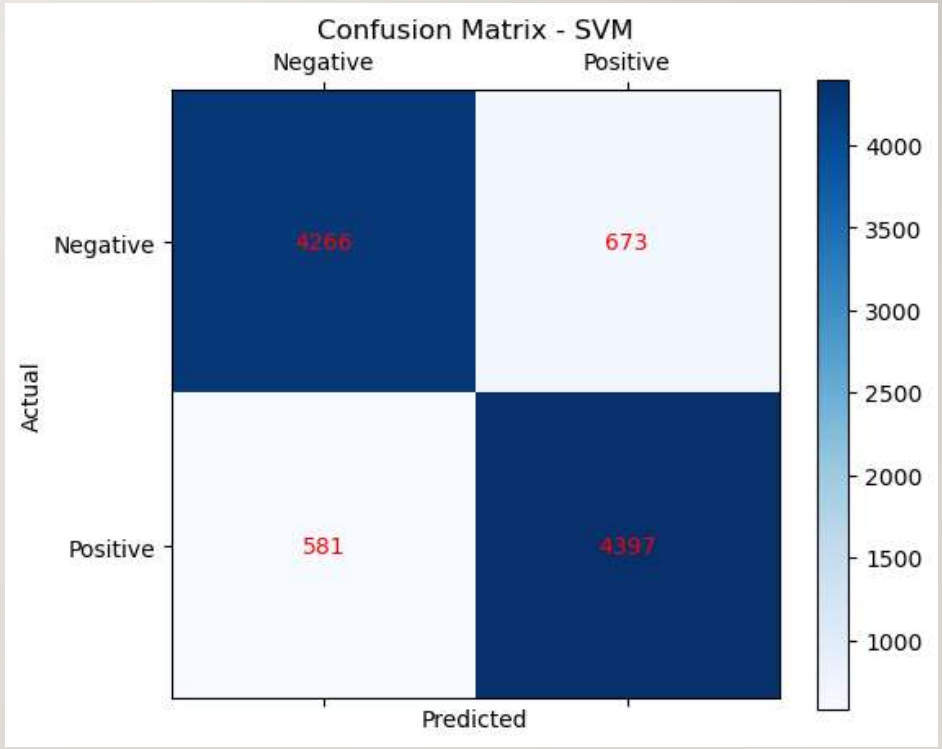
- Geometrijsko razdvajanje: SVM traži optimalnu granicu (hiperravan) koja fizički razdvaja podatke različitih klasa u prostoru
- Izuzetno je efikasan kod podataka sa mnogo karakteristika, poput onih dobijenih putem TF-IDF metode
- Prednost je visok kvalitet klasifikacije i dobra generalizacija, dok je mana to što treniranje može biti sporije na velikim skupovima podataka

NB vs SVM

NB



SVM



LSTM (LONG SHORT-TERM MEMORY)

Rekurentna neuronska mreža sa mehanizmom memorije

LSTM uvodi ćelijsko stanje ct koje omogućava dugoročno pamćenje informacija, dok se protok informacija kontroliše pomoću gate mehanizama. LSTM koristi sledeće gate-ove:

• Forget gate	• Input gate	• Candidate state	• Output gate

GRU (GATED RECURRENT UNIT)

Pojednostavljena varijanta LSTM-a

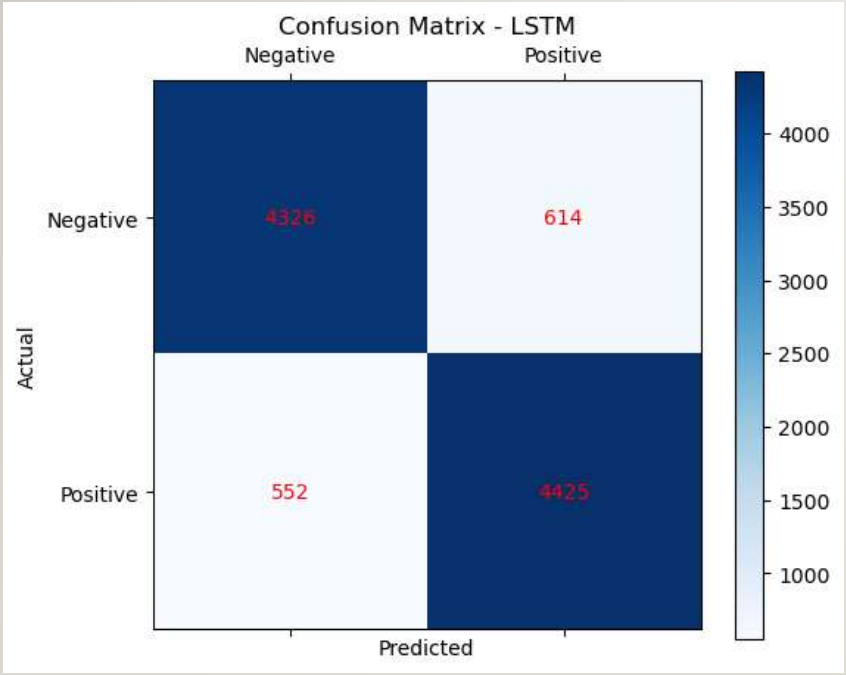
Zbog manje složenosti, GRU se često trenira brže od LSTM-a, dok performanse mogu biti slične

GRU nema posebno ćelijsko stanje ct , već koristi skriveno stanje ht i dva gate-a:

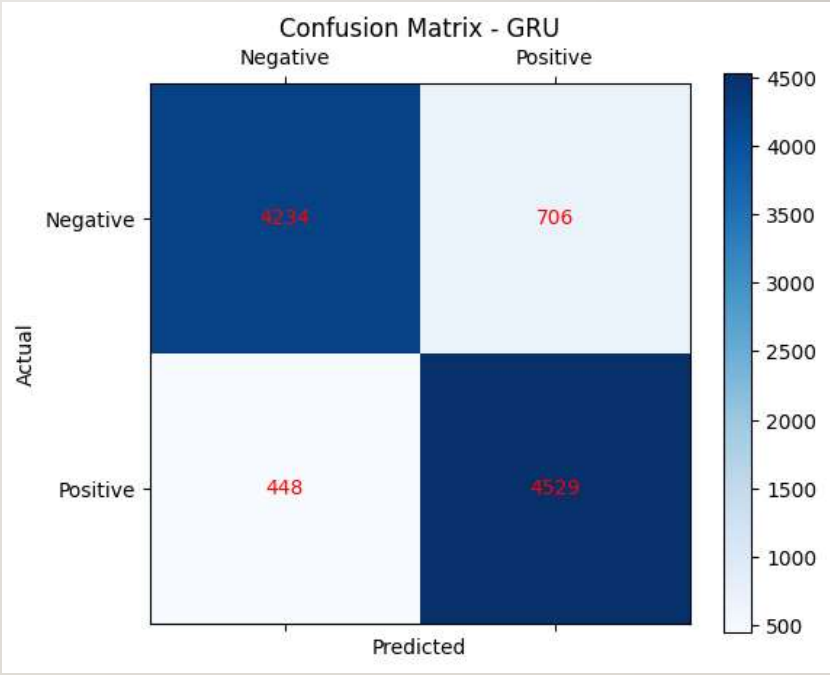
- Update gate
- Reset gate

LSTM vs GRU (128, 0.4)

LSTM



GRU



REZULTATI

- NB i SVM daju solidne bazne rezultate
- LSTM i GRU ostvaruju bolje performanse
- Najbolji model: GRU (64 jedinice, dropout = 0.2)

Poređenje performansi modela za analizu sentimenta

Model	Accuracy	F1-score
Naive Bayes	0.8499	0.85
SVM	0.8736	0.87
LSTM (64, d=0.2)	0.8664	0.87
LSTM (128, d=0.4)	0.8824	0.88
GRU (64, d=0.2)	0.8851	0.89
GRU (128, d=0.4)	0.8836	0.88
MyLSTM	0.8803	0.88
MyGRU	0.8798	0.88

ZAKLJUČAK

- Neuronski modeli nadmašuju klasične pristupe
- GRU se pokazao kao najstabilniji model
- Na osnovu dobijenih rezultata može se zaključiti da izbor arhitekture i hiperparametara ima ključnu ulogu u kvalitetu klasifikacije, kao i da jednostavniji modeli mogu nadmašiti složenije u zavisnosti od prirode podataka.

HVALA NA PAŽNJI!