Domaći zadatak 2

Izveštaj o fine-tuningu RoBERTa modela za klasifikaciju teksta

1. Učitavanje dataseta

Koristila sam “ag_news” skup podataka. Dostupan je putem Hugging Face datasets biblioteke, što olakšava njegovo preuzimanje i pripremu za trening. AG News je popularan dataset koji se koristi za klasifikaciju tekstova, posebno za zadatke klasifikacije vesti. Dataset je namenjen za evaluaciju modela u prepoznavanju i klasifikaciji vesti u različite kategorije. Ovaj dataset sadrži:

- **Tekstove**: Svaka instanca u datasetu je članak vesti.
- **Kategorije**: Članci su podeljeni u četiri glavne kategorije:
    - **World**: Globalne vesti.
    - **Sports**: Sportske vesti.
    - **Business**: Poslovne vesti.
    - **Science/Technology**: Vesti iz nauke i tehnologije.

Da bi ubrzala trening proces i eksperimentisanje sa hiperparametrima, smanjila sam dataset na 10% originalne veličine za train i test skupove.

2. Tokenizer i model

Koristila sam **RobertaTokenizerFast** za tokenizaciju. To je brža verzija tokenizatora specifična za RoBERTa (Robustly optimized BERT approach) model. Tokenizer je alat koji pretvara tekstualne podatke u numeričke vrednosti koje model može da obradi. Za model je korišćen **RobertaForSequenceClassification** sa 4 izlazna labela, to znači da model može predviđati jednu od četiri moguće klase (npr. world, sports, business i science/technology). Ovo je pretrenirani model iz RoBERTa familije, prilagođen za zadatke klasifikacije sekvenci, kao što je klasifikacija sentimenta, klasifikacija tema, detekcija lažnih vesti itd.

3. Tokenizacija podataka

Tokenizovala sam dataset koristeći funkciju koja koristi padding i truncation, sa maksimalnom dužinom sekvence od 128. Kolona sa oznakama klasa se preimenuje da bi se uskladila sa očekivanjima modela. Model očekuje da se oznake klase nalaze u koloni sa nazivom “labels”.

4. Konfiguracija Trainer-a i treniranje modela

Trainer klasa objedinjuje sve potrebne komponente i olakšava proces treninga, evaluacije, i optimizacije hiperparametara.

TrainingArguments: Argumenti koji definišu kako će se trening odvijati (npr. broj epoha, veličina batch-a, brzina učenja).

Data collator: Koristila sam **DataCollatorWithPadding** iz transformers biblioteke. Omogućava da se svaki batch dinamički podstavi na istu dužinu, što olakšava rad sa batch-evima različitih dužina i optimizuje performanse tokom treniranja modela.

Metrike:Implementirala sam funkciju za računanje važnih metrika uključujući tačnost (accuracy), F1 score, preciznost i odziv. Ove metrike pružaju sveobuhvatan uvid u performanse modela.

5. Eksperimentisanje sa Hiperparametrima

Eksperimentisanje sa hiperparametrima je ključni deo procesa treniranja modela, jer može značajno uticati na performanse modela. Eksperimentisala sam sa sledećim hiperparametrima:

- Learning rates: [1e-5, 3e-5, 5e-5] - Ovi opsezi su tipični za fine-tuning BERT-baziranih modela.
- Batch sizes: [8, 16, 32] - Manji batch size-ovi su često efikasniji za fine-tuning, ali sam htela da istražim i veće vrednosti.
- Broj epoha : 3

6. Rezultati

Tokom eksperimentisanja sa različitim vrednostima hiperparametara, najbolje performanse su postignute sa sledećom konfiguracijom koristeći learning rate od 3e-5 i batch size od 32. Ovaj model je postigao impresivan F1 skor od 0.9461 što je pokazatelj uspešnog prilagođavanja modela na zadatak klasifikacije.

Learning rate od 3e-5 se pokazao kao najbolji, dajući F1 skor od 0.9461. Drugi learning rate-ovi (1e-5 i 5e-5) dali su nešto slabije rezultate, sa F1 skorom oko 0.93. **Previše nizak learning rate (1e-5)**: Može da dovede do sporijeg konvergiranja, što znači da model može biti pod-treniran u zadatom broju epoha, a to može rezultirati nižim performansama. **Previše visok learning rate (5e-5)**: Može da uzrokuje da model brzo nauči, ali i da se preskoče neka važna lokalna minimuma, što može rezultirati prenaučenjem ili nedovoljno dobro uvežbanim modelom. **Optimalni learning rate (3e-5)**: Ova vrednost omogućava balans između brzine konvergencije i preciznosti modela, što je dovelo do najboljih rezultata.

Batch size od 32 dao je najbolje rezultate, dok su manji batch size-ovi (8 i 16) imali blago lošije rezultate. **Manji batch size (8, 16)**: Može da doprinese boljoj generalizaciji modela, ali uz cenu većih varijacija u toku treninga. **Veći batch size (32)**: Obezbeđuje stabilniju i bržu konvergenciju, što može pomoći modelu da brže pronađe optimalna rešenja u prostoru hiperparametara, što rezultira višim F1 skorom. Međutim, postoji rizik od prenaučenosti ako je batch size prevelik, ali u ovom slučaju to nije bio problem.

Zaključak

Najbolji rezultati su postignuti sa learning rate-om od 3e-5 i batch size-om od 32, što pokazuje da je ovaj model pronašao dobar balans između brzine učenja i stabilnosti treninga. Rezultati sugerišu da je pažljivo podešavanje hiperparametara ključno za postizanje visokih performansi u zadacima klasifikacije teksta.