Univerzitet u Beogradu

Matematički fakultet

Sentiment analysis

Predmet: Računarska inteligencija

Autori: Arsić Ana, Šaponjić Tamara

Profesor: dr Vladimir Filipović

Asistent: Stefan Kapunac

Datum: Februar 2026

Sadržaj

1 Uvod

Sentiment analiza (eng. *sentiment analysis*) predstavlja zadatak obrade prirodnog jezika u kome se tekst klasifikuje prema izraženom stavu autora, najčešće kao pozitivan ili negativan. U ovom radu razmatramo binarnu sentiment klasifikaciju filmskih recenzija sa IMDb-a.

Cilj rada je poređenje klasičnih metoda mašinskog učenja i rekurentnih neuronskih mreža na istom skupu podataka. Korišćeni modeli su:

- **Naive Bayes (NB)** – brza i jednostavna probabilistička metoda koja se često koristi za klasifikaciju teksta kada su ulazne karakteristike TF-IDF ili Bag-of-Words.
- **Support Vector Machine (SVM)** – robustan klasifikator koji pronalazi hiperravan koja najbolje razdvaja klase; često postiže vrlo dobre rezultate u problemskim domenima klasifikacije teksta sa TF-IDF karakteristikama.
- **LSTM** – rekurentna neuronska mreža sa mehanizmom memorije koja bolje uči dugoročne zavisnosti u sekvencama teksta.
- **GRU** – jednostavnija varijanta LSTM-a, sa manje parametara, često brža za treniranje i ponekad sa sličnim performansama.

Pored korišćenja ugrađenih slojeva, implementirali smo i **ručne (manual) verzije** LSTM i GRU slojeva (MyLSTM i MyGRU) koristeći tenzorske operacije i eksplicitno računanje gate-ova, čime se demonstrira razumevanje unutrašnjeg rada ovih modela.

2 Opis rešenja

2.1 Skup podataka

Korišćen je javno dostupan skup podataka **IMDb Dataset of 50K Movie Reviews** (Kaggle)[1]. Skup sadrži 50 000 recenzija ravnomerno podeljenih na pozitivne i negativne (25 000/25 000), što ga čini balansiranim za binarnu klasifikaciju.

2.2 Pretprocesiranje teksta

Pretprocesiranje obuhvata standardne korake čišćenja i pripreme teksta - uklanjanje HTML tagova, tokenizacija, uklanjanje stop-reči, lematizacija, kako bi se smanjio šum i poboljšala generalizacija modela.

[1] `https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews`

2.3 Reprezentacija teksta

Za NB i SVM modeli koriste **TF-IDF** reprezentaciju teksta. Za LSTM/GRU modeli koriste **sekvencijalnu reprezentaciju** (tokene mapirane na indekse), embedding sloj i zatim rekurentni sloj (LSTM/GRU), nakon čega sledi izlazni sloj.

2.4 Modeli

U okviru rada implementirana su četiri pristupa za binarnu klasifikaciju sentimenta: Naive Bayes, SVM, LSTM i GRU. Prva dva pristupa spadaju u klasične metode mašinskog učenja, dok druga dva predstavljaju neuronske modele za rad sa sekvencama. Za fer poređenje, svi modeli su trenirani i evaluirani na istom skupu podataka.

2.4.1 Naive Bayes (NB)

Naive Bayes je probabilistički klasifikator zasnovan na Bajesovoj teoremi. Model procenjuje verovatnoću pripadnosti klase C na osnovu ulaznih karakteristika X:

$$P(C|X) = \frac{P(X|C) \cdot P(C)}{P(X)}.$$

U zadacima klasifikacije teksta najčešće se koristi Multinomial Naive Bayes, gde se tekst posmatra kao skup reči, a karakteristike predstavljaju učestalosti ili TF-IDF vrednosti. Glavna pretpostavka modela je da su karakteristike međusobno nezavisne uslovljeno klasom, što u praksi nije strogo tačno, ali model često daje vrlo dobre rezultate zbog jednostavnosti i robusnosti.

Prednosti NB modela su brzina treniranja i mala memorijska zahtevnost, dok je mana ograničena sposobnost hvatanja složenih jezičkih obrazaca (npr. negacije i kontekst).

2.4.2 Support Vector Machine (SVM)

Support Vector Machine je klasifikator koji traži hiper-ravan koja maksimalno razdvaja klase u prostoru karakteristika. U slučaju linearne varijante, cilj je pronaći granicu odluke:

$$w^T x + b = 0$$

tako da je margina između klasa maksimalna. SVM je poznat kao veoma efikasan model za klasifikaciju teksta kada se koristi TF-IDF reprezentacija, jer dobro radi u visokodimenzionalnim prostorima.

Prednost SVM-a je visok kvalitet klasifikacije i dobra generalizacija, dok je mana to što treniranje može biti sporije na velikim skupovima podataka i što model ne daje direktno verovatnoće (u slučaju LinearSVC).

2.4.3 LSTM

Long Short-Term Memory (LSTM) predstavlja rekurentnu neuronsku mrežu dizajniranu za rešavanje problema nestajanja gradijenta (eng. *vanishing gradient*) kod standardnih RNN mreža. LSTM uvodi ćelijsko stanje c_t koje omogućava dugoročno pamćenje informacija, dok se protok informacija kontroliše pomoću gate mehanizama.

LSTM koristi sledeće gate-ove:

- **Forget gate** f_t: određuje koje informacije iz prethodnog stanja treba zaboraviti,
- **Input gate** i_t: određuje koje nove informacije treba upisati,
- **Candidate state** $\tilde{c}_t$: kandidovana nova informacija,
- **Output gate** o_t: određuje koje informacije se iznose kao izlaz.

U ovom radu implementirana je i ručna verzija LSTM sloja (MyLSTM) korišćenjem TensorFlow operacija, pri čemu su svi gate-ovi eksplicitno izračunavani u svakom vremenskom koraku.

2.4.4 GRU

Gated Recurrent Unit (GRU) je pojednostavljena varijanta LSTM-a, sa manjim brojem parametara. GRU nema posebno ćelijsko stanje c_t, već koristi skriveno stanje h_t i dva gate-a:

- **Update gate** z_t: određuje koliko prethodnog stanja ostaje,
- **Reset gate** r_t: određuje koliko prethodnog stanja se koristi pri formiranju kandidovanog stanja.

Zbog manje složenosti, GRU se često trenira brže od LSTM-a, dok performanse mogu biti slične. Kao i za LSTM, implementirana je i ručna verzija GRU sloja (MyGRU), gde su update i reset gate eksplicitno računati.

2.4.5 Arhitektura neuronskih modela

Za LSTM i GRU korišćena je standardna arhitektura:

- Embedding sloj koji mapira tokene u vektore fiksne dimenzije,
- LSTM ili GRU sloj (ugrađeni ili ručno implementirani),
- Dropout sloj radi regularizacije,
- Izlazni Dense sloj sa jednom jedinicom i sigmoid aktivacijom.

Pošto je zadatak binarna klasifikacija, izlazni sloj ima jedan neuron, a funkcija gubitka je binary cross-entropy.

3 Eksperimentalni rezultati

3.1 Eksperimentalno okruženje

- Operativni sistem: Linux
- Python verzija: 3.10
- Biblioteke: TensorFlow/Keras, scikit-learn, NLTK, NumPy, Matplotlib

3.2 Rezultati

Tabela 1: Performanse klasičnih modela za analizu sentimenta

Model	**Accuracy**	**F1-score**
Naive Bayes	0.8499	0.85
SVM	0.8736	0.87

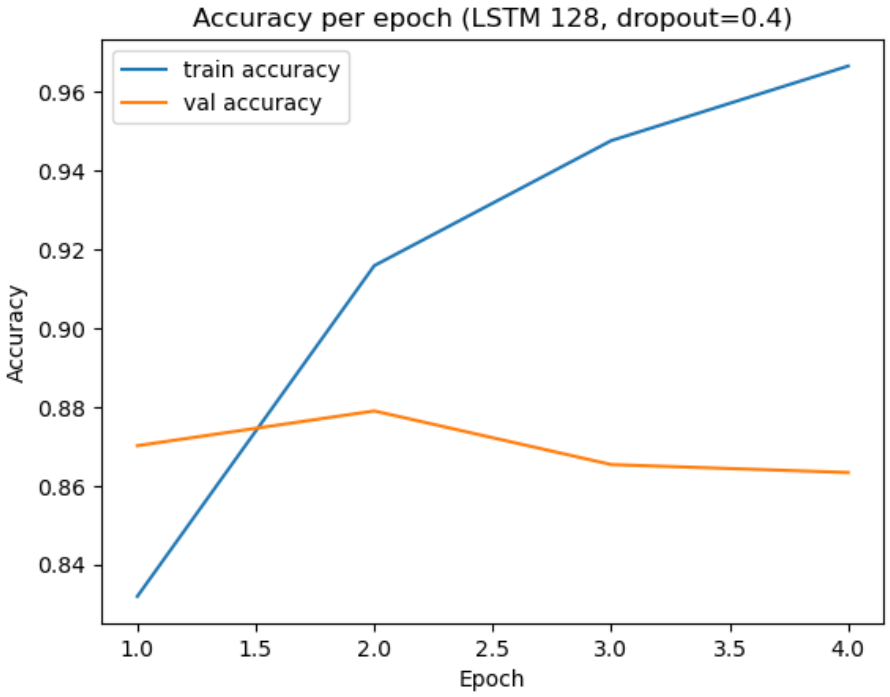
Tabela 2: Uticaj hiperparametara na performanse LSTM i GRU modela

Model	**Accuracy**	**F1-score**
LSTM (64, dropout=0.2)	0.8664	0.87
LSTM (128, dropout=0.4)	0.8824	0.88
GRU (64, dropout=0.2)	0.8851	0.89
GRU (128, dropout=0.4)	0.8836	0.88

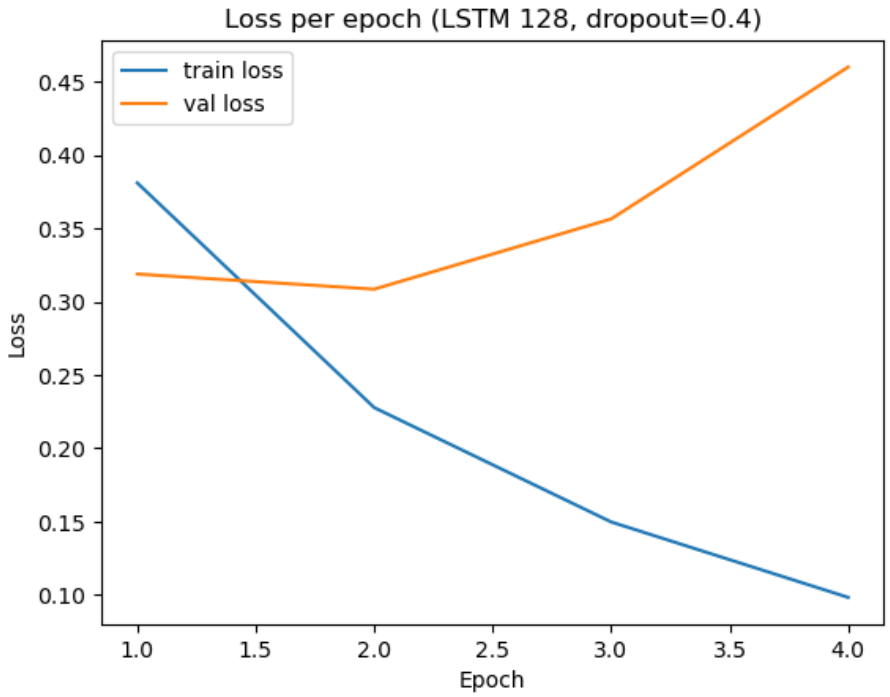
Tabela 3: Performanse ručno implementiranih RNN modela

Model	**Accuracy**	**F1-score**
MyLSTM	0.8803	0.88
MyGRU	0.8798	0.88

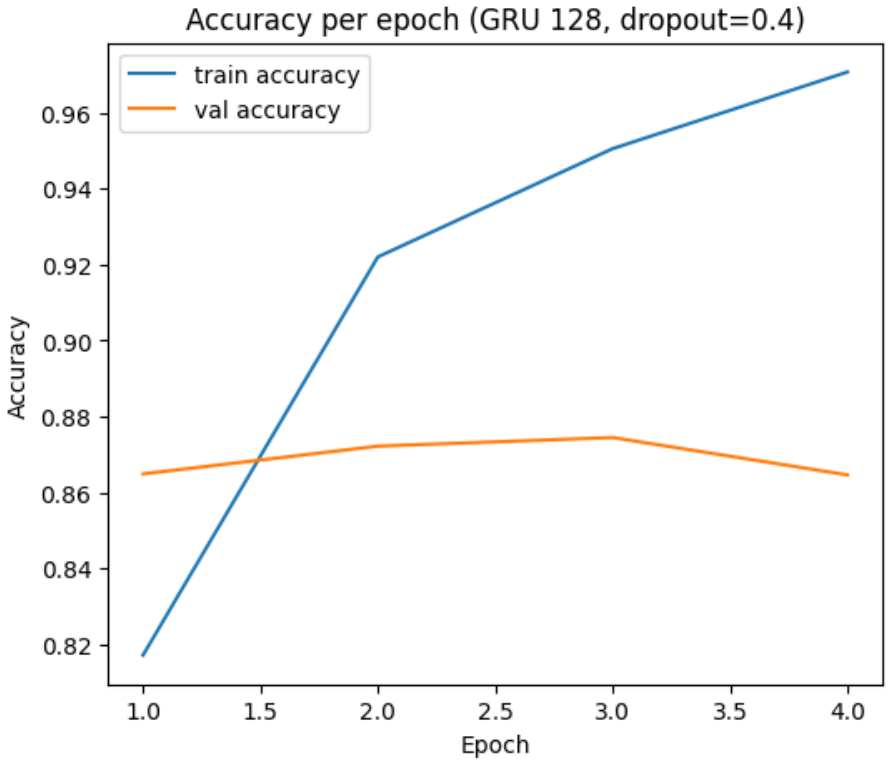
3.3 Grafici i vizuelizacije



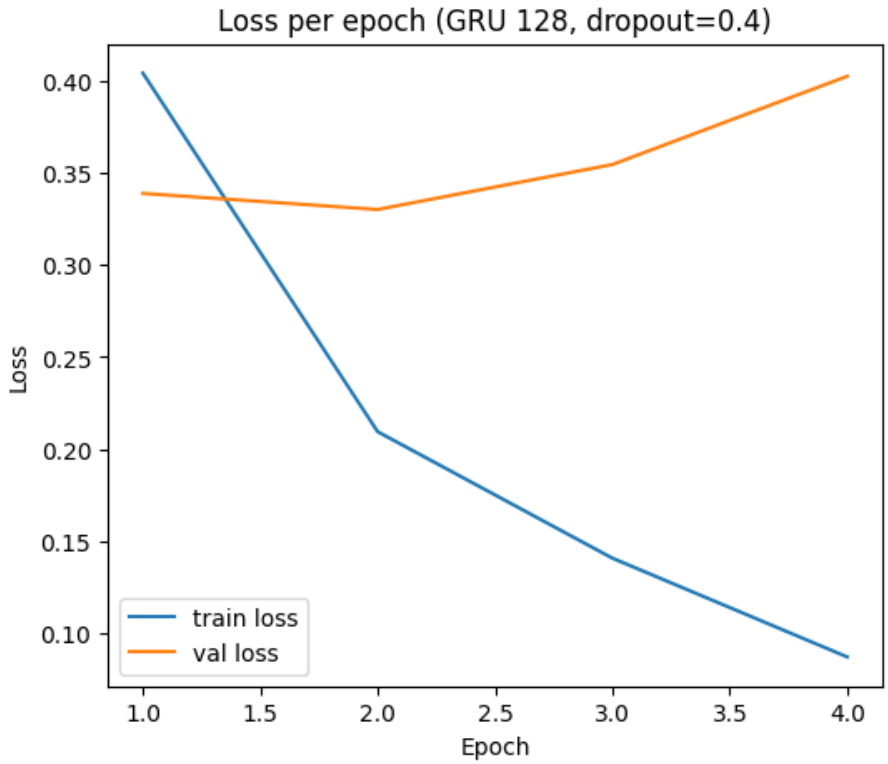
Slika 1: Kriva učenja – tačnost tokom epoha za LSTM (128 jedinica, dropout=0.4)



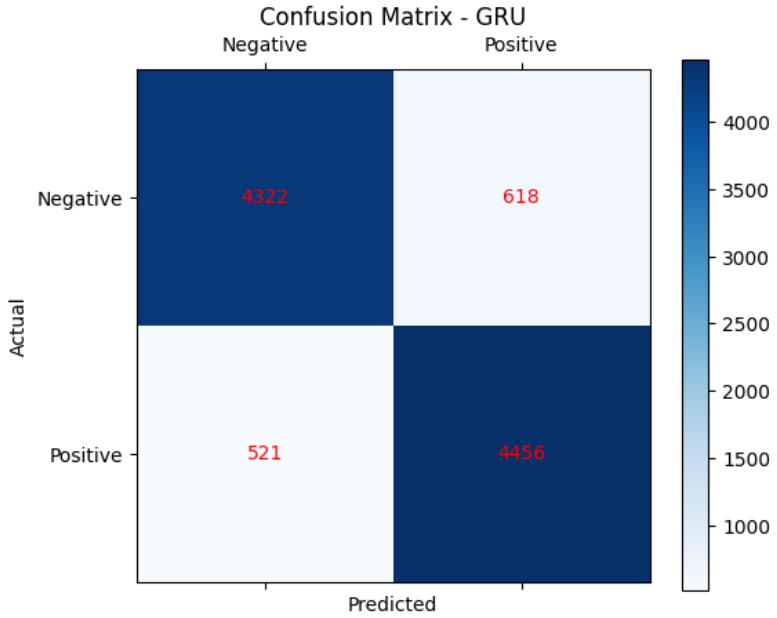
Slika 2: Kriva učenja – gubitak tokom epoha za LSTM (128 jedinica, dropout=0.4)



Slika 3: Kriva učenja – tačnost tokom epoha za GRU (128 jedinica, dropout=0.4)



Slika 4: Kriva učenja – gubitak tokom epoha za GRU (128 jedinica, dropout=0.4)



Slika 5: Matrica konfuzije za najbolji model: GRU(64, 0.2)

3.4 Diskusija

Dobijeni rezultati su u skladu sa poznatim rezultatima iz literature za IMDb 50k skup podataka. U radovima koji koriste TF-IDF reprezentaciju, tipično se navodi da Naive Bayes postiže oko 85% tačnosti, dok SVM dostiže oko 87–88%. Takođe, standardni LSTM i GRU modeli najčešće ostvaruju rezultate u opsegu 86–89% tačnosti. U našem radu dobijene vrednosti tačnosti i F1-score se nalaze upravo u ovom opsegu, što potvrđuje korektnu implementaciju i fer eksperimentalno poređenje.

Na osnovu dobijenih rezultata i grafičkih prikaza može se uočiti da se GRU model sa 64 skrivene jedinice i dropout=0.2 pokazao kao najbolji u pogledu ukupnih performansi. Matrica konfuzije ovog modela pokazuje dobru balansiranost između klasa, uz približno jednak broj grešaka kod pozitivnih i negativnih recenzija, što ukazuje na stabilnu generalizaciju.

Radi fer poređenja LSTM i GRU arhitektura, analizirane su krive učenja za modele sa identičnim hiperparametrima (128 skrivenih jedinica, dropout=0.4). Kod složenijih neuronskih modela uočena je pojava overfitting-a, što se manifestuje rastom validacionog gubitka uprkos poboljšanju trening performansi, ali je ovaj efekat ublažen primenom regularizacije i early stopping mehanizma.

GRU model sa istim parametrima pokazuje nešto stabilnije validacione krive u odnosu na LSTM, što se može objasniti manjim brojem parametara i jednostavnijom arhitekturom. Ovi rezultati potvrđuju da, u zavisnosti od zadatka i podataka, jednostavniji modeli mogu imati prednost u odnosu na složenije arhitekture.

Svi modeli pokazuju poteškoće u klasifikaciji recenzija koje sadrže negacije, implicitni sentiment ili veoma kratak tekst, što predstavlja uobičajeno ograničenje pristupa zasno-

vanih na leksičkim i sekvencijalnim reprezentacijama.

4 Zaključak

Rezultati eksperimenata pokazuju da Naive Bayes model, iako jednostavan i veoma brz za treniranje, ostvaruje solidne performanse, ali zaostaje u odnosu na ostale metode zbog pretpostavke nezavisnosti reči. SVM model sa TF-IDF reprezentacijom postiže značajno bolje rezultate i pokazuje se kao veoma dobar bazni model, naročito u problemima klasifikacije teksta sa visokodimenzionalnim ulaznim prostorom.

Neuronski modeli zasnovani na LSTM i GRU arhitekturama ostvaruju bolje ukupne performanse u odnosu na klasične metode, što potvrđuje njihovu sposobnost da efikasnije modeluju redosled reči i kontekst u tekstu. Analiza uticaja hiperparametara pokazala je da povećanje broja skrivenih jedinica i primena adekvatne regularizacije mogu poboljšati performanse LSTM modela, dok se kod GRU modela kao optimalna pokazala jednostavnija konfiguracija sa manjim brojem jedinica.

Najbolje performanse u ovom radu ostvario je GRU model sa 64 skrivene jedinice i dropout=0.2, koji je postigao najveće vrednosti accuracy i F1-score, uz stabilno ponašanje tokom treninga i dobru generalizaciju, što je potvrđeno i analizom matrice konfuzije.

Pored korišćenja ugrađenih neuronskih slojeva, implementirane su i ručne verzije LSTM i GRU modela. Iako su njihove performanse blago ispod ili uporedive sa bibliotečkim implementacijama, ovi modeli imaju značajnu edukativnu vrednost, jer omogućavaju detaljnije razumevanje unutrašnjih mehanizama rekurentnih neuronskih mreža.

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.

Naši rezultati su u skladu sa poznatim rezultatima iz literature za IMDb 50k skup. NB i SVM modeli postižu tačnosti oko 85–88%, dok standardni LSTM/GRU modeli bez dodatnih mehanizama tipično dostižu 86–89%. U tom smislu, dobijeni rezultati potvrđuju da su implementirani modeli stabilni i konkurentni u odnosu na standardne pristupe opisane u literaturi.

5 Literatura

1. IMDb Dataset of 50K Movie Reviews (Kaggle): `https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews`

2. Scikit-learn dokumentacija (TF-IDF, SVM, Naive Bayes): `https://scikit-learn.org/stable/`

3. TensorFlow/Keras dokumentacija (Embedding, LSTM, GRU): https://www.tensorflow.org/api_docs

4. Dhurba Subedi, Nabin Lamichhane, Nabaraj Subedi (2025.) *Sentiment Analysis Using Machine Learning Algorithms on IMDb Movie Review Dataset*. https://nepjol.info/index.php/jes2/article/view/70138/60783

5. Shubham Kumar, Neetu Singla (2024). *Sentiment Analysis on IMDb Review Dataset Using LSTM Networks*. https://www.researchgate.net/publication/377024012_Sentiment_Analysis_on_IMDB_Review_Dataset

6. Islam, M. T., Parvin, F., Sazan, S. A., Amir, T. B. (2024). *Comparative Analysis of Sentiment Classification on IMDb 50k Movie Reviews: A Study Using CNN, LSTM, CNN-LSTM, and BERT Models*.

7. Dai, A. M., Le, Q. V. (2015). *Semi-supervised Sequence Learning*. Advances in Neural Information Processing Systems (NeurIPS). https://arxiv.org/abs/1511.01432

8. Tang, D., Qin, B., Liu, T. (2015). *Document Modeling with Gated Recurrent Neural Network for Sentiment Classification*. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP). https://arxiv.org/abs/1506.02057