QA using transformers

Simona Jevtovic

Opis problema

- Pravljenje sistema koji generise odgovor na zadato pitanje, postoje dve vrste takvih sistema:

  1. generativni sistemi
  2. **ekstraktivni sistemi**

- Ekstraktivni sistemi se mogu napraviti uz pomoc modela koji ce pronaci odgovor u tekstu koji mu je prosledjen zajedno sa pitanjem

- To se postize treniranjem modela nad velikom kolicinom podataka

Treniranje modela

- Trening modela je kompleksan i odvija se u vise delova:
  1. **pre-training**: treniranje modela nad velikom kolicinom podataka kako bi stekao bazicno razumevanje teksta
  2. **fine tuning:** prilagodjavane na konkretan zadatak koji ce obavljati

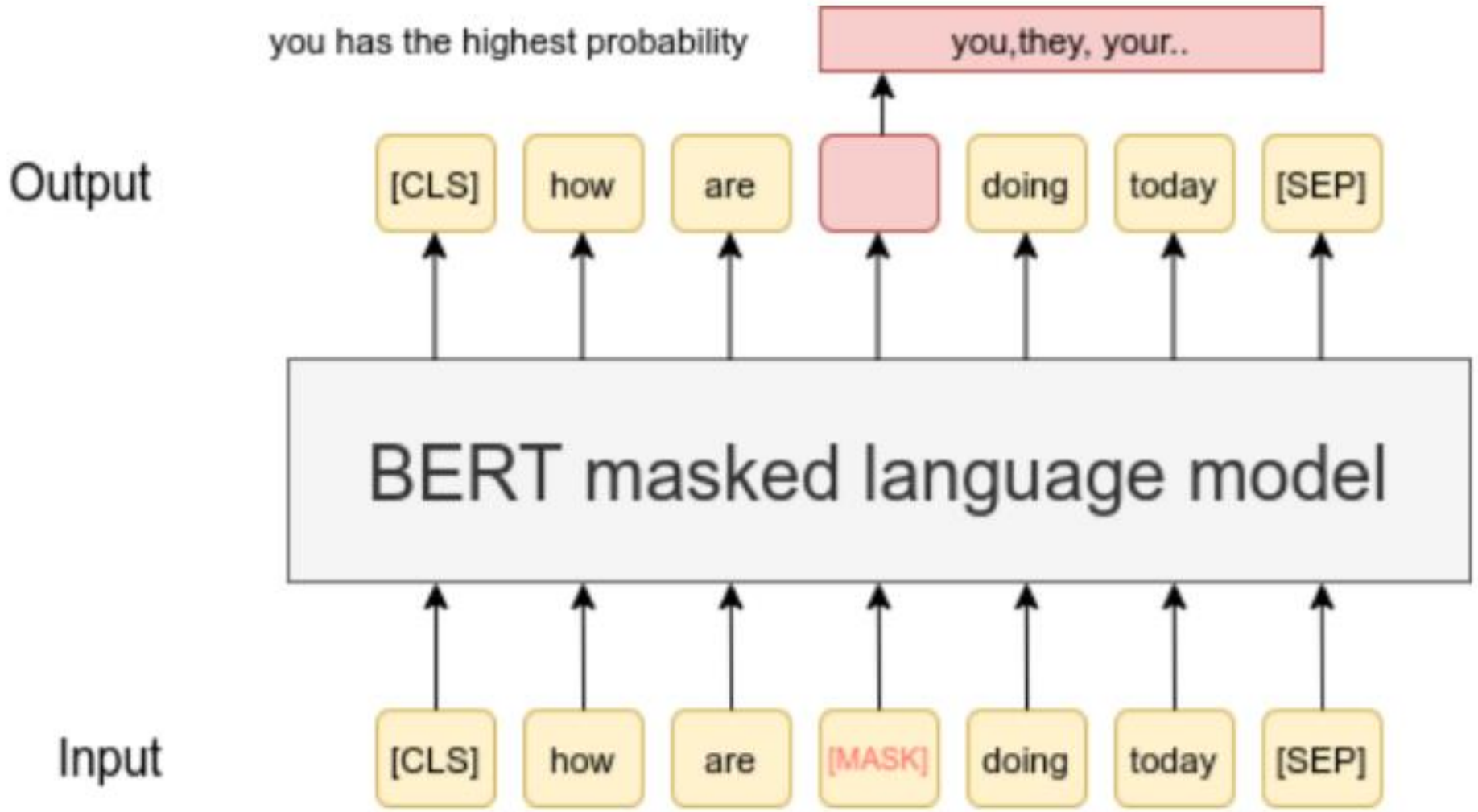
- Postoje pretrenirani modeli dostupni za koriscenje, neki od njih su:

  BERT, RoBERTa, DistilBERT, ALBERT, Electra..

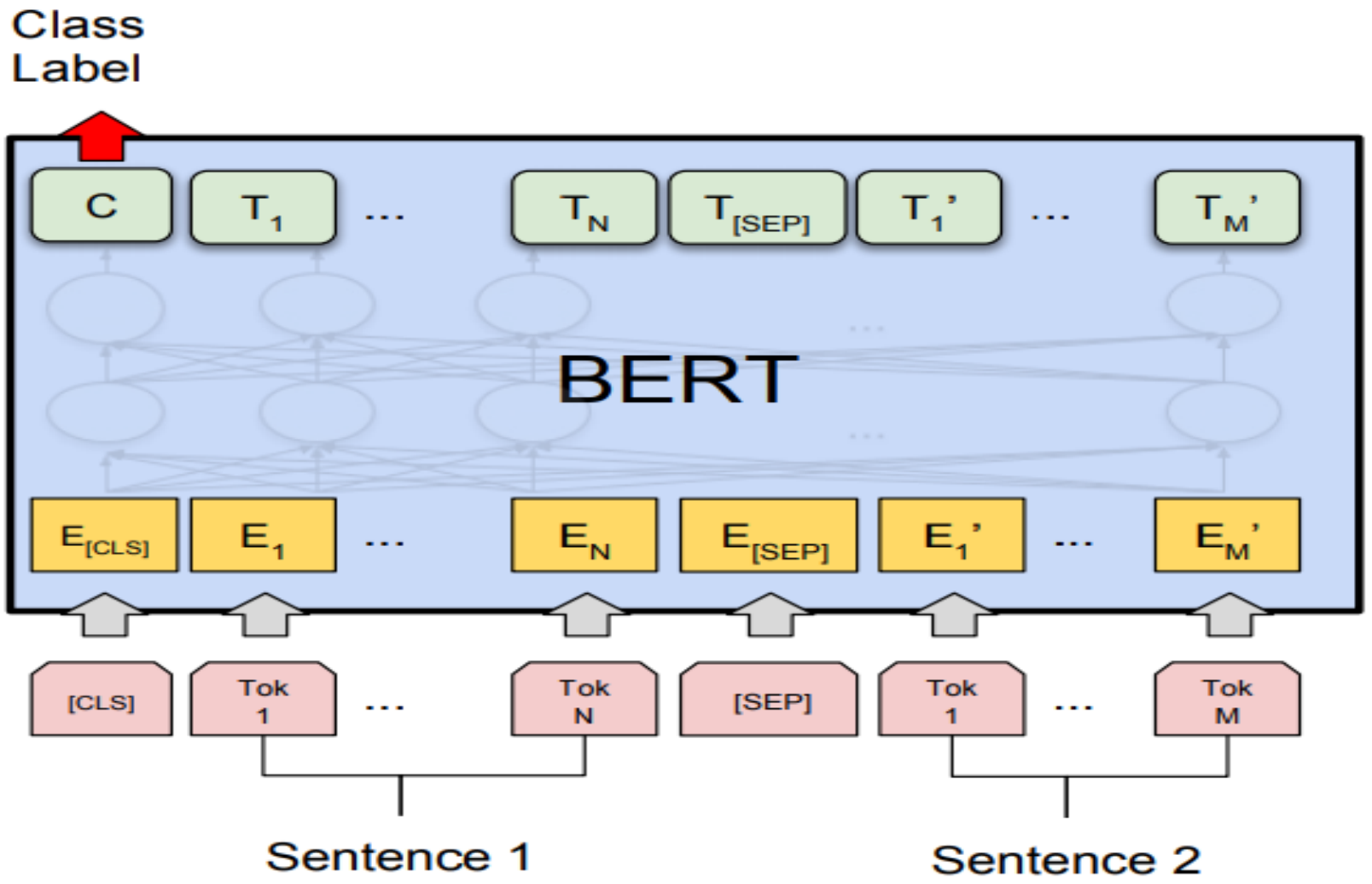
BERT

- Zasnovan je na arhitekturi transformera, sastoji se iz skupa enkodera
- Bidirekcioni model, sto mu omogucava da bolje razume kontekst svake reci u recenici i da se fokusira na reci koje su vazne
- Pretreniranje se sastoji iz dva dela:
  1. Masked Language Modeling (MLM)
  2. Next Sentence Prediction (NSP)

Masked Language Modeling



Next Sentence Prediction



Fine tuning

- SQUAD skup podataka za trening sadrzi vise od 100 000 parova odgovora i pitanja
- Modelu se prosledjuje pitanje i kontekst u kom se nalazi odgovor
- Cilj je da model predvidi pocetni i krajnji token koji oznacavaju pocetak I kraj odgovora u kontekstu
- Da bi model razumeo ulazne podatke potrebno je izvrsiti pretprocesiranje, ulazni podaci se transformisu u tokene(uz pomoc tokenizer-a)

Fine tuning

- Razlicite kombinacije parametara learning rate(1e-5, 3e-5), number of epochs(2, 3) and batch size(8,16)

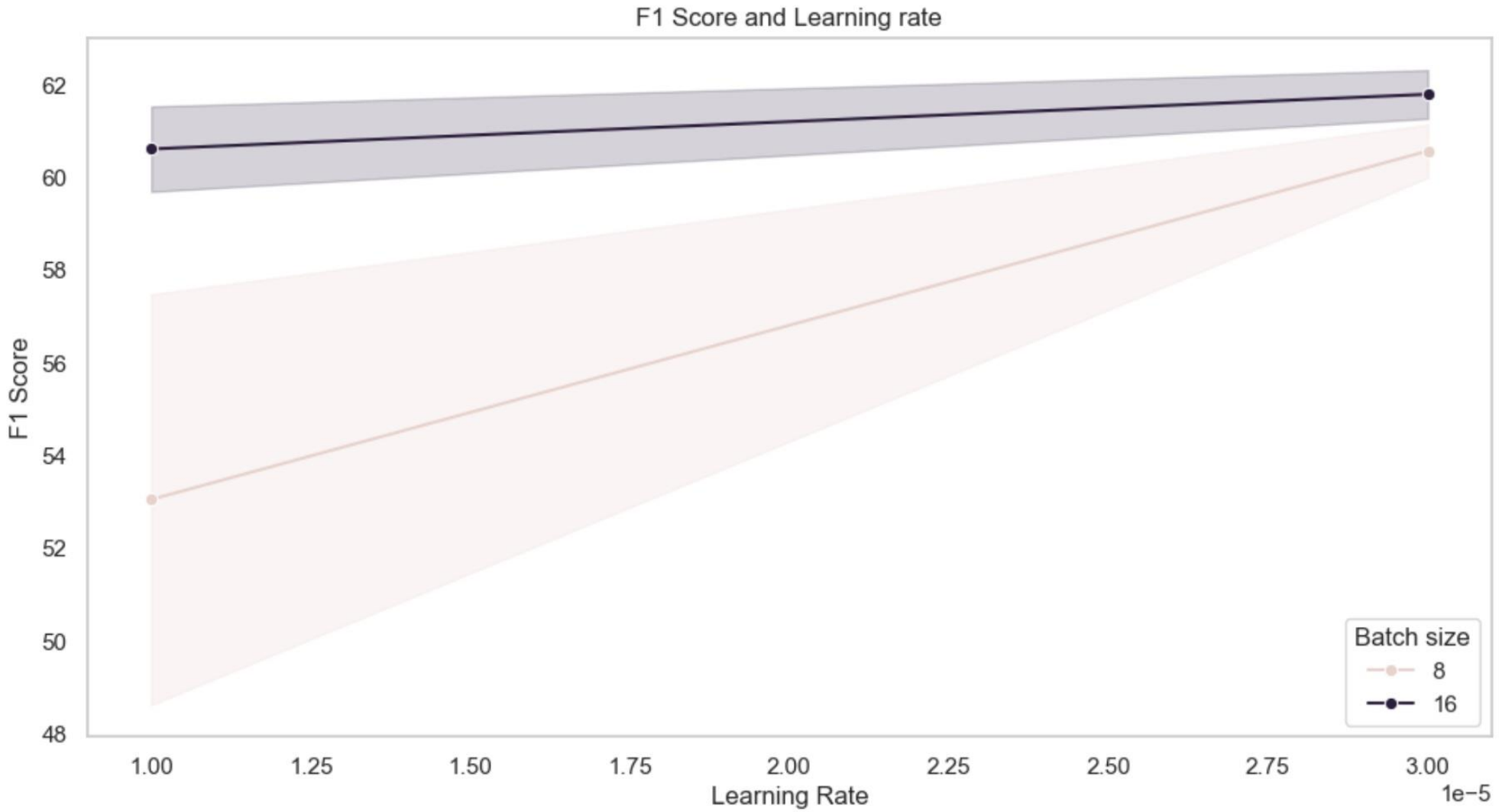
- Skup podataka za trening sadrzi 5000 instanci a skup za validaciju 500 instanci

- Tokom treniranja pracena je vrednost funkcije gubitka (loss function)

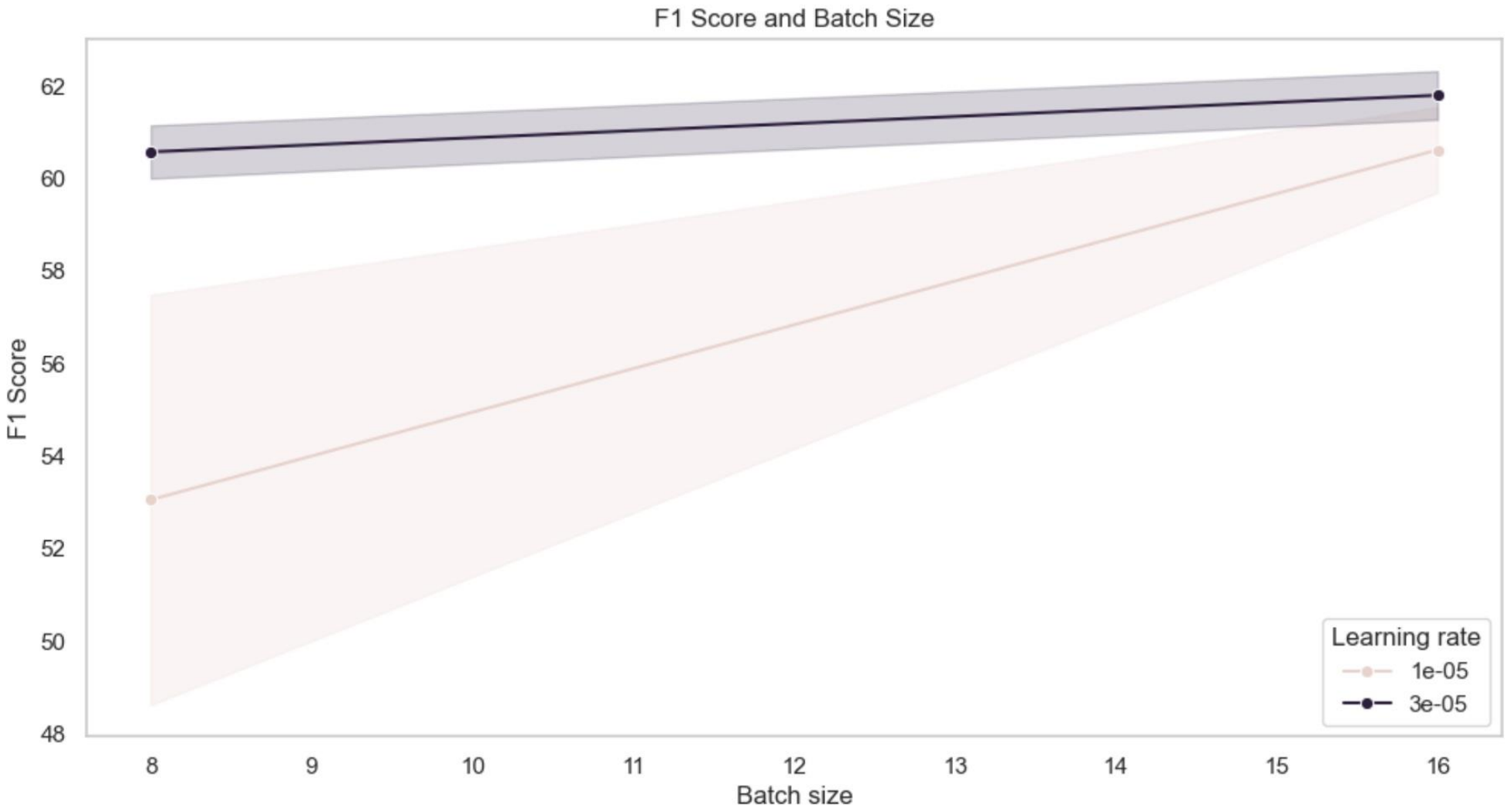
Evaluacija

- Predikcije modela su vrednosti koje oznacavaju verovatnocu da je svaki token iz konteksta pocetni ili krajnji token odgovora
- Takav izlaz potrebno je procesirati kako bi se doslo do dela konteksta koji oznacava odgovor
- Prilikom evaluacije uporedjuje se predvidjeni odgovor sa tacnim odgovorom
- Parametri za evaluaciju: **F1 score**, i **Exact Match (EM)**

Evaluacija



Evaluacija



Evaluacija

	learning_rate	batch_size	num_epochs	exact_match	f1
0	0.00001	8	2	36.8	48.619631
1	0.00001	8	3	46.2	57.479887
2	0.00001	16	2	49.4	59.695797
3	0.00001	16	3	50.4	61.534543
4	0.00003	8	2	50.0	61.144712
5	0.00003	8	3	47.4	59.998342
6	0.00003	16	2	48.8	62.320345
7	0.00003	16	3	49.0	61.272149

Hvala! ☺