UNIVERZITET U NIŠU
ELEKTRONSKI FAKULTET

Poređenje performansi pretreniranih RoBERTa modela za analizu sentimenta na Amazon Reviews datasetu

SEMINARSKI RAD

Predmet:

Web Mining

Mentor: prof. dr Miloš Bogdanović

Kandidat: Milica Stojanović (1701)

Niš, 2025.

Sadržaj

1. Uvod

Analiza sentimenta je postala ključni alat u savremenom digitalnom svetu, s obzirom na njenu široku primenu u različitim oblastima kao što su društvene mreže, marketing, istraživanje tržišta i mnoge druge. Ova tehnika iz oblasti obrade prirodnog jezika (NLP) omogućava identifikaciju emocionalnog tona u tekstu, bilo da je pozitivan, negativan ili neutralan. Tako se pruža dublje razumevanje reakcija korisnika na proizvode, usluge, teme i druge aspekte svakodnevnog života. Kompanije koriste analizu sentimenta kako bi pratili utiske korisnika o svojim proizvodima i uslugama, prilagodili svoje strategije i poboljšali korisničko iskustvo. Ova tehnika takođe omogućava potencijalnim kupcima da procene opšte mišljenje o proizvodima ili uslugama pre nego što donesu odluku o kupovini.

Pored toga, analiza sentimenta ima značajnu primenu u identifikaciji i filtriranju nepoželjnih sadržaja kao što su govor mržnje, nacionalizam i drugi problematični tekstovi, što doprinosi očuvanju bezbednosti i integriteta online prostora. U oblasti finansija, ova tehnika može pomoći u praćenju sentimenta investitora i tržišta, što omogućava predviđanje tržišnih trendova i donošenje informisanih investicionih odluka. Političari i javne organizacije takođe koriste analizu sentimenta za praćenje mišljenja građana i prilagođavanje svojih politika.

Revolucija u obradi prirodnog jezika dolazi sa pojavom transformers arhitekture, koja značajno unapređuje metode razumevanja i generisanja teksta. Ključna komponenta ove arhitekture je mehanizam pažnje, koji omogućava modelima da uče i razumeju dugoročne zavisnosti u tekstualnim podacima. Uz to, transformers modeli omogućavaju paralelnu obradu podataka, što smanjuje potrebe za resursima tokom obuke i primene modela, čime se postižu izuzetni rezultati uz manje troškove. Ova arhitektura takođe omogućava primenu prenesenog učenja, gde se prethodno obučeni modeli dodatno doteruju za specifične zadatke, čime se dodatno poboljšava njihova preciznost i efikasnost.

Ovaj rad se bavi analizom sentimenta koristeći Amazon Reviews dataset i upoređuje tri različita pretrenirana RoBERTa modela: distilroberta-base, distilroberta-base-sst2-distilled i twitter-roberta-base-sentiment. Cilj je da se oceni kako svaki od ovih modela obavlja zadatak prepoznavanja i klasifikacije sentimenta u recenzijama proizvoda.

U analizi su istražene metodologije obuke i evaluacije svakog modela, kao i njihovi rezultati u pogledu ključnih metrika: preciznosti, odziva, F1 tačnosti i korelacione matrice. Ove metrike omogućavaju detaljno upoređivanje performansi modela i identifikovanje njihovih prednosti i slabosti. Na osnovu ovih saznanja, biće pružene preporuke za izbor najefikasnijeg RoBERTa modela za analizu sentimenta, uz razmatranje kako se ovi modeli mogu koristiti u različitim poslovnim i istraživačkim kontekstima.

2. Opis skupa podataka

Za potrebe ovog istraživanja korišćen je skup podataka sa platforme Kaggle, koji sadrži korisničke recenzije proizvoda objavljene na platformi Amazon. Reč je o jednom od najpoznatijih i najobimnijih javno dostupnih skupova za zadatke sentiment analize, jer obuhvata veliki broj raznovrsnih primera koji odražavaju autentična iskustva korisnika iz različitih kategorija proizvoda. Ovaj skup se često koristi kao u istraživanjima iz oblasti obrade prirodnog jezika (NLP) i mašinskog učenja.

Skup podataka sadrži ukupno 4 miliona recenzija, koje su podeljene u dva podskupa: trening skup sa 3 miliona i 600 hiljada primera, koji se koristi za obuku modela, i test skup sa 400 hiljada primera, namenjen evaluaciji i proveri tačnosti modela. Ovakva podela omogućava procenu generalizacione sposobnosti modela — odnosno, koliko dobro model može da prepozna sentiment na podacima koje nije "video" tokom procesa učenja.

Svaka recenzija u skupu podataka ima sledeću strukturu (Slika 2.1.):

- **title** – naslov recenzije koji korisnik ostavlja. U većini slučajeva ovaj naslov sadrži kratak, sažet opis mišljenja o proizvodu.
- **review** – glavni deo recenzije, u kojem korisnik detaljnije iznosi svoje iskustvo i utiske. Ovaj deo je od ključne važnosti za model jer nosi najveći deo semantičke informacije.
- **label** – oznaka sentimenta recenzije, pri čemu je 1 dodeljeno negativnim recenzijama, a 2 pozitivnim. Ova oznaka predstavlja ciljnu promenljivu (target), koju model pokušava da predvidi.

	label	title	review
0	2	Great CD	My lovely Pat has one of the GREAT voices of h...
1	2	One of the best game music soundtracks - for a...	Despite the fact that I have only played a sma...
2	1	Batteries died within a year ...	I bought this charger in Jul 2003 and it worke...
3	2	works fine, but Maha Energy is better	Check out Maha Energy's website. Their Powerex...
4	2	Great for the non-audiophile	Reviewed quite a bit of the combo players and ...

Slika 2.1. Skup podataka

Recenzije su napisane na engleskom jeziku, a obuhvataju širok spektar kategorija — od elektronskih uređaja i kućnih potrepština, preko knjiga i igračaka, do odeće i kozmetike. Zbog takve raznovrsnosti, ovaj skup podataka omogućava treniranje modela koji mogu prepoznati obrasce sentimenta u vrlo različitim kontekstima i stilovima pisanja. Takođe, veličina i raznolikost skupa podataka čine ga izuzetno pogodnim za upotrebu u dubokom učenju, jer veliki broj primera pomaže modelima da uče stabilnije i preciznije.

Pre početka pripreme i obrade podataka, izvršena je detaljna provera integriteta skupa kako bi se obezbedio kvalitet ulaznih informacija za treniranje modela. Analizom je utvrđeno da skup podataka sadrži prazne redove u koloni title, i to 207 primera u trening skupu i 24 primera u test skupu, dok su kolone label i review bile kompletne. Svi identifikovani prazni redovi su uklonjeni iz skupa, čime je obezbeđeno da model ne prima nedovoljno informativne

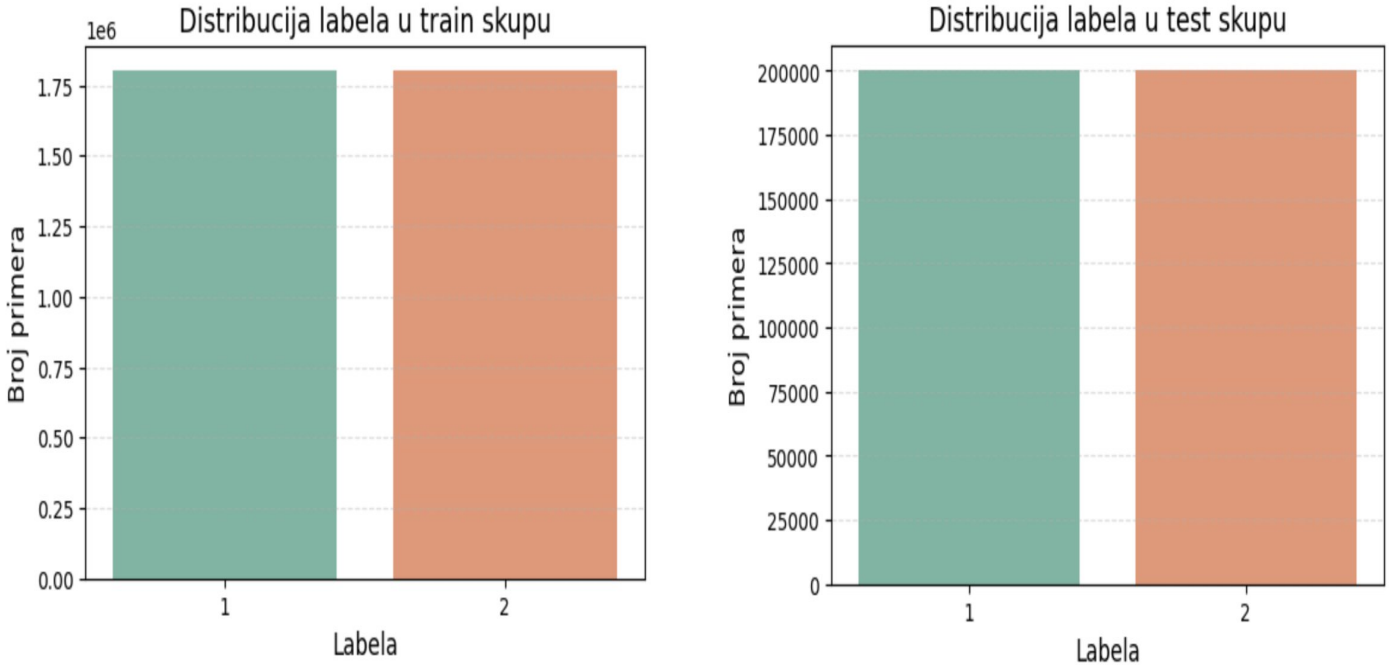
ili nepotpune tekstualne ulaze. Takođe je izvršena i provera duplikata, pri čemu je utvrđeno da u podacima ne postoje duplikati, što dodatno doprinosi kvalitetu i konzistentnosti obuke.

Budući da korisničke recenzije sadrže i naslov (title) i sadržaj (review), odlučeno je da se ove dve kolone spoje u jednu celinu (Slika 2.2.). Na taj način zadržane su sve raspoložive informacije, čime se smanjuje rizik od gubitka važnog konteksta koji može doprineti kvalitetnijoj analizi sentimenta. Naslovi često sadrže sažete emocionalne izraze i ključne reči koje mogu značajno poboljšati performanse modela u prepoznavanju polariteta sentimenta.

	label	text
0	2	Great CD My lovely Pat has one of the GREAT vo...
1	2	One of the best game music soundtracks - for a...
2	1	Batteries died within a year ... I bought this...
3	2	works fine, but Maha Energy is better Check ou...
4	2	Great for the non-audiophile Reviewed quite a ...

Slika 2.2. Skup podataka (title + review)

Još jedna važna karakteristika skupa podataka jeste njegova uravnoteženost. U trening skupu se nalazi 1.800.000 pozitivnih i 1.800.000 negativnih primera, dok se u test skupu nalazi po 200.000 primera svake klase (Slika 2.3.). Ovakva ravnomerna raspodela klasa je od suštinskog značaja jer sprečava pojavu pristrasnosti modela prema jednoj od klasa, čime se obezbeđuje stabilnija i tačnija klasifikacija. Uravnotežen dataset predstavlja dobru osnovu za pouzdanu evaluaciju performansi modela i omogućava realniju procenu njihove sposobnosti da generalizuju na nove podatke.

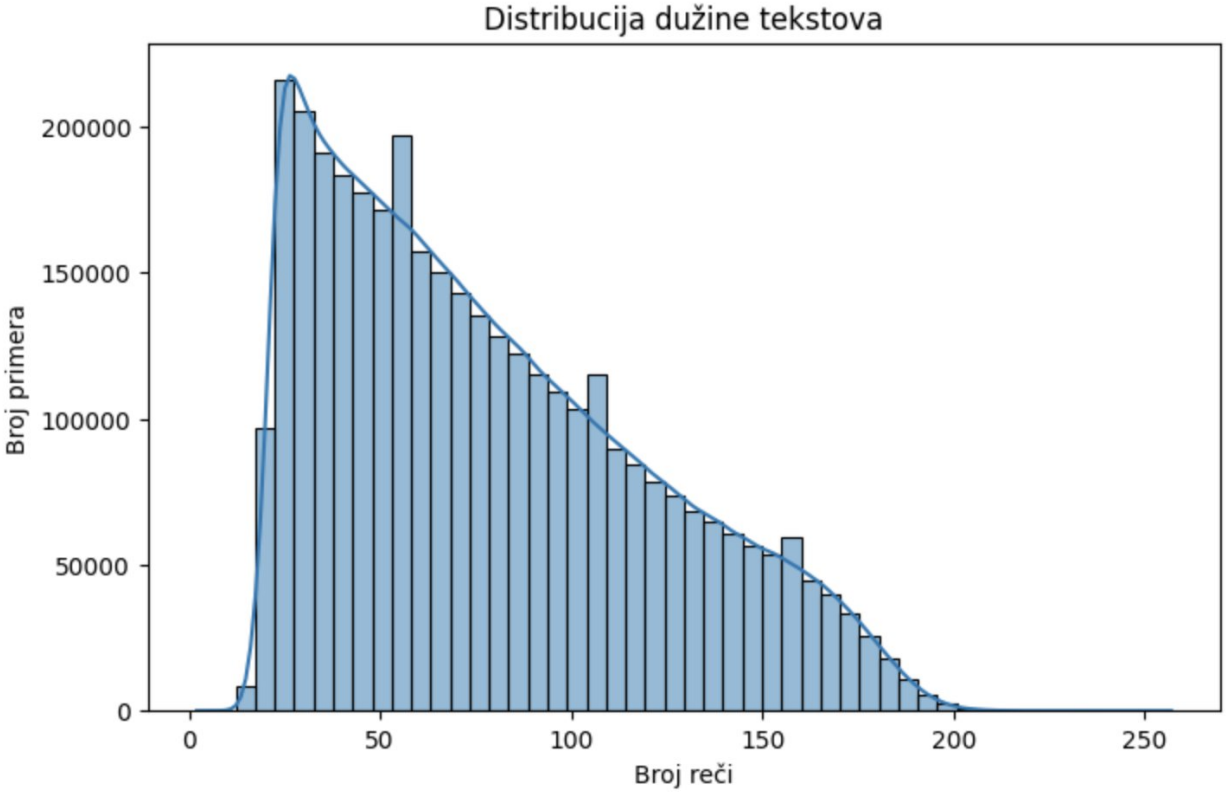


Slika 2.3. Distribucija labela (klasa) za train i test skup

2.1. Analiza dužine tekstova

Kako bi se dobio uvid u karakteristike samih tekstualnih podataka, izvršena je analiza distribucije dužine recenzija, izražene kroz broj reči po primeru.

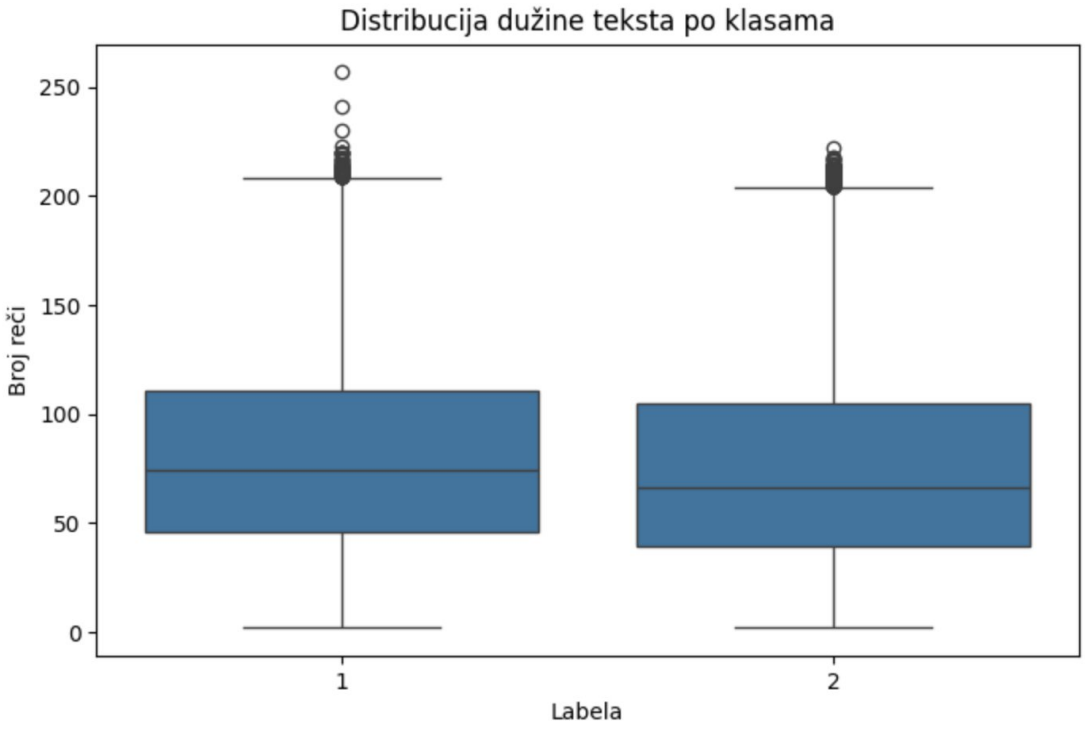
Na prvom grafikonu (Slika 2.1.1.) prikazana je distribucija dužine svih tekstova u trening skupu . Uočava se da najveći broj recenzija sadrži manje od 100 reči, dok se vrh distribucije (moda) nalazi između 20 i 40 reči. Kriva pokazuje postepeni pad, što je očekivano za korisničke recenzije — većina korisnika piše kratke i sažete komentare, dok manji broj piše dugačke tekstove koji prelaze 150 reči. Ova analiza je značajna jer pomaže pri određivanju optimalne maksimalne dužine sekvenci prilikom tokenizacije, čime se izbegava nepotrebno trošenje memorijskih resursa prilikom treniranja modela.



Slika 2.1.1. Distribucija dužine tekstova

Na drugom grafikonu (boxplot prikaz, Slika 2.1.2.), analizirana je dužina tekstova po klasama — pozitivnim i negativnim recenzijama. Uočava se da je distribucija dužina slična u obe klase: većina recenzija je kratka, dok se ekstremne vrednosti (outlieri) odnose na veoma dugačke tekstove koji se javljaju u manjem broju slučajeva. Medijana za obe klase se nalazi oko 70 reči, što ukazuje da ne postoji značajna razlika u dužini teksta između pozitivnih i negativnih recenzija. Ovo je važan nalaz jer potvrđuje da dužina teksta ne utiče direktno na polaritet sentimenta, već da su semantičke informacije ključne za klasifikaciju.

Dodatno, outlieri koji predstavljaju neuobičajeno dugačke tekstove ukazuju na to da će biti potrebno uvesti maksimalnu dužinu tokena u procesu tokenizacije kako bi se ograničila prevelika sekvenca i time optimizovalo korišćenje GPU memorije. Na osnovu ovih rezultata, određena je maksimalna dužina sekvence koja pokriva većinu primera bez značajnog gubitka informacija.



Slika 2.1.2. Distribucija dužine teksta po klasama

2.2. Analiza najčešće korišćenih reči (Word Cloud)

Kako bi se stekao uvid u najčešće korišćene pojmove u recenzijama i identifikovali obrasci koji se spontano pojavljuju u jeziku korisnika, primenjena je vizualizacija pomoću word cloud tehnike. Ova tehnika prikazuje reči proporcionalno njihovoj učestalosti u korpusu — što je reč češće prisutna, to je prikazana krupnije i istaknutije. Time se omogućava brza, intuitivna analiza sadržaja bez potrebe za dubokom statističkom obradom u ovoj fazi.

Word cloud je generisan posebno za pozitivne i negativne recenzije, kako bi se uočile razlike u jeziku i izražavanju korisnika. Kod pozitivnih recenzija, najzastupljenije reči uključuju izraze poput “good”, “great”, “love”, “best”, “excellent” i slične pozitivno obojene termine (Slika 2.2.1.). Ove reči jasno odražavaju zadovoljstvo i afirmativno iskustvo korisnika, i predstavljaju snažan signal za klasifikaciju pozitivnog sentimenta.

Slika 2.2.1 Najčešće reči u pozitivnim recenzijama

Nasuprot tome, kod negativnih recenzija, dominiraju reči kao što su “bad”, “poor”, “waste”, “disappointed”, “broke” i slične, koje ukazuju na nezadovoljstvo, loše iskustvo i negativne emocije (Slika 2.2.2.).

Slika 2.2.2. Najčešće reči u negativnim recenzijama

Ovakav kontrast u učestalosti ključnih reči potvrđuje da lingvistički obrasci između pozitivnih i negativnih recenzija zaista postoje i jasno su izraženi, što je veoma korisno u kontekstu učenja modela. Posebno je važno što su ove razlike intuitivne i lako uočljive, što ukazuje da modeli bazirani na kontekstualnom razumevanju jezika (kao što su transformeri) imaju dovoljno informacija da uspešno razlikuju klase.

Osim toga, word cloud analiza može poslužiti i kao alat za proveru kvaliteta skupa podataka — na primer, da li se pojavljuju irelevantne reči (poput URL-ova, brojeva, oznaka proizvoda), što bi ukazivalo na potrebu za dodatnim čišćenjem teksta. U ovom slučaju, podaci su prethodno detaljno obrađeni, pa se dominantno pojavljuju semantički relevantni pojmovi povezani sa korisničkim iskustvima.

2.3. Čišćenje podataka

Pre nego što se tekstualni podaci mogu koristiti za treniranje modela mašinskog učenja, neophodno je izvršiti proces čišćenja podataka. Tekstualni skupovi podataka često sadrže elemente koji nisu informativni za zadatak klasifikacije — kao što su nepotrebni simboli, URL adrese, brojevi, stop reči i višak razmaka. Uklanjanjem takvih elemenata smanjuje se šum u podacima i poboljšava kvalitet reprezentacije teksta, što direktno utiče na performanse modela. Drugim rečima, kvalitet ulaznih podataka ima presudan značaj za uspeh svake NLP analize, pa je čišćenje teksta ključno za dobijanje tačnih i pouzdanih rezultata.

U okviru ovog rada razvijena je posebna funkcija za čišćenje teksta:

```
def clean_text(text):

    if not isinstance(text, str):
        return ""

    # Make text lowercase
    text = text.lower()

    # Remove text in square brackets
    text = re.sub('\[.*?\]', '', text)

    # Remove links
    text = re.sub('https?://\S+|www\.\S+', '', text)

    # Remove punctuation
    text = re.sub('[^a-zA-Z0-9\s]+', '', text)

    # Remove words containing numbers
    text = re.sub('\w*\d\w*', '', text)

    # Remove stop words
    stop_words = set(stopwords.words('english'))
    words = text.split()

    filtered_words = [word for word in words if word not in stop_words]
    text = ' '.join(filtered_words)

    # Remove extra whitespace
    text = re.sub('\s+', ' ', text).strip()

    return text
```

Ova funkcija obavlja sledeće korake:

- Provera tipa podataka:
  Na početku se proverava da li je ulazni podatak string. Ako nije, vraća se prazan string kako bi se izbegle greške u kasnijim koracima obrade.
- Pretvaranje teksta u mala slova:
  Svi karakteri se konvertuju u mala slova (lowercase). Na taj način se obezbeđuje doslednost i izbegava da se iste reči u različitim oblicima (npr. “Good” i “good”) tretiraju kao različite.
- Uklanjanje teksta unutar uglastih zagrada:
  Tekst unutar [...] često sadrži reference, dodatne oznake ili metapodatke koji nisu relevantni za sentiment analizu.
- Uklanjanje URL adresa:
  Linkovi nemaju semantičko značenje u kontekstu sentimenta i predstavljaju šum, pa se uklanjaju pomoću regularnih izraza.
- Uklanjanje interpunkcije:
  Interpunkcijski znaci (.,!? itd.) nisu neophodni za većinu modela klasifikacije teksta i njihovo uklanjanje pojednostavljuje korpus.
- Uklanjanje reči koje sadrže brojeve:
  Brojevi često nisu informativni za sentiment analizu i mogu negativno uticati na proces tokenizacije.

- Uklanjanje stop reči:
  Stop reči (npr. “the”, “is”, “in”) su vrlo česte u jeziku, ali same po sebi ne doprinose razumevanju sentimenta. Njihovo uklanjanje smanjuje dimenzionalnost i poboljšava fokus modela na informativnije reči.
- Uklanjanje viška razmaka:
  Nakon svih transformacija, višak praznina i razmaka se uklanja kako bi tekst bio uredno formatiran i spreman za dalju obradu.

Ovim procesom značajno je smanjen šum u podacima, čime je obezbeđena čistija i standardizovana tekstualna reprezentacija. Takva reprezentacija je pogodna za sledeće korake obrade, poput tokenizacije i vektorizacije, što direktno doprinosi stabilnijem treniranju modela i boljim rezultatima.

Pre same podele podataka na skupove, izvršena je standardizacija oznaka klasa. Pošto su originalne vrednosti oznaka bile 1 i 2, kako bi se uskladilo sa standardnim formatom koji koriste mnogi modeli binarne klasifikacije, izvršena je zamena oznake 2 u 1, a oznaka 1 u 0. Na ovaj način:

- klasa 0 predstavlja negativne recenzije,
- dok klasa 1 označava pozitivne recenzije.

Ova transformacija pojednostavljuje dalji rad i omogućava lakšu integraciju sa bibliotekama za mašinsko učenje i modelima koji očekuju binarne oznake u formatu 0 i 1. Takođe, ovakav pristup je u skladu sa uobičajenom praksom u zadacima analize sentimenta.

2.4. Podela podataka

Nakon što su podaci očišćeni i pripremljeni, sledeći korak bio je njihova podela na odgovarajuće skupove kako bi se omogućilo pravilno treniranje i objektivna evaluacija modela. Podaci su podeljeni na skup za treniranje (train), skup za validaciju (validation) i test skup (test).

Podela je izvršena korišćenjem funkcije **train_test_split** iz biblioteke **scikit-learn**, uz stratifikaciju na osnovu oznaka klasa (label), čime je očuvana proporcionalna zastupljenost pozitivnih i negativnih primera u svim skupovima. Na ovaj način obezbeđeno je da svi skupovi imaju reprezentativnu raspodelu podataka, što doprinosi stabilnijem i realnijem treniranju i evaluaciji modela.

Od ukupnog skupa za treniranje, 90 % podataka je korišćeno za obuku modela, dok je 10 % izdvojeno za validaciju. Test skup je prethodno pripremljen i korišćen za nezavisnu evaluaciju finalnog modela. Ova podela omogućava praćenje performansi modela tokom treniranja i sprečava pojavu overfittinga.

Rezultujuće veličine skupova bile su sledeće:

- Train skup: 3.240.000 primera
- Validation skup: 360.000 primera
- Test skup: 400.000 primera

Ovakav balansiran pristup podeli podataka čini evaluaciju modela pouzdanijom, jer omogućava merenje generalizacione sposobnosti modela na neviđenim primerima.

3. Tokenizacija

Tokenizacija predstavlja proces pretvaranja sirovog teksta u numerički format koji je razumljiv modelu mašinskog učenja. Budući da neuronske mreže ne mogu direktno da obrađuju tekstualne podatke, tokenizator ima zadatak da svaku reč, podrečicu ili simbol iz teksta preslika u odgovarajući brojčani identifikator. Na taj način se tekstualne sekvence pretvaraju u nizove brojeva, odnosno tokene, koji čine ulazne podatke za model.

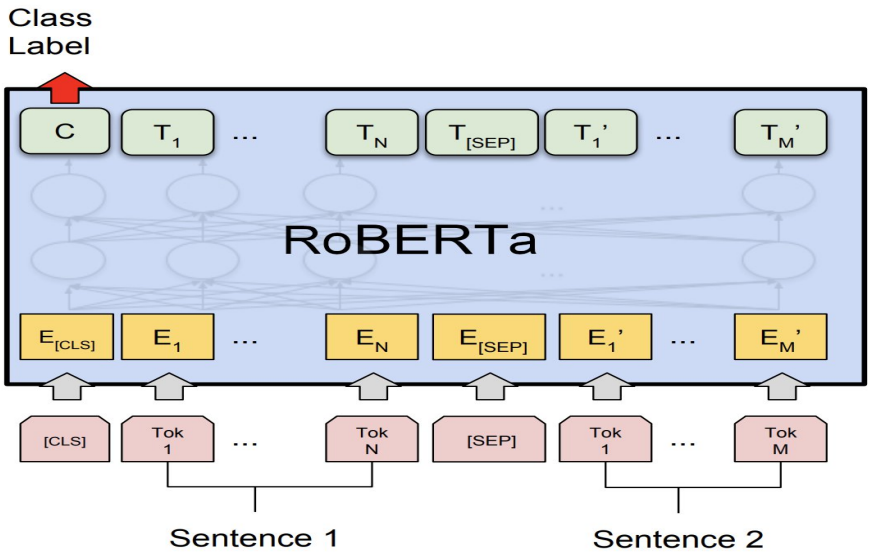
U ovom radu korišćen je tokenizator RoBERTa (Slika 3.1.), koji je posebno prilagođen istoimenom pretrenom modelu razvijenom od strane kompanije Meta AI.

Tokenizator je učitan korišćenjem komande:

tokenizer = RobertaTokenizer.from_pretrained("roberta-base")

čime se preuzima unapred definisan vokabular i pravila segmentacije teksta identična onima korišćenim tokom treniranja originalnog RoBERTa modela. Time se obezbeđuje konzistentnost između načina na koji se tekst obrađuje i formata podataka koji model očekuje.

Za razliku od klasičnih pristupa tokenizaciji, gde se tekst deli na cele reči, RoBERTa koristi napredni algoritam **Byte-Pair Encoding** (**BPE**). On ne posmatra reči kao nedeljive celine, već ih razbija na manje podjedinice (**subword**-e), koje su zajedničke mnogim rečima. Ovaj proces se zasniva na principu učestalosti: najpre se tekst tokenizuje na pojedinačne znakove, a zatim se u više iteracija spajaju parovi znakova koji se najčešće pojavljuju zajedno, dok se ne formira stabilan skup podrečica. Tako nastaje vokabular koji sadrži kombinacije koje se javljaju u velikom broju reči i predstavljaju osnovne gradivne jedinice jezika.



Slika 3.1. RoBERTa tokenizator

Na primer, reč “unbelievable” može biti podeljena na tokene “un”, “believ” i “able”. Ovaj način segmentacije omogućava modelu da prepozna značenje i u slučaju da se u tekstu pojavi nova reč, poput “believability”, jer već poznaje njene delove. Umesto da tretira svaku reč kao zasebnu celinu, model uči značenje manjih jezičkih jedinica i kombinuje ih u različitim kontekstima. To mu omogućava da razume morfološke sličnosti između reči sa zajedničkim korenom, kao što su “beautiful”, “beauty” i “beautifully”, gde svi oblici dele podjedinicu “beaut”.

Na ovaj način, BPE pristup omogućava efikasnije učenje i generalizaciju. Model postaje sposoban da razume značenje novih reči koje se nisu nalazile u trening skupu, jer ih može "sastaviti" od poznatih delova. Takođe, veličina vokabulara ostaje kontrolisana – umesto da sadrži milione celih reči, on obuhvata samo nekoliko desetina hiljada najfrekventnijih subword jedinica. To direktno doprinosi bržem učenju, manjoj potrošnji memorije i boljoj opštoj sposobnosti razumevanja jezika.

RoBERTa tokenizator automatski:

- dodaje specijalne tokene početka i kraja sekvence (<s>, </s>),
- generiše attention_mask (1 za stvarne tokene, 0 za padding),
- ne koristi token_type_ids (segment ID-eve), za razliku od BERT-a, što pojednostavljuje ulazni skup podataka.

Nakon definisanja tokenizatora, implementirana je funkcija koja se koristi za obradu celog skupa podataka:

```
def tokenize_function(batch):
    return tokenizer(
        batch["clean_text"],
        padding="max_length",
        truncation=True,
        max_length=128
    )
```

Primenom parametara padding="max_length" i truncation=True obezbeđeno je da sve sekvence budu jednake dužine — kraće su dopunjene do maksimalnog broja tokena (128), dok su duže sekvence skraćene na istu granicu. Ova uniformnost omogućila je efikasno procesiranje podataka u batch režimu tokom treniranja modela. Pored toga, tokenizator automatski dodaje specijalne tokene koji označavaju početak i kraj sekvence, kao i attention masku koja modelu omogućava da razlikuje stvarne delove teksta od dopunjenih (padding) tokena. Zatim su svi tekstualni skupovi konvertovani u odgovarajući format pomoću biblioteke datasets:

```
train_dataset = Dataset.from_pandas(train_df[['clean_text', 'label']])
val_dataset = Dataset.from_pandas(val_df[['clean_text', 'label']])
test_dataset = Dataset.from_pandaas(test_df[['clean_text', 'label']])
```

Nakon toga, tokenizacija je primenjena nad svakim skupom:

```
train_dataset = train_dataset.map(tokenize_function, batched=True)
val_dataset = val_dataset.map(tokenize_function, batched=True)
test_dataset = test_dataset.map(tokenize_function, batched=True)
```

Nakon tokenizacije, podaci su sačuvani na disk, što omogućava njihovo kasnije učitavanje bez potrebe za ponovnim procesiranjem — značajno ubrzavajući eksperimentisanje sa različitim modelima.

4. Treniranje

Proces treniranja modela predstavlja ključnu fazu u okviru ovog istraživanja, jer se upravo u tom koraku meri sposobnost pretreniranih jezičkih modela da se prilagode konkretnom zadatku analize sentimenta. Nakon što su podaci očišćeni, tokenizovani i konvertovani u numerički format pogodan za ulaz u neuronsku mrežu, sprovedeno je fino podešavanje (fine-tuning) odabranih RoBERTa modela na skupu Amazon korisničkih recenzija. Cilj treniranja bio je da se pretrenirani modeli, koji su prethodno učeni na ogromnim količinama opšteg tekstualnog materijala, dodatno optimizuju za prepoznavanje emocionalnog tona recenzija – pozitivnog ili negativnog.

4.1. Izbor modela

U okviru ovog rada izabrane su tri varijante RoBERTa arhitekture, koje se međusobno razlikuju po stepenu specijalizacije, veličini i domenu pretreniranja. Cilj ovakvog pristupa bio je da se omogući objektivno poređenje performansi modela različitih kapaciteta i namena, pod jednakim eksperimentalnim uslovima.

Svi modeli su preuzeti sa platforme Hugging Face, koja omogućava direktan pristup pretreniranom jezgru modela i njegovom tokenizeru putem funkcije *from_pretrained().* Tokom inicijalizacije modela, postavljen je parametar *num_labels=2*, jer se u okviru ovog istraživanja radi o binarnoj klasifikaciji sentimenta (pozitivno i negativno), dok je argument *ignore_mismatched_sizes=True* omogućio prilagođavanje modela različitim izlaznim dimenzijama u odnosu na originalni broj klasa u pretreniranoj verziji.

```
from transformers import AutoModelForSequenceClassification

model_name = "cardiffnlp/twitter-roberta-base-sentiment"

model = AutoModelForSequenceClassification.from_pretrained(
    model_name,
    num_labels=2,
    ignore_mismatched_sizes=True
)
```

4.1.1. distilroberta-base

Model DistilRoBERTa-Base predstavlja kompaktniju i optimizovanu verziju originalnog RoBERTa modela, razvijenu primenom tehnike model distilacije (engl. knowledge distillation). Ova tehnika podrazumeva da se veći, kompleksniji „učitelj“ model (teacher) koristi za prenos znanja na manji i efikasniji „učenik“ model (student), koji zadržava većinu performansi uz znatno smanjenu veličinu i računarsku složenost. Konkretno, DistilRoBERTa ima šest slojeva u odnosu na dvanaest kod RoBERTa-Base, što mu omogućava oko 40% manje parametara (≈82 miliona), dok pritom ostvaruje oko 95% tačnosti originalnog modela na zadacima razumevanja jezika.

DistilRoBERTa koristi istu osnovnu arhitekturu i tokenizator kao RoBERTa, zadržavajući prednosti self-attention mehanizma i bidirekcionog konteksta, ali uz značajno manju memorijsku potrošnju i kraće vreme treniranja. Ovakav pristup čini ga naročito pogodnim za situacije gde su resursi ograničeni, ali je i dalje potrebna visoka preciznost modela. U ovom istraživanju, DistilRoBERTa-Base je odabran kao referentni (baseline) model, jer omogućava pouzdano merenje performansi efikasne, ali neneuro-specijalizovane verzije RoBERTa arhitekture. Na taj način, njegova upotreba pruža osnovu za poređenje sa složenijim i domen-specifičnim modelima u zadatku klasifikacije sentimenta Amazon recenzija.

DistilRoBERTa model je javno dostupan putem biblioteke Hugging Face Transformers, a razvijen je od strane tima Hugging Face kao deo istraživanja o destilaciji velikih jezičkih modela zasnovanih na Transformer arhitekturi.

4.1.2. azizbarank/distilroberta-base-sst2-distilled

Model DistilRoBERTa-Base SST-2 Distilled predstavlja unapređenu verziju osnovnog DistilRoBERTa modela, specijalno fino podešenu (fine-tuned) za zadatak analize sentimenta. Dok je DistilRoBERTa-Base univerzalni jezički model, varijanta SST-2 Distilled dodatno je trenirana na poznatom korpusu Stanford Sentiment Treebank 2 (SST-2), koji sadrži hiljade recenzija filmova klasifikovanih kao pozitivne ili negativne. Ovim dodatnim korakom, model ne samo da razume strukturu engleskog jezika, već i uči specifične obrasce emocionalne polarizacije — reči, izraze i sintagme koje signaliziraju stav autora.

Prednost ovakvog pristupa je u tome što model već poseduje „predznanje“ o semantičkim nijansama sentimenta. Time se ubrzava proces konvergencije prilikom dodatnog treniranja na novom skupu podataka, kao što su Amazon recenzije, jer model polazi od već formirane sposobnosti prepoznavanja pozitivnih i negativnih tonova. U praksi, to znači da je model stabilniji, brže postiže dobre rezultate i često dostiže višu tačnost u odnosu na generičke modele, posebno u binarnoj klasifikaciji.

DistilRoBERTa-SST2-Distilled zadržava sve arhitektonske karakteristike osnovnog DistilRoBERTa (šest slojeva, 82 miliona parametara), ali njegova klasifikaciona glava već ima parametre optimizovane za sentiment zadatke. U ovom radu korišćen je kao specijalizovani model, čiji cilj je da pokaže koliko dodatna fine-tuning faza na srodnom domenu (recenzije filmova) može doprineti performansama na drugačijem, ali tematski bliskom korpusu (recenzije proizvoda). Model je javno dostupan preko Hugging Face platforme.

4.1.3. cardiffnlp/twitter-roberta-base-sentiment

Model CardiffNLP Twitter-RoBERTa-Base Sentiment predstavlja domenski specijalizovanu verziju RoBERTa arhitekture, razvijenu od strane Cardiff NLP grupe sa ciljem da omogući analizu sentimenta na tekstovima preuzetim sa društvene mreže Twitter. Ovaj model je treniran na ogromnom skupu od 58 miliona tvitova, koristeći self-supervised pristup, a zatim fino podešen (fine-tuned) za klasifikaciju sentimenta na TweetEval benchmarku — jednom od najpoznatijih standarda za evaluaciju NLP modela u domenu društvenih mreža.

Za razliku od generičkih RoBERTa modela, koji su obučeni na formalnijim tekstovima poput vesti i članaka, Twitter-RoBERTa je osmišljen da razume neformalne, kratke i emocionalno bogate izraze tipične za onlajn komunikaciju. Tokom treninga, model je naučio

da prepoznaje semantičke obrasce koji uključuju sleng, skraćenice, emotikone i sarkastične konstrukcije, čime je posebno prilagođen zadacima gde je kontekst izražen implicitno. Zbog toga se ističe u zadacima koji zahtevaju razumevanje „tonu poruke“ čak i kada se sentiment ne iskazuje eksplicitno.

U okviru ovog rada, model cardiffnlp/twitter-roberta-base-sentiment korišćen je kao predstavnik domen-specifičnih modela, s ciljem da se ispita koliko se znanje stečeno na neformalnim tekstovima društvenih mreža može uspešno preneti na formalniji domen — recenzije proizvoda na Amazonu. Iako se pretpostavlja da razlika u tipu jezika može uticati na performanse, ovaj model je koristan jer pokazuje sposobnost transfera znanja između domena, što je jedno od ključnih istraživačkih pitanja u savremenom NLP-u.

4.2. Podešavanje i proces treniranja modela

Nakon definisanja i inicijalizacije modela, sproveden je proces fine-tuninga (dodatnog treniranja) pretreniranih RoBERTa arhitektura na pripremljenom Amazon Reviews datasetu. Trening je realizovan korišćenjem biblioteke Hugging Face Transformers, konkretno klase Trainer, koja obezbeđuje potpunu kontrolu nad procesom obuke i evaluacije modela, uz minimalnu količinu koda.

Za kontrolu hiperparametara i podešavanje eksperimentalnih uslova korišćena je klasa **TrainingArguments**, kojom su precizno definisani svi parametri treniranja. U okviru istraživanja korišćene su sledeće postavke:

```
training_args = TrainingArguments(
    output_dir="./twitter-roberta-base-sentiment",
    eval_strategy="epoch",
    save_strategy="epoch",
    learning_rate=2e-5,
    per_device_train_batch_size=128,
    per_device_eval_batch_size=128,
    num_train_epochs=5,
    weight_decay=0.01,
    load_best_model_at_end=True,
    logging_dir="./logs3",
    logging_steps=100
)
```

Ovim podešavanjima određeno je da se evaluacija i čuvanje modela obavljaju na kraju svake epohe ("epoch"), čime se omogućava kontinuirano praćenje performansi i izbor najbolje verzije modela na osnovu rezultata sa validacionog skupa.

Brzina učenja (learning_rate=2e-5) postavljena je na relativno nisku vrednost kako bi se osiguralo stabilno konvergiranje modela bez naglih oscilacija u gubitku (loss funkciji). Batch veličina od 128 omogućila je efikasno iskorišćenje grafičke memorije, dok je broj epoha (5) izabran kao optimalan balans između trajanja treniranja i rizika od prenaučenosti (overfittinga). Parametar weight_decay=0.01 uveden je radi regularizacije modela, sprečavajući preveliko prilagođavanje trenažnim podacima.

Konačna konfiguracija i pokretanje procesa treniranja realizovani su pomoću **Trainer** klase:

```
trainer = Trainer(
    model=model,
    args=training_args,
    train_dataset=train_dataset,
    eval_dataset=val_dataset,
    tokenizer=tokenizer,
    compute_metrics=compute_metrics
)

trainer.train()
```

Klasa Trainer automatizuje proces obuke tako što interno upravlja forward i backward propagacijom, izračunavanjem gradijenata, ažuriranjem težina, evaluacijom modela i logovanjem metrika. Funkcija compute_metrics definisana je tako da meri osnovne pokazatelje performansi: tačnost (accuracy), preciznost (precision), odziv (recall) i F1-score, što omogućava sveobuhvatnu procenu kvaliteta modela.

Trening svih modela obavljen je na Data Crunch serveru, koji poseduje četiri NVIDIA Tesla GPU jedinice, svaka sa 16 GB grafičke memorije. Zahvaljujući paralelizaciji kroz CUDA i PyTorch DistributedDataParallel, proces treniranja je značajno ubrzan, omogućavajući istovremenu obradu više batch-ova i optimalno iskorišćenje svih GPU resursa. U poređenju sa lokalnim okruženjima, treniranje na ovakvom sistemu omogućava bržu konvergenciju, stabilnije performanse i lakše upravljanje većim batch veličinama.

U toku eksperimenta praćeni su logovi treniranja u realnom vremenu pomoću TensorBoard-a (putem direktorijuma ./logs), čime je omogućeno vizuelno praćenje promena vrednosti funkcije gubitka i metrika kroz epohe. Nakon završetka procesa, model sa najboljim validacionim rezultatima automatski je sačuvan pomoću opcije load_best_model_at_end=True.

5. Evaluacija modela i analiza rezultata

Nakon uspešno završenog procesa treniranja RoBERTa modela za sentiment analizu, sprovedena je sveobuhvatna evaluacija sa ciljem procene efikasnosti, stabilnosti i verodostojnosti modelskih predikcija.

Analiza performansi obuhvatila je primenu standardnih metrika za klasifikaciju, kao što su accuracy, precision, recall i F1-score, kao i dodatne tehnike vizuelne i analitičke provere poput konfuzionih matrica, ROC i Precision-Recall krivih, distribucije predikcija po klasama i analize stvarnih grešaka modela. Ovakav pristup omogućava ne samo numeričku procenu uspešnosti, već i dublje razumevanje ponašanja modela u realnim uslovima, posebno u slučajevima dvosmislenih i kompleksnih tekstualnih primera. Na ovaj način sagledani su i

ukupni rezultati svakog modela, identifikovane specifične slabosti, kao i prednosti pristupa baziranih na kompaktnim RoBERTa varijantama za zadatke sentiment analize.

5.1. Metrike evaluacije

Evaluacija modela za sentiment analizu zahteva korišćenje odgovarajućih metrika koje mogu pouzdano da odraze koliko je model uspešan u prepoznavanju emocionalnog tona teksta. U ovom radu primenjene su standardne metrike klasifikacije: accuracy, precision, recall i F1-score, koje u kombinaciji pružaju celovitu sliku kvaliteta modela.

Accuracy predstavlja osnovnu metriku koja meri udeo ispravno predviđenih primera u odnosu na ukupan broj primera u test skupu.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Gde su :

- TP (True positive): tačno pozitivne vrednosti,
- TN (True Negatives) – tačno negativne vrednosti,
- FP (False Positives) – lažno pozitivne vrednosti,
- FN (False Negatives) – lažno negativne vrednosti.

Prednost ove metrike je njena intuitivnost i jednostavno tumačenje. Međutim, ona može biti varljiva u slučaju neravnoteže klasa (class imbalance), jer model može delovati „tačan" i kada favorizuje dominantnu klasu. Zbog toga se accuracy koristi u kombinaciji sa ostalim metrikama.

Precision meri kvalitet pozitivnih predikcija modela — tj. koliko primera označenih od strane modela kao pozitivnih zaista pripada pozitivnoj klasi.

$$Precision = \frac{TP}{TP + FP}$$

Visoka preciznost znači da model retko daje lažno pozitivne rezultate. Ova metrika je posebno važna u sistemima gde je pogrešno označavanje negativnih primera kao pozitivnih skupo (npr. spam detekcija).

Recall meri koliko je model uspešan u pronalaženju svih stvarno pozitivnih primera.

$$Recall = \frac{TP}{TP + FN}$$

Visok recall ukazuje da model retko propušta pozitivne primere, što je važno kod zadataka gde je ključno detektovati sve relevantne slučajeve (npr. medicinska dijagnostika). U kontekstu sentiment analize, visoki recall znači da model uspešno prepoznaje većinu izraženih emocija.

F1-score predstavlja harmonijsku sredinu između precision-a i recall-a i koristi se kada je važno pronaći balans između ova dva aspekta performansi.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

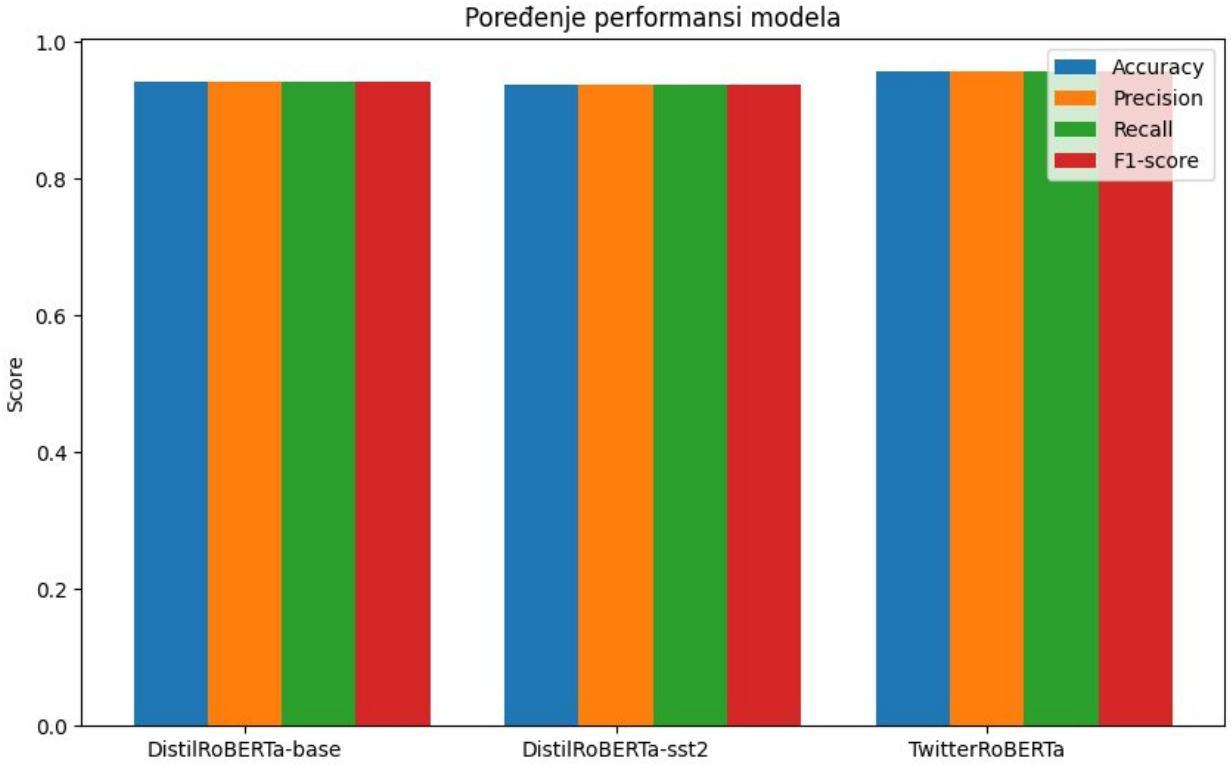
Ova metrika je robusna u prisustvu neravnoteže klasa i predstavlja stabilan pokazatelj ukupnog kvaliteta modela. Što je F1 vrednost veća, to je model bolje izbalansiran u prepoznavanju pozitivnih primera bez generisanja prevelikog broja grešaka.

5.2. Rezultati evaluacije modela

Upoređene su performanse tri RoBERTa modela optimizovana za sentiment analizu. U nastavku je prikazana tabela sa postignutim rezultatima.

Model	Accuracy	Precision	Recall	F1-score
DistilRoBERTa Base	0.942	0.9421	0.9420	0.9420
DistilRoBERTa SST2	0.936	0.9360	0.9360	0.9360
TwitterRoBERTa	**0.956**	**0.9560**	**0.9560**	**0.9560**

Pored tabele, na grafiku u nastavku prikazano je vizuelno poređenje performansi modela, što omogućava intuitivan uvid u njihove razlike (Slika 5.2.1.).



Slika 5.2.1. Poređenje performansi modela

Rezultati pokazuju da su svi modeli ostvarili veoma visoke performanse, što ukazuje na to da su RoBERTa arhitekture izuzetno efikasne u zadatku sentiment analize. Međutim, postoje jasne razlike u njihovoj uspešnosti:

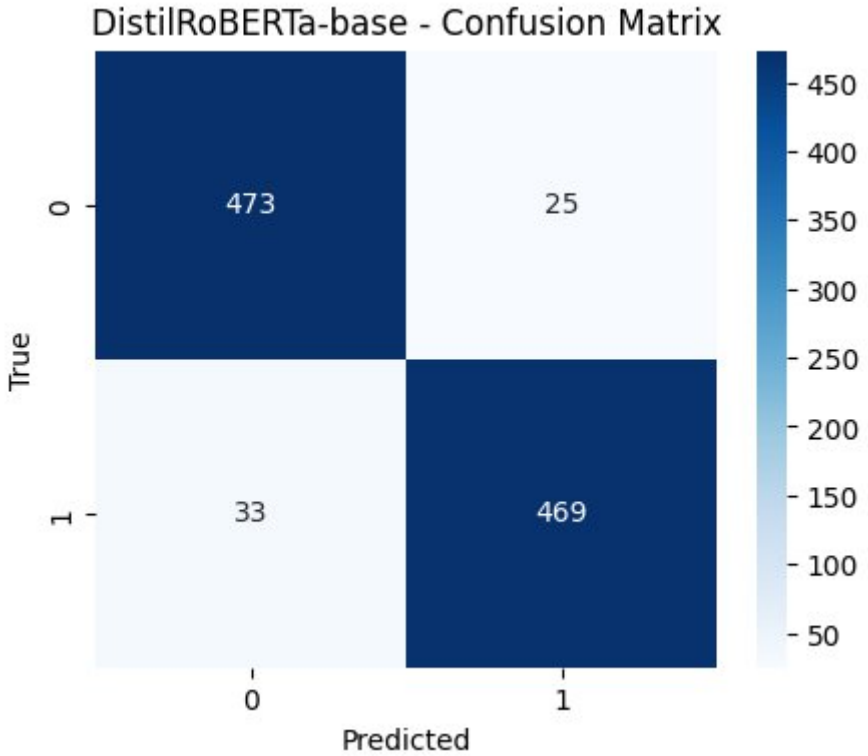
TwitterRoBERTa model pokazao je najbolje rezultate u svim metrikama, sa ukupnom tačnošću od **95.6**%, čime je nadmašio ostale modele. Očekivano, s obzirom na to da je treniran na velikom broju kratkih, kolokvijalnih tekstova sa društvenih mreža, što ga čini posebno pogodnim za analizu neformalnog izraza i brzih emocionalnih reakcija.

DistilRoBERTa Base ostvario je vrlo stabilne rezultate (94.2% accuracy) i predstavlja odličan kompromis između brzine i performansi. Kao distilovana verzija standardnog RoBERTa modela, nudi značajno manju potrošnju resursa i brže izvođenje.

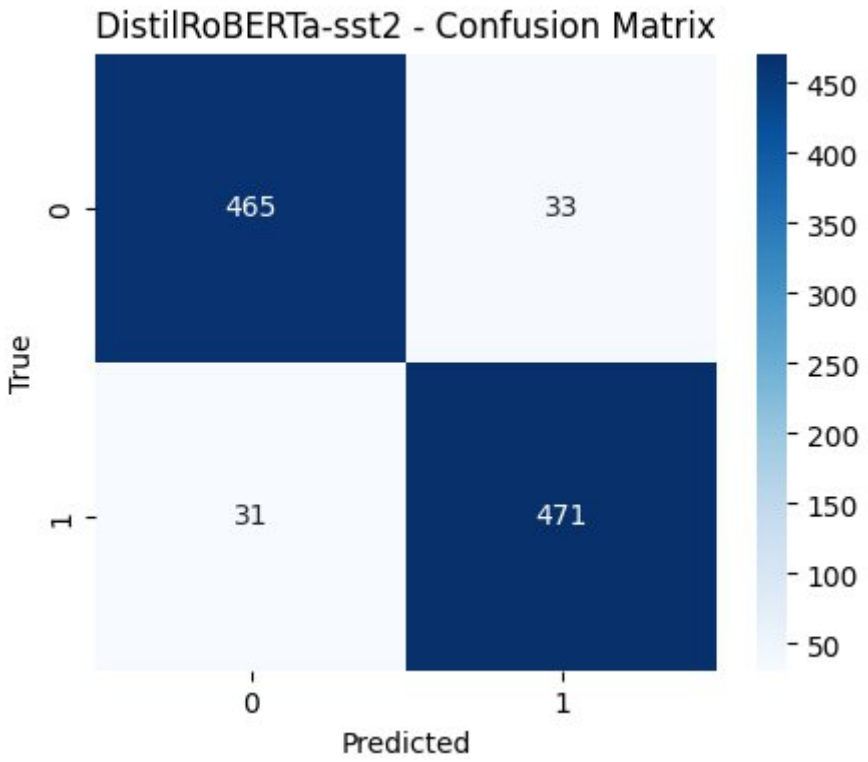
DistilRoBERTa SST2, iako ostvaruje najniži rezultat među testiranim modelima (93.6% accuracy), i dalje pruža pouzdane performanse. Razlog za nešto slabiji učinak leži u domenskom nepoklapanju — model je treniran na formalnijim tekstovima (filmske recenzije), pa mu je potrebno više prilagođavanja na neformalan jezik.

5.3. Konfuziona matrica

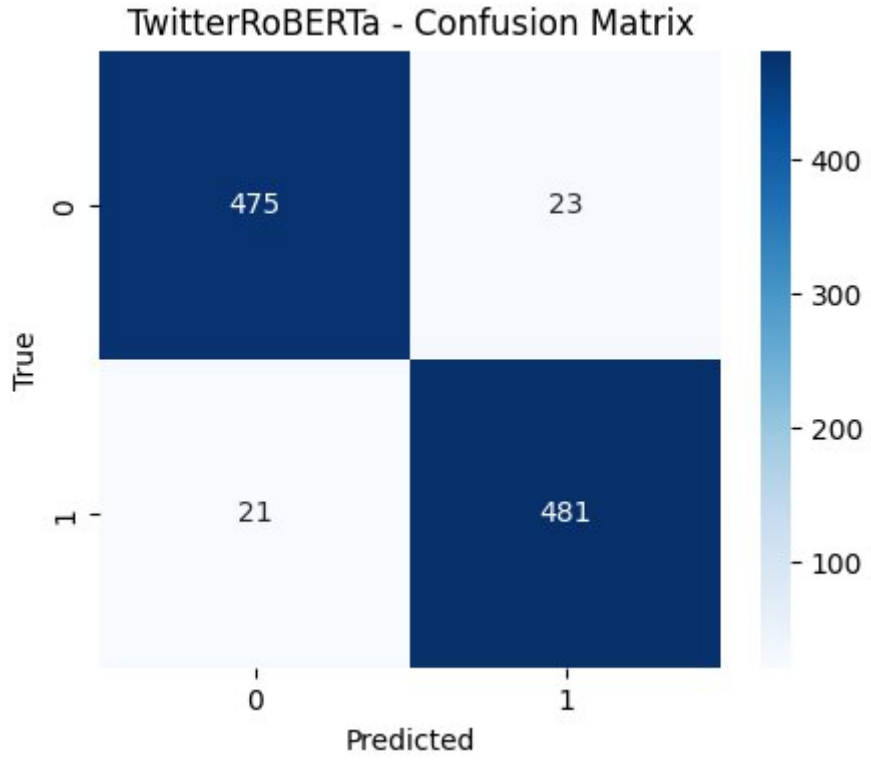
Pored osnovnih numeričkih metrika, važan deo evaluacije modela predstavlja analiza konfuzione matrice. Konfuziona matrica daje detaljan prikaz odnosa između stvarnih (istinitih) i predikovanih klasa, omogućavajući identifikovanje tipičnih grešaka modela i njihovog ponašanja pri klasifikaciji različitih kategorija sentimenta.



Ovaj model distilroberta-base pokazuje stabilnije rezultate u odnosu na SST2 varijantu, sa manjim brojem pogrešnih pozitivnih predikcija (25 FP). Istovremeno, blago povećanje lažno negativnih primera (33 FN) ukazuje da model u pojedinim slučajevima ne prepoznaje pozitivno obojen sentiment, naročito kod manje izraženih emocija. U celini, model postiže odlične rezultate uz očuvan balans performansi i efikasnosti.



Model distilroberta-base-sst2-distilled pokazuje uravnoteženu sposobnost detekcije i pozitivnih i negativnih primera, međutim beleži nešto veći broj grešaka u odnosu na ostale modele. Primećuje se da model nešto češće pogrešno predviđa pozitivnu klasu (33 FP) u odnosu na negativnu (31 FN). To ukazuje da ponekad interpretira neutralan ili slabije negativan ton kao pozitivan. To je očekivano imajući u vidu da je treniran na formalnijem tekstualnom domenu (SST-2), pa slabije generalizuje na kraći i neformalniji tekst.



Model twitter-roberta-base-sentiment postiže najbolje rezultate od sva tri modela. Ima najmanji broj pogrešnih klasifikacija, što potvrđuje prethodne kvantitativne metrike. Odličan odnos FP i FN grešaka ukazuje da model vrlo precizno razlikuje i pozitivan i negativan sentiment. Najmanji broj FN vrednosti (21) jasno pokazuje superiornost u prepoznavanju pozitivnog sentimenta — što je naročito važno kod tekstova iz prakse gde su impulsi i emocionalni izrazi često kratki i intenzivni. Ovaj rezultat je logičan s obzirom da je model treniran na velikoj količini društvenomrežnog sadržaja, koji je stilski vrlo blizak podacima sa test skupa.

5.4. ROC i PR krive

Za dodatnu evaluaciju performansi modela korišćene su ROC (Receiver Operating Characteristic) krive i PR (Precision–Recall) krive. Ove vizuelne metode omogućavaju procenu ponašanja modela pri različitim pragovima klasifikacije, što daje širi uvid u njihov kvalitet u odnosu na samo jedinstveni prag (npr. 0.5).

ROC kriva prikazuje odnos između:

- True Positive Rate (TPR) – osetljivost modela (Recall)
- False Positive Rate (FPR) – udeo negativnih primera pogrešno označenih kao pozitivni

$$TPR = \frac{TP}{TP + FN}, \qquad FPR = \frac{FP}{FP + TN}$$

Idealni model teži uglu (0,1) — visoka osetljivost, niska stopa lažnih alarma. Površina ispod ROC krive (AUC – Area Under Curve) je indikator globalne sposobnosti modela da razdvoji klase:

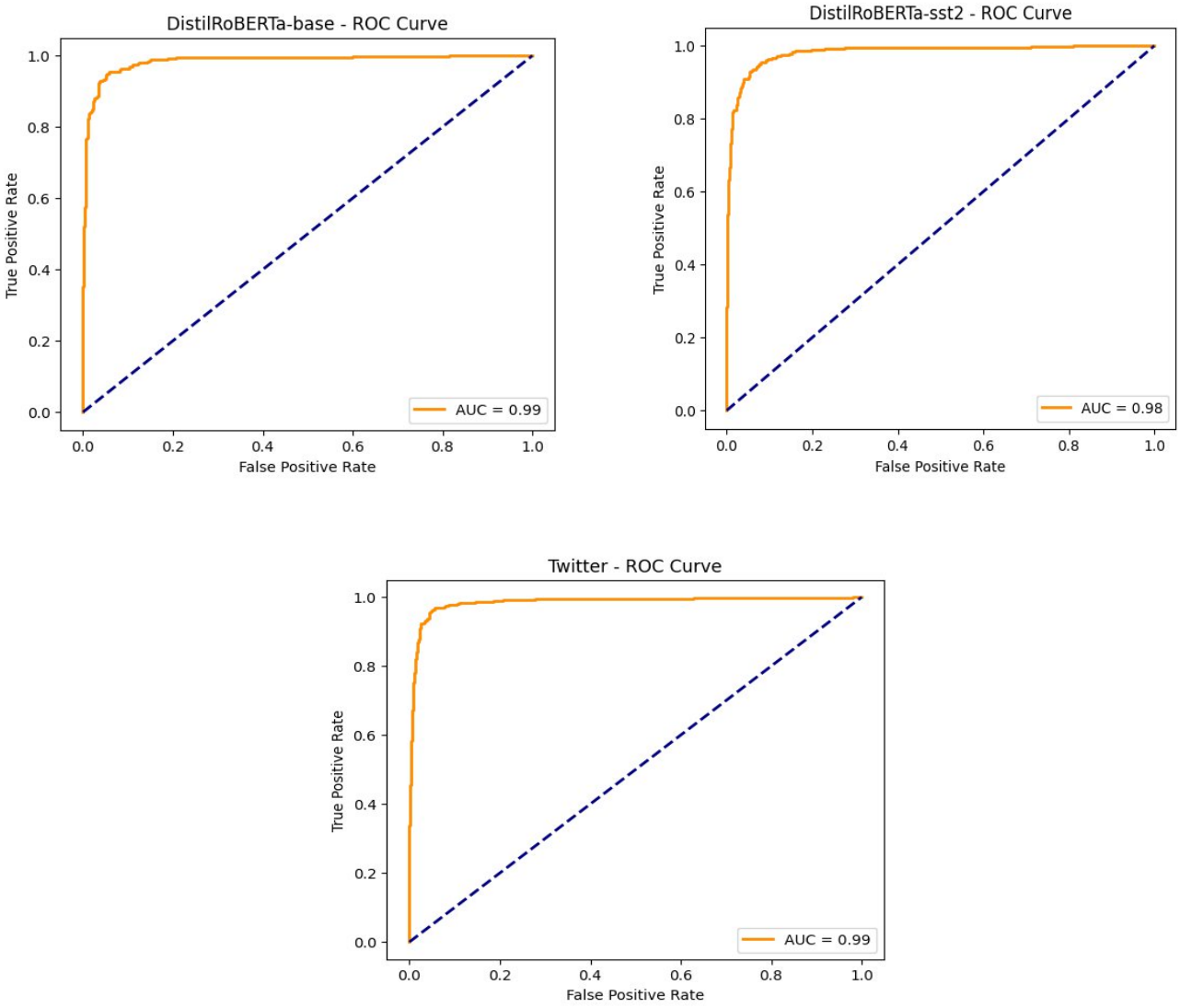
- AUC = 1.0 → savršen model
- AUC = 0.5 → slučajno pogađanje (npr. bacanje novčića)

PR kriva prikazuje odnos između Precision (preciznost) i Recall (odziv)

$$Precision = \frac{TP}{TP + FP}, \qquad Recall = \frac{TP}{TP + FN}$$

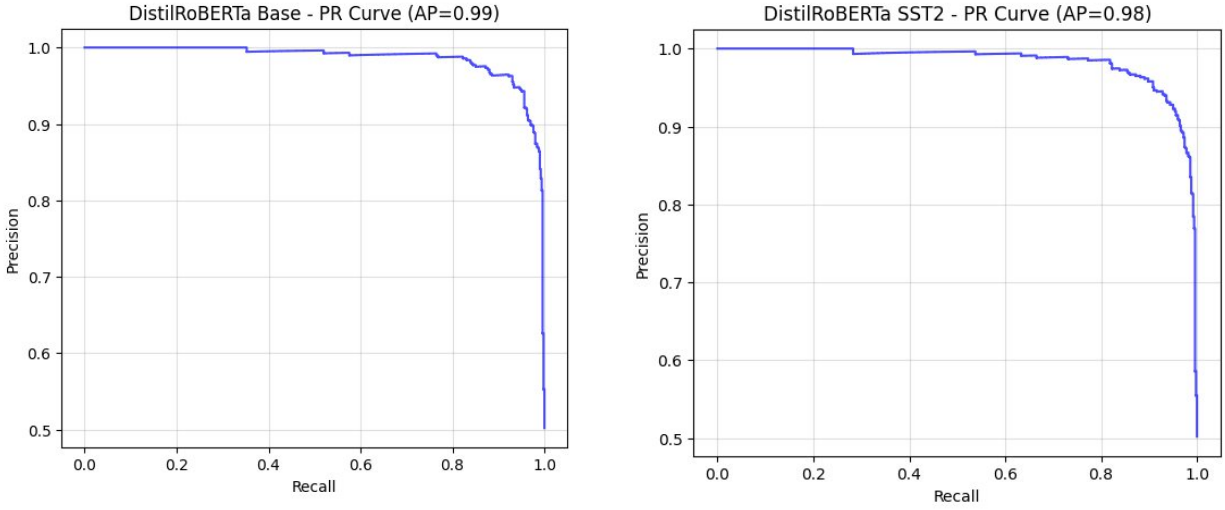
Ova kriva je posebno važna kod neravnoteže klasa, jer ROC može izgledati dobro čak i kada jedna klasa dominira. U sentiment analizi — gde balans obično postoji — PR kriva služi kao dodatna potvrda stabilnosti modela. Površina ispod PR krive označava Average Precision (AP). Što je AP bliži 1, model bolje balansira preciznost i odziv.

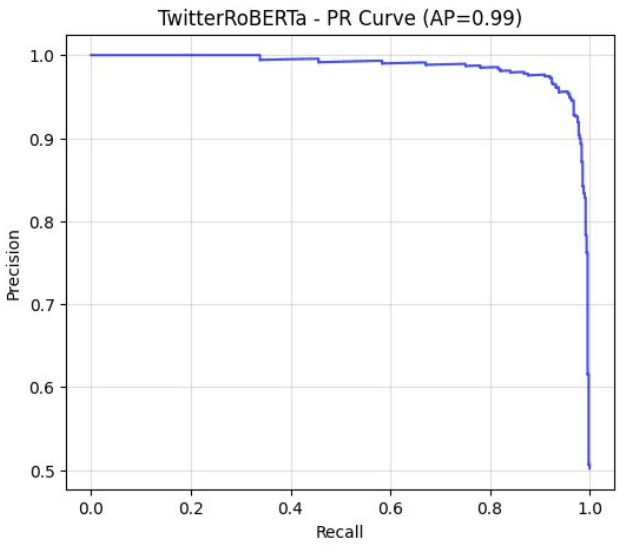
Sve ROC krive (Slika 5.4.1.) pokazuju izrazitu zakrivljenost ka gornjem levom uglu, što znači da modeli veoma dobro razdvajaju pozitivne i negativne primere.TwitterRoBERTa i DistilRoBERTa Base postižu gotovo idealan rezultat (0.99), dok DistilRoBERTa SST2 za nijansu zaostaje, ali i dalje ostvaruje odličan učinak (0.98).



Slika 5.4.1. ROC krive modela

PR krive (Slika 5.4.2.) potvrđuju stabilne performanse — kod sva tri modela, preciznost ostaje izuzetno visoka čak i pri većim vrednostima odziva. TwitterRoBERTa i DistilRoBERTa Base ponovo dominiraju, dok SST2 pokazuje nešto brži pad preciznosti pri maksimalnom recall-u.





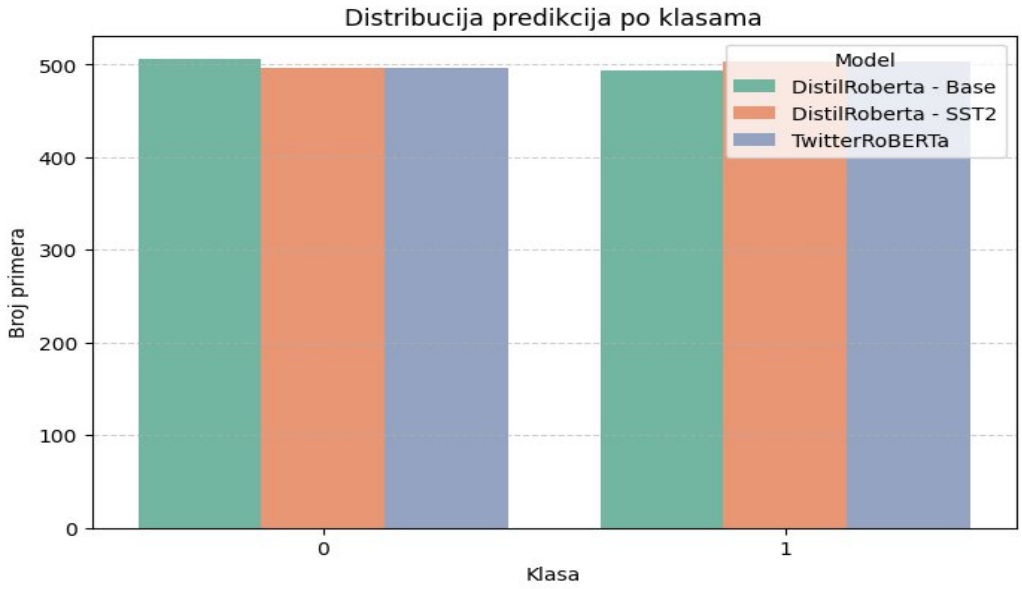
Slika 5.4.2. PR krive modela

5.5. Distribucija predikcije po klasama

Pored numeričkih metrika i vizuelizacija kao što su ROC i PR krive, važan deo evaluacije predstavlja analiza distribucije modelskih predikcija po klasama. Cilj ove analize je da se proveri da li modeli pokazuju sklonost ka favorizovanju jedne klase (bias), odnosno da li postoji disbalans u predviđanjima koji bi mogao negativno uticati na realnu primenu modela.

Distribucija predikcija predstavlja ukupan broj primera koje je model svrstao u svaku klasu (pozitivnu i negativnu), bez obzira na njihovu stvarnu vrednost. Idealno, distribucija predikcija treba da reflektuje realnu distribuciju u test skupu — u ovom slučaju relativno ravnomernu zastupljenost pozitivnih i negativnih primera.

Rezultati analize (Slika 5.5.1.) pokazuju da sva tri modela održavaju uravnoteženu distribuciju predikcija, bez izražene pristrasnosti ka određenoj klasi. Ovo je posebno važno u sentiment analizi, jer modeli sa nepravilnom distribucijom mogu davati iskrivljene zaključke, npr. preterano pozitivan ili negativan ton.



Slika 5.5.1. Distribucija predikcije po klasama

5.6. Analiza grešaka i primeri pogrešnih predikcija

Da bi se detaljnije razumele vrste grešaka koje modeli prave, analizirani su konkretni primeri iz test skupa gde su predikcije bile netačne. U tabelama su prikazani uzorci teksta, njegova stvarna oznaka i predikovana klasa, kao i dužina teksta. Primeri su izabrani tako da ilustruju tipične izazove pri sentiment analizi.

Model ponekad pogrešno interpretira neutralno-kritičke naučne i politicke recenzije kao pozitivne. Dugački tekstovi sa kompleksnim rečenicama i formalnim stilom predstavljaju izazov. Pozitivne preporuke sa suptilnim tonom („recommend ordering…“) ponekad se pogrešno tumače kao negativne zbog praktičnih zamerki u tekstu.

	index	true	pred	text	length
13	233	1	0	typical thomas video james biggest problem dvd...	569
44	753	1	0	title revised say advancement homeland securit...	562
15	250	1	0	effective book discussing physical death child...	521
1	71	1	0	recommend ordering size larger regular size or...	518
27	401	0	1	good politics science book starts enticing psy...	485
52	901	0	1	may modernday classic horror film based origin...	468
37	554	1	0	statin levels based researchedread pill formul...	461
35	546	1	0	discontinued another reviewer recently advised...	434
54	947	0	1	coyote genius usual girl goes big city cliches...	402
47	846	1	0	nietzsche antichrist nietzsche fed christianit...	381

Često pozitivno označava informativne, deskriptivne i istorijske sadržaje, iako sentiment nije pozitivan. Model distilroberta-base-sst2-distilled je treniran na filmskim recenzijama, pa se vidi da mu nedostaje robusnost za širi domen. Bori se sa tekstovima bez direktno emotivnog izražavanja.

	index	true	pred	text	length
38	576	0	1	informative idiosyncratic coverage historical ...	700
13	233	1	0	typical thomas video james biggest problem dvd...	569
16	250	1	0	effective book discussing physical death child...	521
2	71	1	0	recommend ordering size larger regular size or...	518
21	340	0	1	transformers dark moon accounts michael baydir...	502
28	401	0	1	good politics science book starts enticing psy...	485
56	901	0	1	may modernday classic horror film based origin...	468
37	554	1	0	statin levels based researchedread pill formul...	461
35	546	1	0	discontinued another reviewer recently advised...	434
60	947	0	1	coyote genius usual girl goes big city cliches...	402

Iako najuspešniji, model twitter-roberta-base-sentiment i dalje greši pre svega u slučajevima: mešovitih tonova gde se istovremeno javljaju kritičke i pohvalne ocene, stilistički složenih, književno-kritičkih tekstova sa metaforama i implicitnim vrednosnim sudovima. Referenci i preporuka iz filmsko-književnog konteksta, gde tačna klasifikacija zavisi od šireg poznavanja dela ili žanra.

	index	true	pred	text	length
10	233	1	0	typical thomas video james biggest problem dvd...	569
11	250	1	0	effective book discussing physical death child...	521
21	401	0	1	good politics science book starts enticing psy...	485
37	901	0	1	may modernday classic horror film based origin...	468
27	554	1	0	statin levels based researchedread pill formul...	461
25	546	1	0	discontinued another reviewer recently advised...	434
41	947	0	1	coyote genius usual girl goes big city cliches...	402
33	840	0	1	unusual writing technique joyce quite understa...	381
38	912	1	0	make sure see jean de florette first second pa...	358
16	362	1	0	good idea good looks needs lots work recently ...	348

Najveći izazovi su dugi tekstovi, kompleksna sintaksa, implicitne emocije i dvosmislene ocene, dok kratki direktni komentari modelima predstavljaju najmanji problem.

6. Zaključak

U ovom radu sprovedeno je istraživanje performansi različitih RoBERTa modela primenjenih na zadatak sentiment analize teksta. Proces je obuhvatio pripremu podataka, fino podešavanje modela (fine-tuning), evaluaciju koristeći standardne metrike (accuracy, precision, recall, F1-score), kao i naprednije vizuelne tehnike poput ROC i PR krivih. Dodatno, analiza konfuzionih matrica i pojedinačnih pogrešnih predikcija omogućila je detaljnije razumevanje ponašanja modela i tipova podataka kod kojih dolazi do grešaka.

Eksperimenti su realizovani na Amazon Reviews datasetu za binarnu klasifikaciju sentimenta (pozitivan / negativan). Rezultati jasno ukazuju da svi analizirani modeli ostvaruju vrlo visoke performanse, sa tačnošću iznad 93%. Najbolje rezultate postigao je cardiffnlp/twitter-roberta-base, što je očekivano s obzirom na njegovu obuku na neformalnim i kraćim tekstovima, koji su stilski slični korisničkim komentarima u upotrebljenom skupu podataka.

Model distilroberta-base pokazao se kao izuzetno efikasna i resursno štedljiva alternativa, dok distilroberta-base-sst2 za nijansu zaostaje zbog domenskog nesklada između originalnog trening korpusa (SST-2, čiji je sadržaj formalniji) i Amazon recenzija.

Analiza pogrešnih predikcija pokazala je da se greške najčešće javljaju u tekstovima sa suptilnim, ironičnim i dvosmislenim tonom, kao i kod stilistički složenijih recenzija sa kombinovanim pozitivnim i negativnim elementima. Ovo potvrđuje da interpretacija sentimenta često zavisi od konteksta, implicitnih značenja i šireg znanja, što i dalje ostaje izazov za savremene modele.

Zaključno, RoBERTa-bazirani modeli predstavljaju izuzetno efikasno rešenje za sentiment analizu korisničkih recenzija, sa potencijalom za primenu u realnim sistemima preporuke, reputacionim analizama i automatizovanom upravljanju korisničkim komentarima.

7. Literatura

[1] Barbieri, F., Camacho-Collados, J., Espinosa-Anke, L., & Neves, L. (2020). *TweetEval: Unified benchmark and comparative evaluation for tweet-based classification tasks*. Findings of the Association for Computational Linguistics: EMNLP 2020, 1644–1654. https://doi.org/10.18653/v1/2020.findings-emnlp.148

[2] Guo, X., Jiang, Y., & Sun, X. (2024). *Sentiment analysis based on RoBERTa for Amazon review: An empirical study on decision making*. arXiv preprint arXiv:2411.00796. https://arxiv.org/abs/2411.00796

[3] Rahman, M., Islam, S., & Ahmed, F. (2024). *RoBERTa-BiLSTM: A context-aware hybrid model for sentiment analysis*. arXiv preprint arXiv:2406.00367. https://arxiv.org/abs/2406.00367

[4] Joshy, S., & Sundar, R. (2022). *Analyzing the performance of sentiment analysis using BERT, DistilBERT and RoBERTa*. https://www.researchgate.net/publication/369319715

[5] Barbieri, F., & Camacho-Collados, J. (2021). *cardiffnlp/twitter-roberta-base* [Model]. Hugging Face. https://huggingface.co/cardiffnlp/twitter-roberta-base

[6] Hugging Face. (2020). *distilroberta-base* [Model]. Hugging Face. https://huggingface.co/distilroberta-base

[7] Hugging Face. (2020). *distilroberta-base-sst2* [Fine-tuned model]. Hugging Face. https://huggingface.co/azizbarank/distilroberta-base-sst2-distilled

[8] Dataset: Amazon Reviews (Kaggle) https://www.kaggle.□om/datasets/kritanjalijain/amazon-reviews