SVEUČILIŠTE JOSIPA JURJA STROSSMAYERA U OSIJEKU

FAKULTET ELEKTROTEHNIKE, RAČUNARSTVA I

INFORMACIJSKIH TEHNOLOGIJA OSIJEK

KLASIFIKACIJA EMOCIJA

IZ AUDIO-ZAPISA GOVORA

Projektni zadatak

Karla Binder, univ.bacc.ing.comp.

Marko Budak, univ.bacc.ing.comp.

Matko Grgić, univ.bacc.ing.comp.

Leo Tumbas, univ.bacc.ing.comp.

Osijek, 2024.

SADRŽAJ

1. UVOD

U području strojnog učenja i umjetne inteligencije, čovjek neprestano pokušava svoje ponašanje i obrasce istih prenijeti na stroj. Jedan od glavnih načina prenošenja informacija među ljudima je govor. Na računalu je jedinica uvijek jedinica, ali kod čovjeka isti sadržaj ne mora nužno značiti isti kontekst. Naime, način na koji je nešto izrečeno, tj. emocija koja ide uz sam sadržaj govora, izgovorene riječi, može prenijeti potpuno drugu informaciju na sugovornika. Zbog toga je razumijevanje ljudskih emocija ključan aspekt u sporazumijevanju između čovjeka i stroja. Klasifikacija emocija aktualna je tema gdje se pomoću matematičkih funkcija nastoje otkriti složeni obrasci u akustičnim značajkama ljudskog govora. Mogućnost preciznog raspoznavanja emocija može pridonijeti poboljšanju korisničkog iskustva, unaprjeđivanju sustava komunikacije, razvoju empatije unutar računalnog sustava i dr.

Ovaj rad usmjeren je na klasifikaciju ljudskog govora i pronalaženja obrazaca za otkrivanje jedne od četiri emocija kod čovjeka: ljutnja, sreća, tuga i gađenje, te otkrivanje „neutralnog“ ili „neeskpresivnog“ govora – govora bez konteksta emocije. Problemu se pristupilo stvaranjem tri modela, od kojih svaki sadrži jednu od tri popularne arhitekture povratnih neuronskih mreža: RNN (*Recurrent neural network*), LSTM (*Long-short term memory*) i GRU (*Gated recurrent unit*). Izvlačenjem značajki iz zvučnih zapisa korištenjem *Wav2Vec2 feature extractor* mreže, sustav pokušava klasificirati govor prema emociji izraženoj u zvučnom zapisu.

U prvom poglavlju rada govori se o podacima unutar podatkovnih skupova korištenih za treniranje, evaluaciju, i testiranje modela. Nad istima je provedena eksplorativna analiza kako bi se utvrdile značajke podatkovnih skupova i nepravilnosti koje je u predobradi podataka potrebno ispraviti.

Drugo poglavlje sadrži popis korištenih algoritama i tehnologija unutar projekta. Opisani su načini rada *Wav2Vec2 feature extractor* mreže i svakog od korištenih slojeva.

Unutar trećeg poglavlja predstavljamo rezultate treniranja i evaluacije, rezultatima testiranja odabiremo „najbolji“ model među izgrađenima, te uspoređujemo performanse našeg modela s performansama modela izgrađenih u relevantnom znanstvenom radu.

Demonstracija rada, u četvrtom pogljavlju, prikazat će kako izrađeno programsko sučelje omogućuje korisniku klasifikaciju vlastitih zvučnih zapisa prema emocijama.

2. OPIS KORIŠTENIH PODATKOVNIH SKUPOVA

Prvi od dva korištena skupa podataka je LSSED (*Large-Scale Dataset And Benchmark For Speech Emotion Recognition*). LSSED je vrlo opsežan podatkovni skup za prepoznavanje emocija u govoru[1] te je korišten za treniranje i evaluaciju modela. Sadrži 147,025 dvodimenzionalna polja opisnih karakteristika nastalih kao izlazi *Wav2Vec2 feature extractor*[2] mreže. Polja su nastala davanjem zvučnih zapisa rečenica (206 sati i 25 minuta sadržaja) izgovorenih od 820 ljudi *Wav2Vec2 feature extractor* mreži kao ulaz. Svaki primjerak iz podatkovnog skupa označen je jednom od 11 različitih emocija: ljutnja, strah, sreća, tuga, iznevjerenost, dosada, gađenje, uzbuđenost, strah, „neeskpresivni“ govor, te ostalo. Osim oznaka za emocije, svaki primjer sadrži oznaku za spol i dob osobe.

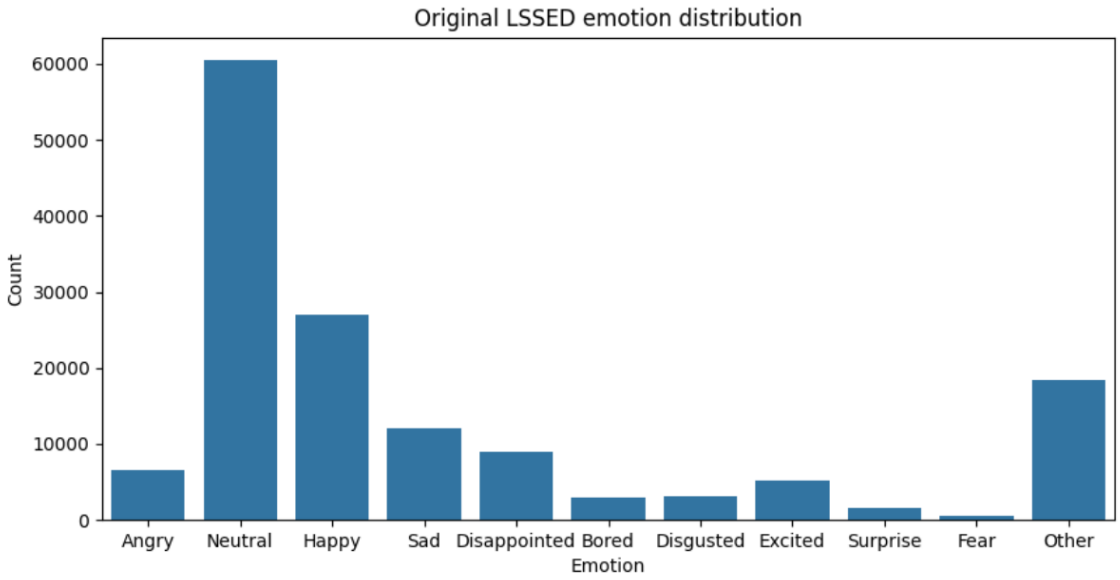
	Vvid	Gender	Age	Emotion
0	P0001__109	Female	Midlife	Angry
1	P0001__112	Female	Midlife	Disappointed
2	P0001__115	Female	Midlife	Disappointed
3	P0001__26	Female	Midlife	Angry
4	P0001__47	Female	Midlife	Disappointed

Slika 2.1. *Prikaz prvih pet ulaza iz LSSED podatkovnog skupa*

Za testiranje modela korišten je SER-EN (*Speech Emotion Recognition (en)*)[3] podatkovni skup koji je kombinacija četiri različita podatkovna skupa: CREMA-D (*Crowd-sourced Emotional Multimodal Actors Dataset*)[4], TESS (*Toronto emotional speech set*)[5], RAVDESS (*The Ryerson Audio-Visual Database of Emotional Speech and Song*)[6], i SAVEE (*Surrey Audio-Visual Expressed Emotion Database*)[7]. Iz sva 4 podatkovna skupa izdvojeni su isključivo audio-zapisi ljudskog govora, svi ostali primjerci nisu dodani u SER-EN podatkovni skup. Oznake emocija sadržane su u samim imenima datoteka.

2.1. LSSED

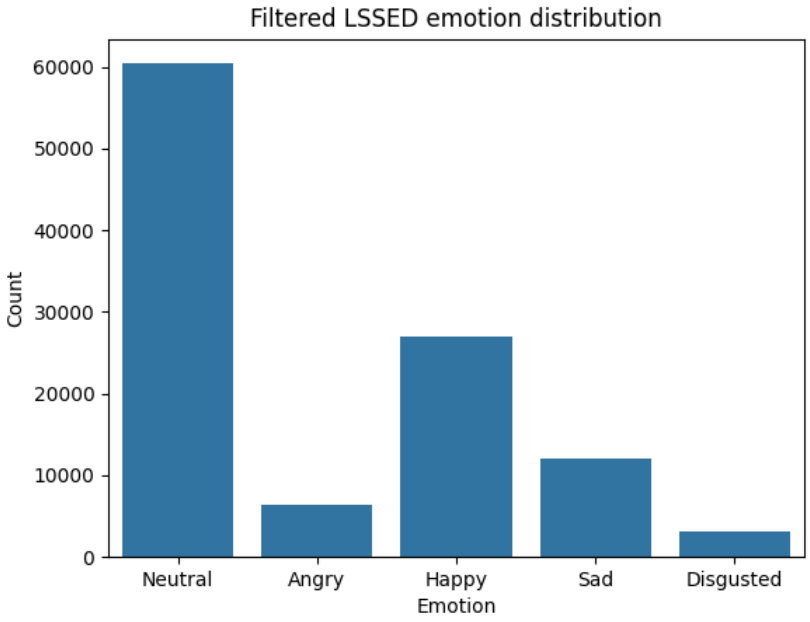
Zadatak analize za LSSED podatkovni skup je upoznati se s podacima unutar skupa te shvatiti i prikazati na koji su način raspoređeni i označeni u različite klase.



Slika 2.2. *Grafički prikaz distribucije primjeraka neobrađenog LSSED podatkovnog skupa prema izraženoj emociji*

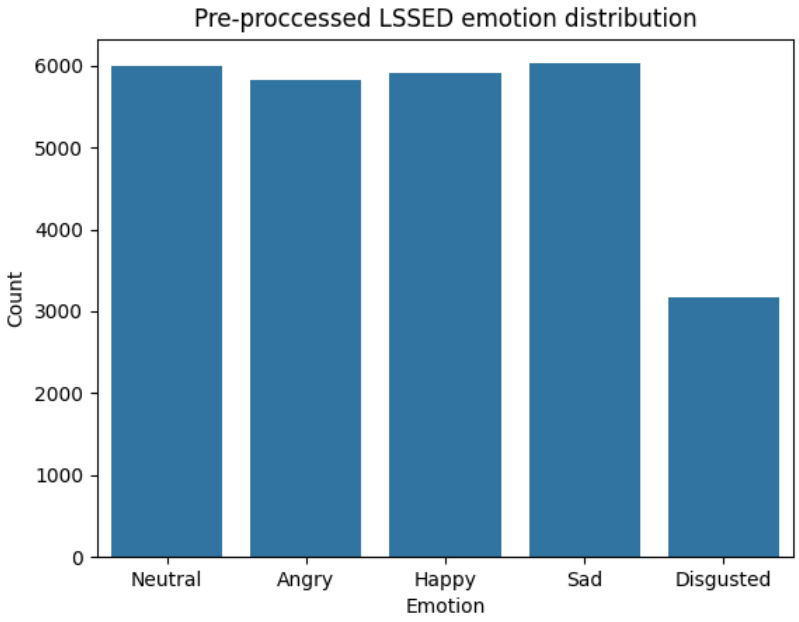
Prema slici 2.2. i prilogu 2.1. možemo uočiti ozbiljne neujednačene učestalosti emocija u LSSED podatkovnom skupu, što može dovesti do lošijeg generaliziranja parametara tijekom treniranja mreže. Također, problem predstavlja i velika količina primjeraka označenih sa *other*; podaci bez oznake specifične emocije nisu nimalo korisni za izradu ovog projektnog zadatka.

Kako bismo dobili vjerodostojniji i „bolji“ model, potrebno je prilagoditi LSSED podatkovni skup. Za početak, izbačeni su svi primjerci kojima je emocija označena kao *other*, a od preostalih primjeraka zadržani su oni čije se oznake za emocije prema Paulu Ekmanu svrstavaju u „osnovne“ emocije[8]: *angry*, *neutral*, *happy*, *sad*, *surprise*, *fear*, te oznaka za „neekspresivni“ govor. Prema prilogu 2.1. možemo uočiti kako među preostalim oznakama *surprise* i *fear* nemaju niti približno dovoljno primjeraka kako bi se model mogao kvalitetno trenirati na njima, pa su i primjerci označeni tim oznakama izbačeni iz skupa. Grafički prikaz distribucije primjeraka ovako dobivenog podatkovnog skupa vidljiv je na slici 2.3.



Slika 2.3. *Grafički prikaz distribucije primjeraka pročišćenog LSSED podatkovnog skupa prema izraženoj emociji*

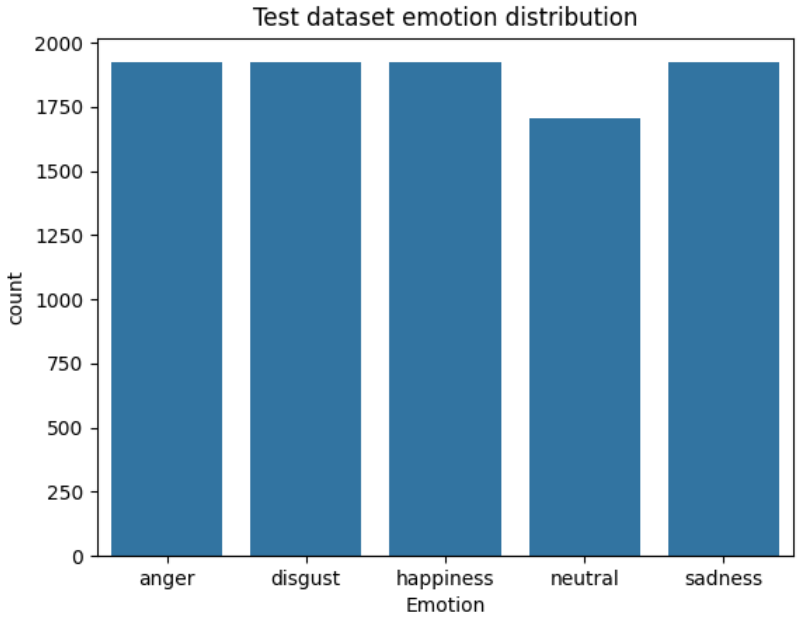
Podatkovni skup, prema prilozima 2.2. i 2.3., sadrži više primjeraka ženskog govora, te više primjeraka mladih i srednjovječnih osoba. Najviše primjeraka, prema slici 2.3., označeno je oznakom za „neekspresivni“ govor. Budući da podatkovni skup i dalje nije prihvatljivo ujednačen, provedena su daljnja uzorkovanja skupa prema trima kategorijama kojima su primjerci opisani te je dobiven potpuno predobrađeni podatkovni skup čije su relevantne distribucije vidljive na slici 2.4. te prilozima 2.4. i 2.5. Podatkovni skup podjeljen je na skup za treniranje i skup za evaluaciju u omjeru 80:20 te je ova podjela prikazana na prilogu 2.6.



Slika 2.4. *Distribucija primjeraka potpuno predobrađenog LSSED podatkovnog skupa prema izraženoj emociji*

2.2. SER-EN

Za testiranje modela korišten je SER-EN podatkovni skup, koji sveukupno sadrži 9395 *wav* datoteka označenih emocijama na koje treniramo modele ovog projektnog zadatka. Ovaj podatkovni skup nije bilo potrebno predobrađivati, već samo izbaciti primjerke označene emocijama na koje ne treniramo modele. Distribucija primjeraka prema oznakama prikazana je na slici 2.5.



Slika 2.5. *Distribucija primjeraka SER-EN podatkovnog skupa prema izraženoj emociji*

3. OPIS KORIŠTENIH TEHNOLOGIJA I ALGORITAMA

3.1. Google Colab

Google Colab, skraćenica od "Colaboratory", je besplatna usluga u oblaku koju pruža Google za izvođenje Python kôda[9], posebno orijentirana na podršku za rad s Jupyter bilježnicama. Ova platforma omogućuje korisnicima pristup virtualnim strojevima s NVIDIA GPU-ovima, što je korisno za izvođenje zahtjevnih operacija, poput rada s dubokim neuronskim mrežama.

3.2. Jupyter bilježnice

Jupyter bilježnice su interaktivna okolina za izvođenje kôda, pisanje teksta i vizualizaciju podataka[10]. Omogućavaju korisnicima stvaranje dokumenata koji kombiniraju programski kôd (najčešće u Pythonu) s formatiranim tekstom, slikama te rezultatima izvođenja kôda. Ova interaktivnost olakšava eksperimentiranje i analizu podataka.

3.3. PyTorch

PyTorch je biblioteka otvorenog izvora za duboko učenje[11] koja pruža mogućnost razvoja i treniranja neuronskih mreža. Razvijen od strane Facebookovog istraživačkog tima, PyTorch se ističe po svojoj dinamičkoj građi, što omogućava korisnicima da definiraju i prilagode svoje modele tijekom izvođenja kôda, što olakšava eksperimentiranje i brzo prototipiranje.

3.4. RNN, LSTM, i GRU slojevi

Za izradu mreže ovog projektnog zadatka korištena su tri različita modela mreže. Prvi korišteni slojevi su RNN[12] slojevi, vrsta neuronske mreže koja vrlo dobro radi sa podacima u nizu. Za razliku od uobičajenih neuronskih mreža, povratne sadrže povratnu vezu te su korisne za zadatke kod kojih su za izlaz bitna prethodna stanja ulaza. Nedostatak RNN-a je eksplodirajući, odnosno nestajući gradijent do kojeg može doći ako imamo preveliku odnosno premalu vrijednost težine u izlazu iz povratne veze. Vrijednosti za neki x broj ulaza mogu težiti beskonačnosti (eksplodirajući gradijent) ili 0 (nestajući gradijent).

U tom slučaju moguće je koristiti LSTM[13] slojeve. LSTM koristi tri vrste „vrata“ za kontrolu vrijednosti: ulazna vrata, vrata zaboravljanja i vrata izlaza. Ulazna vrata određuju koliko će nove informacije biti zadržano u memorijskoj ćeliji. Vrata zaboravljanja određuju koji će se dio informacija iz trenutne memorijske ćelije sačuvati, a izlazna vrata određuju koji

dio memorijske ćelije će se proslijediti kao izlaz. Pogodan je za korištenje u obradi prirodnog govora i prepoznavanju govora.

Treći slojevi slični su LSTM-u, a zovu se GRU[14]. Za razliku od LSTM-a koji ima troja vrata za ažuriranje stanja izlaza za dugoročnu i kratkoročnu memoriju, GRU ima dvoja vrata i samo jedan izlaz koji objedinjuje dugoročnu i kratkoročnu memoriju.

3.5. *Wav2Vec2 feature extractor* mreža

Wav2Vec2 je model za prepoznavanje govora razvijen od strane FAIR-a (*Facebook AI Research*, danas *Meta AI*)[15]. *Feature extractor* u *Wav2Vec2* se odnosi na dio modela koji je odgovoran za pretvaranje audio signala (*wav* datoteke) u opisne značajke pogodne za daljnju obradu i analizu. *Wav2Vec2 feature extractor* koristi CNN (*Convolutional neural network*) slojeve kako bi naučio reprezentacije značajki iz sirovih audio signala.

3.6. Struktura klasifikatora

Klasifikator stvoren za potrebe ovog projektnog zadatka sastoji se od jednog *LayerNorm* sloja[16], tri RNN/LSTM/GRU sloja, i jednog *Linear* sloja[17]. *LayerNorm* sloj upotrijebljen je analogno upotrebi *BatchNorm* slojeva kod CNN mreža: normalizira vrijednosti slojeva ulaznih podataka kako bi omogućili glađi gradijent, brže treniranje, i bolju generalizaciju klasifikatora. RNN/LSTM/GRU slojevi korišteni su na *many-to-one* način: iz cijelog izlaza slojeva uzeti su samo izlazi koji su nastali nakon što je cijeli ulazni podatak prošao kroz slojeve. *Linear* sloj korišten je za smanjivanje dimenzionalnosti izlaza sa veličine koju kao izlaz daju RNN/LSTM/GRU slojevi na polje sa samo 5 vrijednosti. Primjer definicije klasifikatora Python kôdom prikazan je na slici 3.1.

```python
class EmotionClassifier(nn.Module):
  def __init__(self):
    super().__init__()

    self.norm = nn.LayerNorm(normalized_shape = 1024)
    self.rnn1 = nn.LSTM(input_size = 1024, hidden_size = 1024,
                        num_layers = 3, batch_first = True,
                        bidirectional = False)
    self.linear1 = nn.Linear(1024, 5)

  def forward(self, x, length):
    out = self.norm(x)
    out, _ = self.rnn1(out)

    # Many-to-one RNN mod
    try:
      _ = out.shape[2]
      indices = [i for i in range(out.shape[0])]
      out = out[indices, np.subtract(length, 1), :]
    except:
      out = out[np.subtract(length, 1), :]

    out = self.linear1(out)
    return out
```

Slika 3.1. *Primjer definiranja klasifikatora*

3.7. Treniranje

Za treniranje mreže korištena je *categorical cross entropy* funkcija gubitka[18]. Ovakav pristup prikladan je za klasifikacijske zadatke poput ovog projektnog zadatka. Zbog nebalansiranog podatkovnog skupa (prema slici 2.4. vidljivo je kako klasa *disgust* ima manje primjeraka u odnosu na ostale emocije) funkciji su za svaku klasu dodjeljene težine izračunate prema (3-1) kako primjerci klase *disgust* ne bi bili „zanemareni“ prilikom treniranja.

$$težina_i = \frac{ukupna\ količina\ primjeraka\ podatkovnog\ skupa}{količina\ primjeraka\ klase\ i} \tag{3-1}$$

Za optimizator je korišten *Adam*[19], točnije *AMSGrad* verzija koja koristi maksimum prošlih kvadriranih gradijenata umjesto eksponencijalnog prosjeka za ažuriranje parametara, u pokušaju rješavanja problema konvergiranja od kojih optimizatori bazirani na Adam algoritmu često pate[20]. Korišten je i *StepLR scheduler* stope učenja[21], uz pomoć koje je moguće promijeniti vrijednost stope učenja u ovisnosti o broju proteklih epoha treniranja. Na slici 3.2. prikazan je Python kôd kojim su inicijalizirani opisani algoritmi.

```
loss_module = nn.CrossEntropyLoss(weight = torch.FloatTensor(
    [21545./4663., 21545./4724., 21545./4800.,
     21545./4822., 21545./2536.]).to(DEVICE),
                                  label_smoothing = 0.1)
optimizer = torch.optim.Adam(model.parameters(), lr = 0.00005,
                             amsgrad = True, fused = True)
scheduler = torch.optim.lr_scheduler.StepLR(optimizer,
                                            step_size = 25, gamma = 0.5)
```

Slika 3.2. *Prikaz Python kôda za inicijalizaciju algoritama korištenih pri treniranju*

Za procjenu performansi modela tijekom treniranja i evaluacije, koja slijedi nakon svake epohe treniranja, korištene su metrike: funkcija gubitka, točnost, preciznost, odziv, i *F1-score*. Jednadžbe za računanje metrika, osim funkcije gubitka, nalaze se u prilogu 3.1.

3.8. Interpretacija rezultata treniranja modela

Prvi trenirani model je klasifikator s 3 RNN sloja, čije je treniranje provedeno na 75 epoha. Grafovi metrika praćenih tijekom treniranja, za koje su vrijednosti izračunate tijekom evaluacije nakon svake epohe treniranja, nalaze se u prilozima 3.2. – 3.6. Iz priloga 3.2., 3.5., te 3.6. uočljivo je iz čestih rasta i padova vrijednosti kako RNN slojevi nisu dovoljno složeni za kvalitetnu generalizaciju parametara na ulazima dugačkih nizova. Također, iz priloga 3.3. vidljivo je kako je oko 55. epohe model počeo patiti od *overfitting*-a. Za evaluaciju su odabrani modeli nastali nakon 22., 37.-42., 45.-50., te nakon 52.-54. epoha.

Sljedeći je trenirani klasifikator s 3 LSTM sloja, čije je treniranje provedeno na 40 epoha. Grafovi praćenih metrika nalaze se u prilozima 3.7. – 3.11. Uspoređujući grafove s grafovima prošlog klasifikatora, odmah je uočljivo kako je LSTM prikladniji za zadatak klasifikacije govora prema izraženoj emociji. Iz priloga 3.8. vidljivo je kako se kod ovog klasifikatora *overfitting* počeo pojavljivati već oko 30. epohe. Gledajući grafikone u prilozima 3.7. i 3.9., za evaluaciju su odabrani modeli nastali nakon 17.-28. epoha.

Posljednji trenirani klasifikator je model s 3 GRU sloja, čije je treniranje provedeno na 50 epoha. Grafovi praćenih metrika nalaze se u prilozima 3.12. – 3.16. Iz priloga 3.13. lako je uočljivo kako je treniranje moglo biti prekinuto puno ranije, no to ne predstavlja problem jer su stanja modela spremana nakon svake epohe. Gledajući priloge 3.12., 3.14., te 3.16., očito je kako je i ovaj klasifikator „bolji“ od klasifikatora s RNN slojevima, no nije jasno koliko je točno „bolji“ od klasifikatora s LSTM slojevima. Za evaluaciju su odabrani modeli nastali nakon 13.-25. epoha.

4. EVALUACIJA IZRAĐENIH MODELA

Klasifikatori su testirani na SER-EN podatkovnom skupu, tako da su povezani u veću mrežu u kojoj izlaz *Wav2Vec2 feature extractor* mreže čini ulaz u trenirani klasifikator. Mreža je definirana prema Python kôdu prikazanom na slici 4.1.

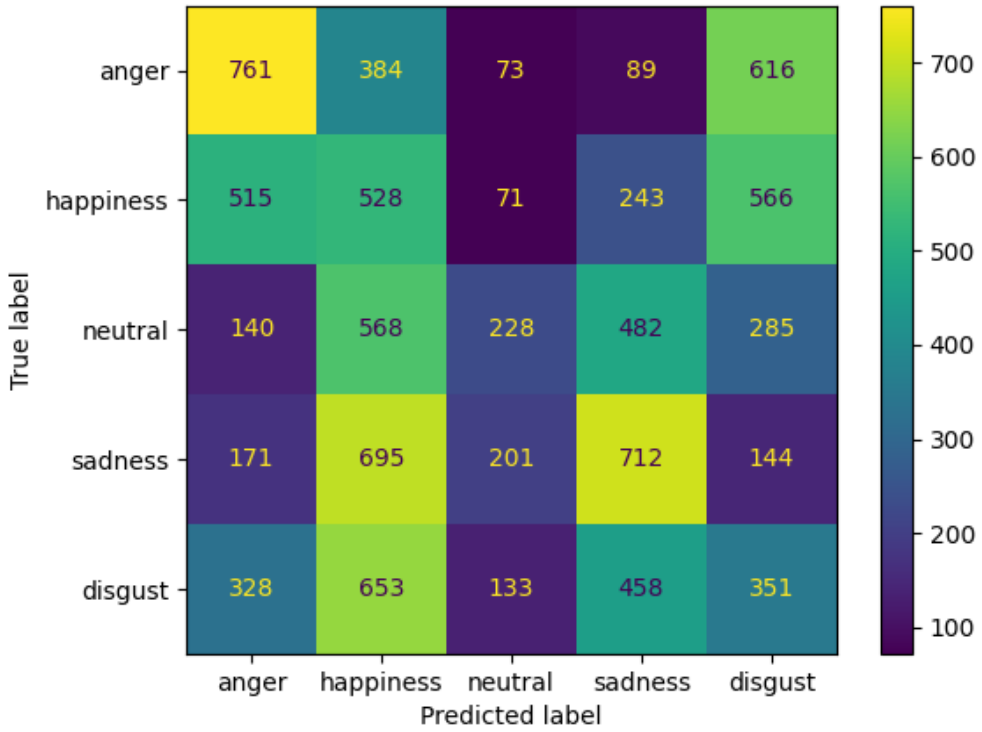
```
class Emotioner(nn.Module):
  def __init__(self, feature_extractor, wav2vec2_model, emotion_classifier, sampling_rate = 16000):
    super().__init__()
    self.feature_extractor = feature_extractor
    self.wav2vec2_model = wav2vec2_model
    self.emotion_classifier = emotion_classifier
    self.sampling_rate = sampling_rate

  def extract_features(self, wav_array, sampling_rate):
    wavs_token = self.feature_extractor([wav_array], sampling_rate = sampling_rate,
                                        padding = True, do_normalize = True,
                                        return_tensors = 'pt').to(DEVICE)
    outputs = self.wav2vec2_model(**wavs_token, output_hidden_states = True)
    w2vlastfeat = outputs['last_hidden_state'].squeeze().detach().cpu().numpy()
    feature_array = torch.FloatTensor(w2vlastfeat).to(DEVICE)
    return feature_array

  def forward(self, wav_array):
    features = self.extract_features(wav_array, self.sampling_rate)
    output = self.emotion_classifier(features, features.shape[0])
    _, pred_label = torch.max(output.data, dim = 0)
    return (output, pred_label)
```

Slika 4.1. *Prikaz Python kôda koji definira mrežu korištenu prilikom testiranja*

Testiranjem klasifikatora s RNN slojevima, najbolje performanse je pokazao klasifikator nastao nakon 47. epohe. Metrike ovog klasifikatora izračunate tijekom testiranja prikazane su tablicom 4.1., a matrica zabune prikazana je na slici 4.2.

TOČNOST:	27.46%
PRECIZNOST:	28.82%
ODZIV:	27.46%
F1-SCORE:	27.26%

Tablica 4.1. *Metrike klasifikatora s RNN slojevima*



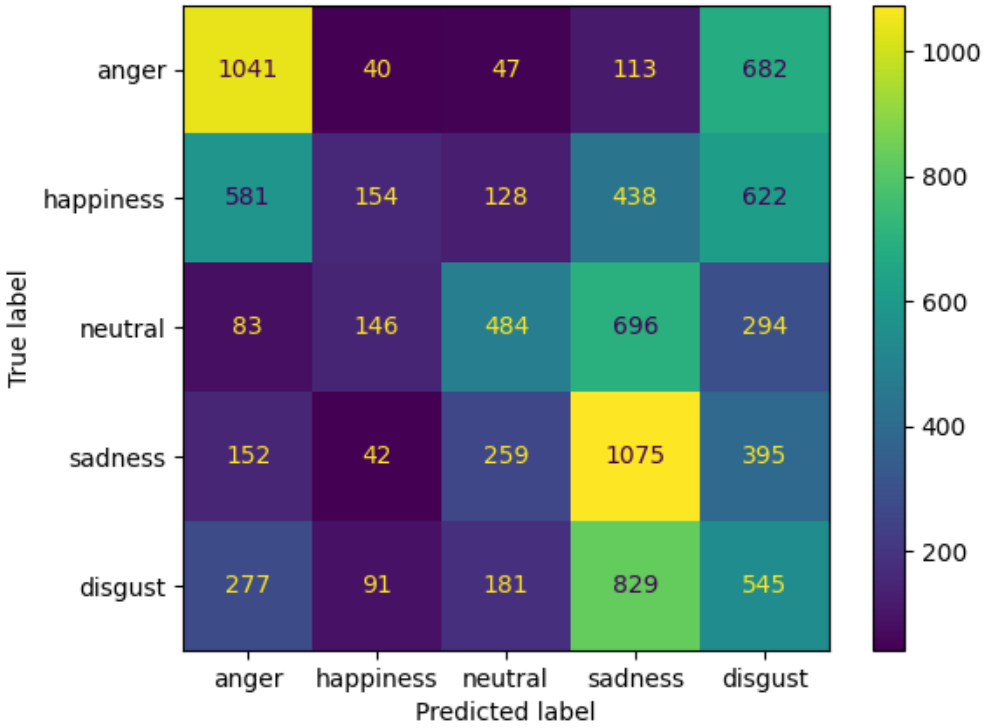
Slika 4.2. *Matrica zabune klasifikatora s RNN slojevima*

Iz matrice prikazane na slici 4.2. vidljivo je kako model najbolje klasificira snimke u kojima je izražena ljutnja (39.57% točnih klasifikacija) i tuga (37.03% točnih klasifikacija), ali istovremeno s njima ima i neke od najekstremnijih problema: ljutnju krivo klasificira kao gađenje u 32.03% slučajeva, a tugu krivo klasificira kao sreću u 36.14% slučajeva. Najlošiji je u klasifikaciji gađenja (18.25% točnih klasifikacija) i u klasifikaciji „neekspresivnog“ govora (svega 13.39% točnih klasifikacija).

Među klasifikatorima s LSTM slojevima, najbolje performanse pokazao je klasifikator nastao nakon 28. epohe. Metrike klasifikatora izračunate tijekom testiranja prikazane su u tablici 4.2., a matrica zabune prikazana je na slici 4.3.

TOČNOST:	35.11%
PRECIZNOST:	36.01%
ODZIV:	35.11%
F1-SCORE:	33.07%

Tablica 4.2. *Metrike klasifikatora s LSTM slojevima*



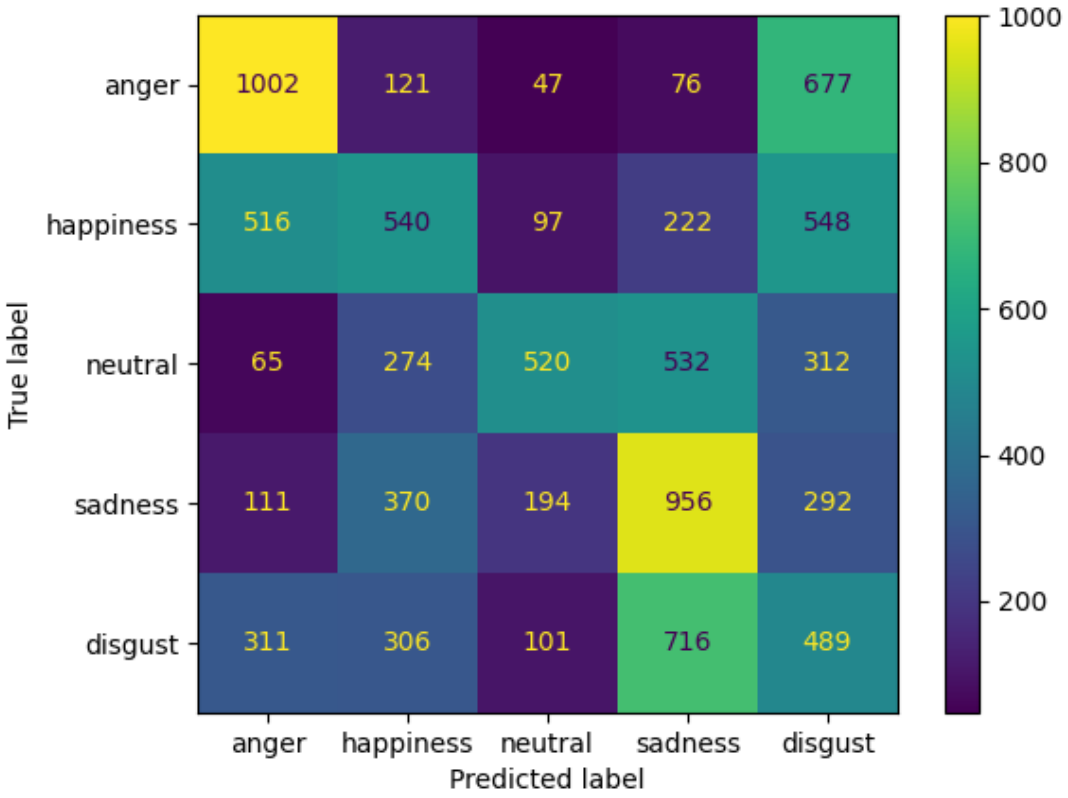
Slika 4.3. *Matrica zabune klasifikatora s LSTM slojevima*

Prema matrici zabune prikazanoj na slici 4.3., vidljivo je kako je model izvrstan u klasificiranju zapisa u kojima je izražena tuga (55.90% točnih klasifikacija) i ljutnja (54.13% točnih klasifikacija). S druge strane, sreću gotovo pa ni ne prepoznaje (svega 8.01% točnih klasifikacija), već ju najčešće klasificira kao gađenje (u 32.35% slučajeva) ili kao ljutnju (u 30.21% slučajeva).

Posljednji klasifikatori su oni s GRU slojevima, kod kojih je najbolje performanse pokazao klasifikator nastao nakon 26. epohe. Metrike klasifikatora izračunate tijekom testiranja prikazane su u tablici 4.3., a matrica zabune prikazana je na slici 4.4.

TOČNOST:	37.33%
PRECIZNOST:	39.06%
ODZIV:	37.33%
F1-SCORE:	37.34%

Tablica 4.3. *Metrike klasifikatora s GRU slojevima*



Slika 4.4. *Matrica zabune klasifikatora s GRU slojevima*

Prema matrici zabune na slici 4.4., klasifikator s GRU slojevima najbolje klasificira zapise u kojima je izražena ljutnja (52.11% točnih klasifikacija) i tuga (49.71% točnih klasifikacija). Što je još važnije, razlika između klase koju najbolje prepoznaje (ljutnja – 52.11%) i klase koju najlošije prepoznaje (gađenje – 25.43%) je najmanja od svih testiranih modela. Iz tog razloga, ali i iz usporedbe metrika sva tri klasifikatora, uočljivo je kako je klasifikator s GRU slojevima najbolji izrađeni klasifikator. Na slici 4.5. prikazana je usporedba matrice zabune klasifikatora kojeg su stvorili tvorci LSSED podatkovnog skupa[1] i matrice zabune klasifikatora s GRU slojevima izrađenog za ovaj projektni zadatak.

	Angry	Neutral	Happy	Sad
Angry	32.43%	47.23%	16.56%	3.78%
Neutral	1.45%	86.85%	9.54%	2.17%
Happy	2.77%	59.85%	35.18%	2.20%
Sad	2.12%	71.04%	9.83%	17.01%

True label						
	Anger	52,11%	6,29%	2,44%	3,95%	35,21%
	Happiness	26,83%	28,08%	5,04%	11,54%	28,50%
	Neutral	3,82%	16,09%	30,53%	31,24%	18,32%
	Sadness	5,77%	19,24%	10,09%	49,71%	15,18%
	Disgust	16,17%	15,91%	5,25%	37,23%	25,43%
		Anger	Happiness	Neutral	Sadness	Disgust
			Predicted label			

Slika 4.5. *Usporedba matrica zabune*

Prema matrici zabune klasifikatora tvoraca LSSED podatkovnog skupa, uočljivo je kako je model vrlo sklon klasificirati primjerak kao „neekspresivan“ govor, zbog čega je iznimno točan u prepoznavanju te klase, ali istovremeno najveće pogreške u klasifikaciji su upravo pogrešno klasificiran zapis kao „neekspresivan“ govor. Ova pojava nije neočekivana, jer je klasifikator treniran na svim primjercima *angry*, *neutral*, *happy*, i *sad* klasa u LSSED podatkovnom skupu, a prema slici 2.1. uočljivo je kako gotovo polovina svih primjeraka u podatkovnom skupu su označeni kao „neekspresivan“ govor. Naš klasifikator prepoznaje „neekspresivan“ puno lošije u usporedbi sa prvim klasifikatorom, ali iako je treniran na 5 klasa, u usporedbi sa prvim klasifikatorom koji je treniran na 4 klase, najniža točnost za jednu klasu (gađenje – 25.43%) je veća od najniže točnosti za jednu klasu prvog klasifikatora (tuga – 17.01%). Na slici 4.6. prikazane su točnosti svih klasifikatora koje su tvorci LSSED podatkovnog skupa trenirali i testirali.

Algorithm	Backbone	WA	UA
FCN-Attention[23]	ALEXNet	0.570	0.250
MTS-3branches[24]	ALEXNet	0.570	0.250
MTS-5branches[24]	ALEXNet	0.570	0.250
MTS-3branches[24]	ResNet152	0.585	0.296
MTS-5branches[24]	ResNet152	0.582	0.311
ADV-Real[25]	VGG16	0.570	0.250
ADV-Fake[25]	VGG16	0.570	0.250
ADV-Real[25]	ResNet152	0.548	0.381
ADV-Fake[25]	ResNet152	0.453	0.339
VGG[20]	VGG11	0.595	0.337
VGG[20]	VGG13	0.604	0.393
VGG[20]	VGG16	0.585	0.313
VGG[20]	VGG19	0.585	0.370
ResNet[21]	ResNet18	0.594	0.382
ResNet[21]	ResNet34	0.598	0.355
ResNet[21]	ResNet50	0.587	0.377
ResNet[21]	ResNet101	0.592	0.332
ResNet[21]	ResNet152	0.601	0.396
PyResNet	ResNet50	**0.615**	**0.420**
PyResNet	ResNet101	**0.616**	**0.428**
PyResNet	ResNet152	**0.624**	**0.429**

Slika 4.6. *Rezultati testiranja klasifikatora treniranih na LSSED podatkovnom skupu*

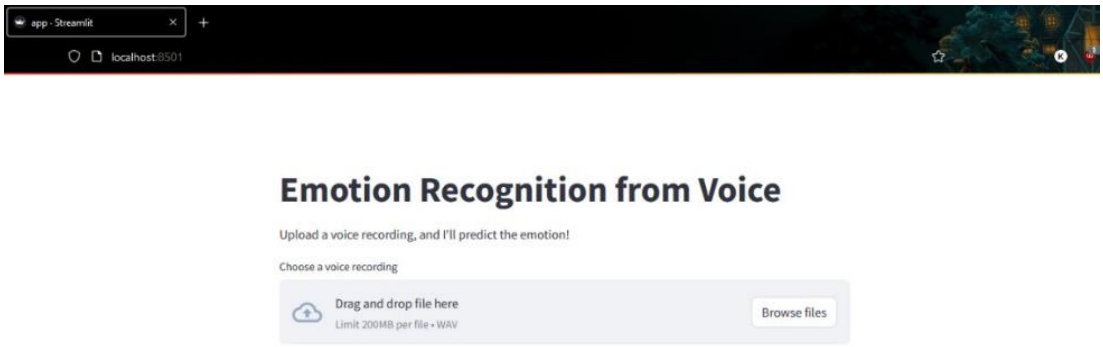
Prema slici 4.6., prvo uočljivo je kako su svi testirani klasifikatori mnogo dublji od našeg najboljeg klasifikatora: najplići klasifikator je *VGG11* – svega 11 slojeva, dok je najdublji *ResNet152* – 152 sloja. UA ovdje označava *unweighted accuracy*, odnosno točnost koja se izračunava prema formuli iz priloga 3.1., dok WA označava *weighted accuracy*, odnosno točnost izračunatu prema formuli (4-1).

$$WA = \frac{\sum_i^5 \frac{TP_i}{TP_i + FN_i} + \frac{TN_i}{TN_i + FP_i}}{5} \quad (4\text{-}1)$$

Za naš model, UA je jednak 37.33%, a WA je jednak 54.25%, čime je naš model usporediv sa većinom modela sa slike 4.6, potencijalno je i bolji od većine njih jer su modeli sa slike trenirani na samo 4 klase, a klasifikator ovog projektnog zadatka je treniran na 5 klasa. Time smo dodatno dokazali da kvalitetno predobrađeni podatkovni skup više utječe na performanse modela i sposobnost modela da kvalitetno generalizira parametre, nego broj slojeva.

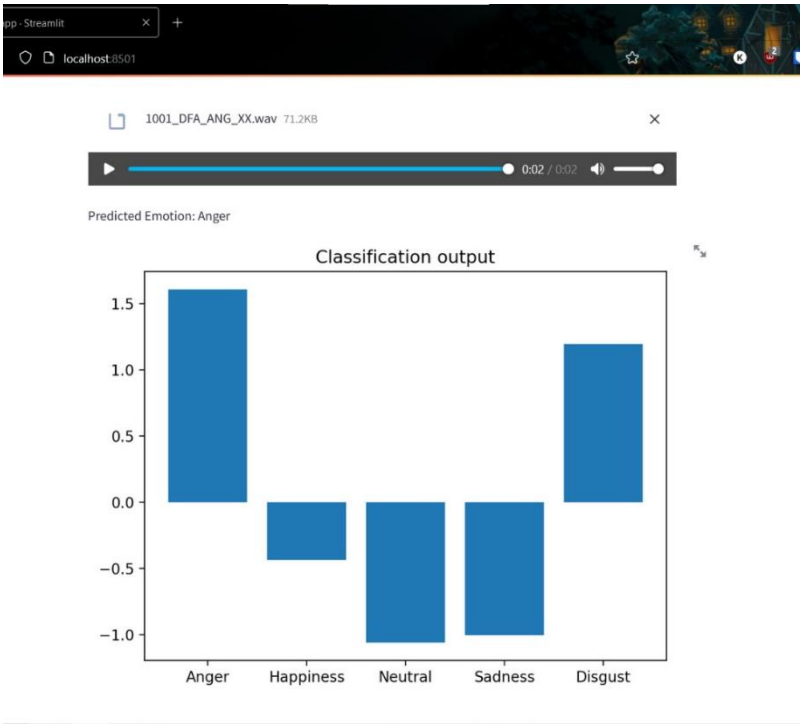
5. DEMONSTRACIJA RADA

Model prikazan na slici 4.1., zajedno s klasifikatorom s GRU slojevima, implementiran je u jednostavnu *Streamlit*[22] aplikaciju čije je sučelje prikazano na slici 5.1.



Slika 5.1. *Prikaz sučelja izrađene aplikacije*

Unutar aplikacije moguće je jednostavno učitati *wav* datoteku s računala korisnika, nakon čega aplikacija automatski obrađuje zapis korištenjem modela, te prikazuje *audio player*, prepoznatu emociju, te izlaz iz modela u obliku grafikona. Sučelje aplikacije s rezultatima klasifikacije prikazano je na slici 5.2.



Slika 5.2. *Prikaz sučelja aplikacije s rezultatima klasifikacije zapisa*

6. ZAKLJUČAK

Metodom nadziranog učenja i korištenjem relevantnog podatkovnog skupa (LSSED), te PyTorch biblioteke uspješno je realiziran model sposoban klasificirati govor osobe u jednu od klasa koje označavaju emocije. Za uspješnu realizaciju potrebno je obraditi dobivene podatke te ih pripremiti za učenje modela. U radu je napravljena usporedba tri relevantne metode za klasifikaciju govora: RNN, LSTM i GRU. Svaka od njih dala je svoje rezultate, a kao najbolji model istaknu se GRU. Najbolji model naknadno je evaluiran na SER podatkovnom skupu.

Model se može unaprijediti uvođenjem dodatnih slojeva neuronskim mreža za koje tijekom izvođenja ovog rada nije bilo dovoljno vremenskih i računalnih resursa. Moguće je dodatno očistiti i unaprijediti ulazni podatkovni skup te ga nadopuniti podacima za još veću mogućnost generalizacije.

LITERATURA

[1] Fan W, Xu X, Xing X, et. al., "LSSED: a large-scale dataset and benchmark for speech emotion recognition," *ICASSP*, pp. 641–645, 2021.

[2] "Wav2Vec2" [online]. Available: https://huggingface.co/transformers/v4.9.2/model_doc/model_doc/wav2vec2.html. [Accessed: 31.1.2024.].

[3] "Speech Emotion Recognition (en)" [online]. Available: https://www.kaggle.com/datasets/dmitrybabko/speech-emotion-recognition-en. [Accessed: 31.1.2024.].

[4] H., Cao, D. G., Cooper, M. K., Keutmann, R. C., Gur, A., Nenkova, R., Verma, "CREMA-D: Crowd-sourced Emotional Multimodal Actors Dataset," *IEEE Trans Affect Comput*, no. 4, vol. 5, pp. 377–390, 2014.

[5] M. K., Pichora-Fuller, K., Dupuis, "Toronto emotional speech set (TESS)." Borealis, 2020.

[6] S. R., Livingstone, F. A., Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)." Zenodo, 05-Apr-2018.

[7] "Surrey Audio-Visual Expressed Emotion (SAVEE) Database" [online]. Available: http://kahlan.eps.surrey.ac.uk/savee/. [Accessed: 31.1.2024.].

[8] "Emotion classification," *Wikipedia*. 29-Dec-2023.

[9] "colab.google" [online]. Available: http://0.0.0.0:8080/. [Accessed: 31.1.2024.].

[10] "Project Jupyter" [online]. Available: https://jupyter.org. [Accessed: 31.1.2024.].

[11] "Features" [online]. Available: https://pytorch.org/features/. [Accessed: 31.1.2024.].

[12] "RNN — PyTorch 2.2 documentation" [online]. Available: https://pytorch.org/docs/stable/generated/torch.nn.RNN.html. [Accessed: 31.1.2024.].

[13] "LSTM — PyTorch 2.2 documentation" [online]. Available: https://pytorch.org/docs/stable/generated/torch.nn.LSTM.html. [Accessed: 31.1.2024.].

[14] "GRU — PyTorch 2.2 documentation" [online]. Available: https://pytorch.org/docs/stable/generated/torch.nn.GRU.html. [Accessed: 31.1.2024.].

[15] "wav2vec" [online]. Available: https://ai.meta.com/research/impact/wav2vec. [Accessed: 31.1.2024.].

[16] "LayerNorm — PyTorch 2.2 documentation" [online]. Available: https://pytorch.org/docs/stable/generated/torch.nn.LayerNorm.html#layernorm. [Accessed: 31.1.2024.].

[17] "Linear — PyTorch 2.2 documentation" [online]. Available: https://pytorch.org/docs/stable/generated/torch.nn.Linear.html#linear. [Accessed: 31.1.2024.].

[18] "CrossEntropyLoss — PyTorch 2.2 documentation" [online]. Available: https://pytorch.org/docs/stable/generated/torch.nn.CrossEntropyLoss.html. [Accessed: 31.1.2024.].

[19] "Adam — PyTorch 2.2 documentation" [online]. Available: https://pytorch.org/docs/stable/generated/torch.optim.Adam.html#adam. [Accessed: 31.1.2024.].

[20] "Papers with Code - AMSGrad Explained" [online]. Available: https://paperswithcode.com/method/amsgrad. [Accessed: 31.1.2024.].

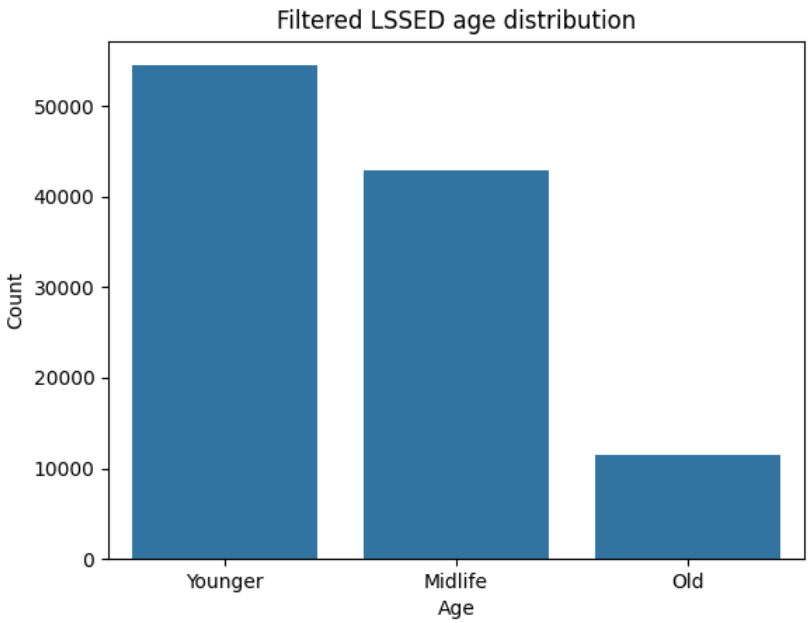
[21] "StepLR — PyTorch 2.2 documentation" [online]. Available: https://pytorch.org/docs/stable/generated/torch.optim.lr_scheduler.StepLR.html#steplr. [Accessed: 31.1.2024.].

[22] "Streamlit • A faster way to build and share data apps" [online], 14-Jan-2021. Available: https://streamlit.io/. [Accessed: 31.1.2024.].

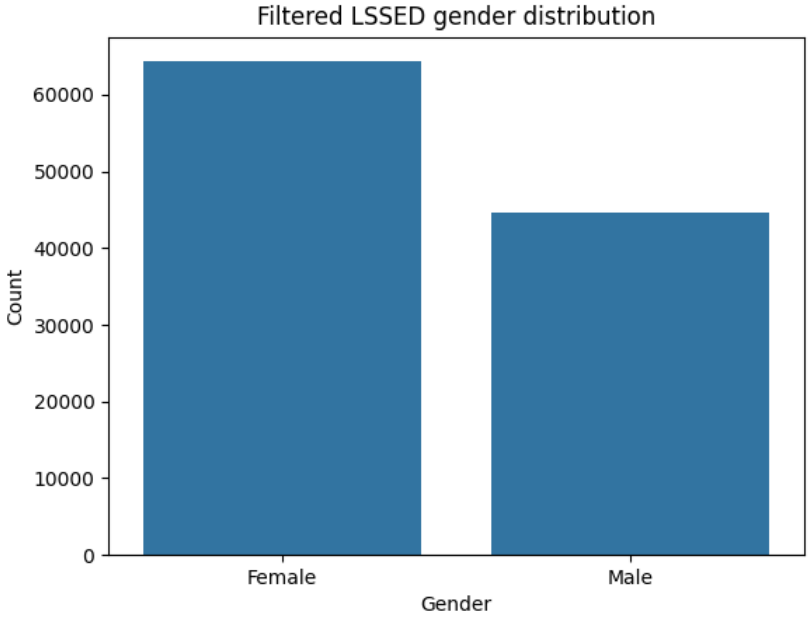
PRILOZI

	TRAINING	TEST	TOTAL
ANGRY	5192	1298	6490
NEUTRAL	48369	12092	60461
HAPPY	21627	5406	27033
SAD	9641	2410	12051
DISAPPOINTED	7124	1781	8905
BORED	2333	583	2916
DISGUSTED	2543	636	3179
EXCITED	4182	1046	5228
SURPRISE	1325	331	1656
FEAR	502	126	628
OTHER	14782	3696	18478
TOTAL	117620	29405	147025

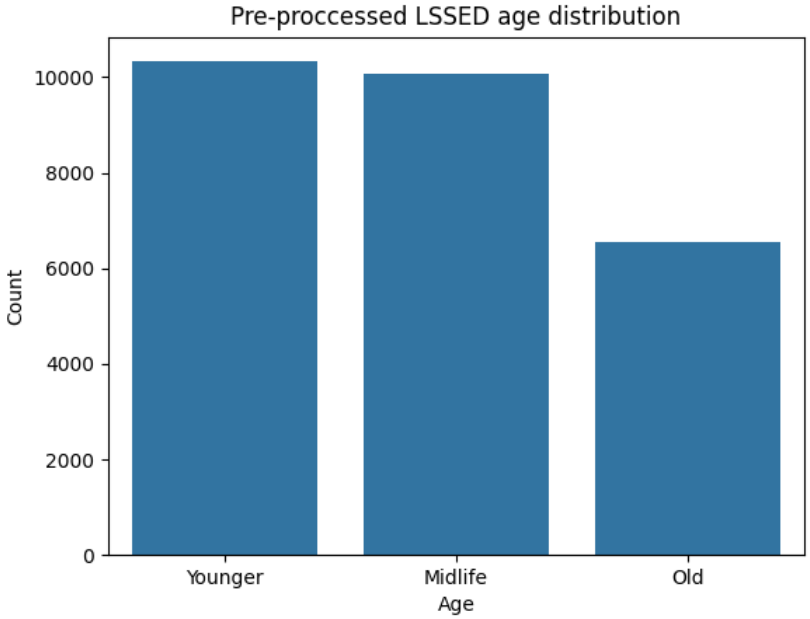
Prilog 2.1. *Distribucija primjeraka neobrađenog LSSED podatkovnog skupa*



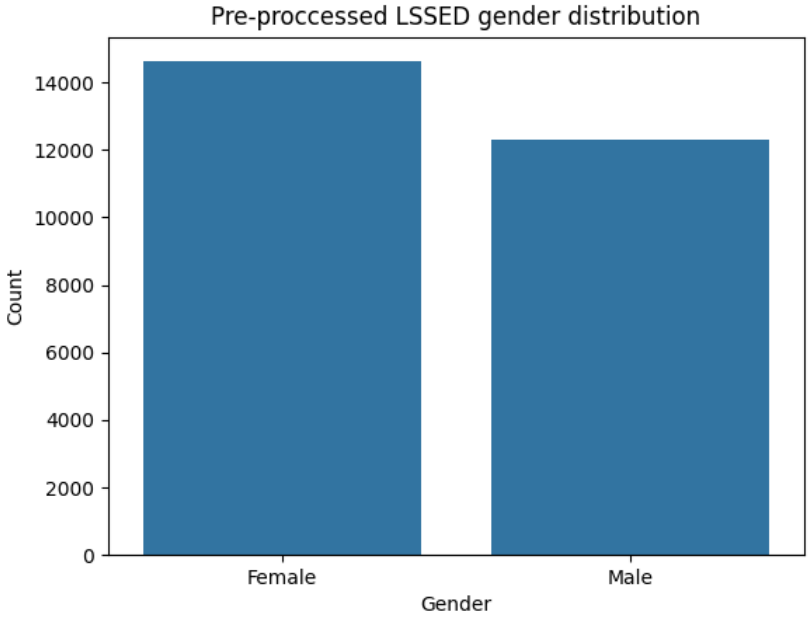
Prilog 2.2. *Distribucija primjeraka pročišćenog LSSED podatkovnog skupa prema dobi govornika*



Prilog 2.3. *Distribucija primjeraka pročišćenog LSSED podatkovnog skupa prema spolu govornika*



Prilog 2.4. *Distribucija primjeraka potpuno predobrađenog LSSED podatkovnog skupa prema dobi govornika*



Prilog 2.5. *Distribucija primjeraka potpuno predobrađenog LSSED podatkovnog skupa prema spolu govornika*

	TRAINING	**EVAL**	**TOTAL**
ANGRY	4663	1164	5827
NEUTRAL	4724	1181	5905
HAPPY	4800	1200	6000
SAD	4822	1205	6027
DISGUSTED	2536	632	3168
TOTAL	21545	5382	26927

Prilog 2.6. *Podjela LSSED podatkovnog skupa na skup za treniranje i skup za evaluaciju*

$$točnost = \frac{količina\ točnih\ klasifikacija}{ukupna\ količina\ klasifikacija}$$

$$težina_i = \frac{količina\ primjeraka\ klase\ i}{ukupna\ količina\ primjeraka}$$

$$preciznost_i = \frac{TP_i}{TP_i + FP_i + 1e^{-10}}$$

$$ukupna\ preciznost = \sum težina_i * preciznost_i$$

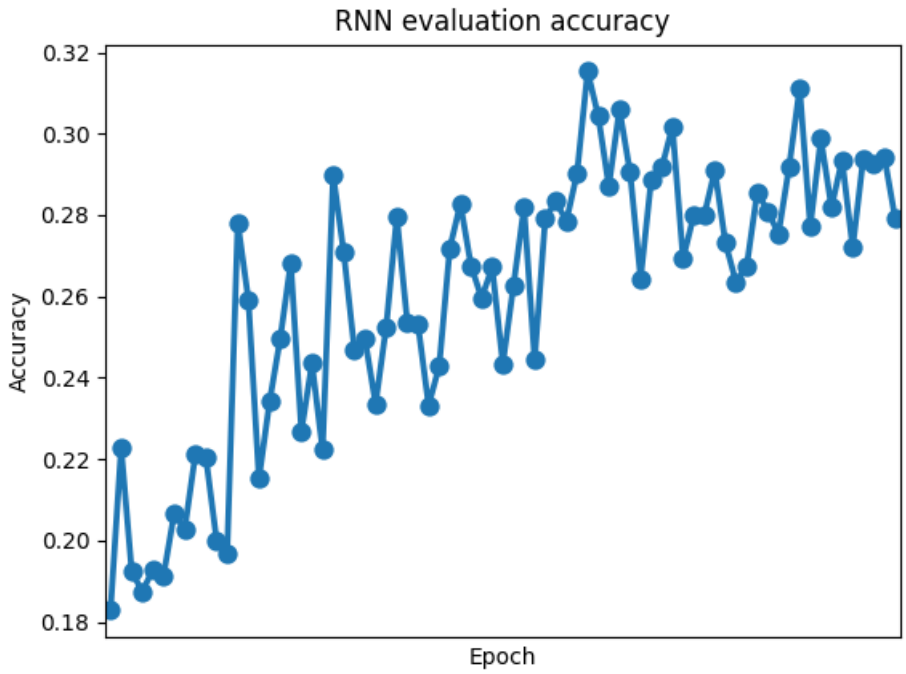
$$odziv_i = \frac{TP_i}{TP_i + FN_i + 1e^{-10}}$$

$$ukupni\ odziv = \sum težina_i * odziv_i$$

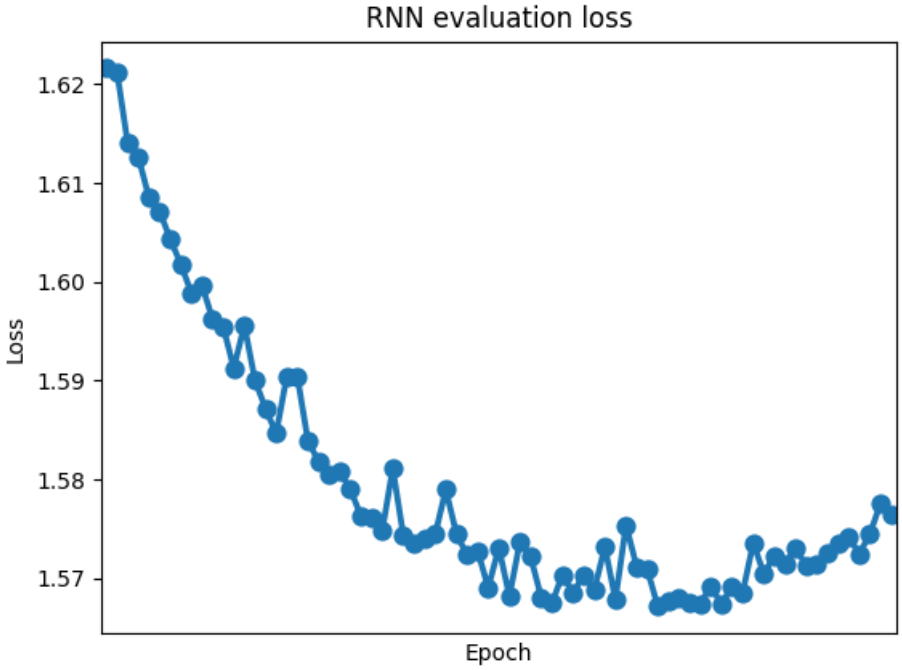
$$F1_i = \frac{2 * preciznost_i * odziv_i}{preciznost_i + odziv_i + 1e^{-10}}$$

$$ukupni\ F1 = \sum težina_i * odziv_i$$

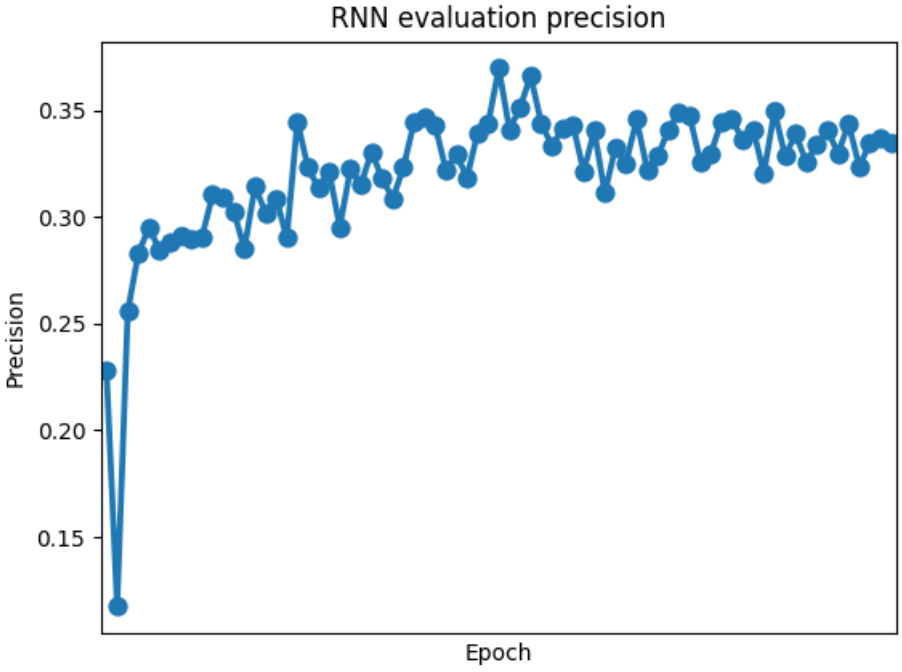
Prilog 3.1. *Jednadžbe korištene za računanje metrika*



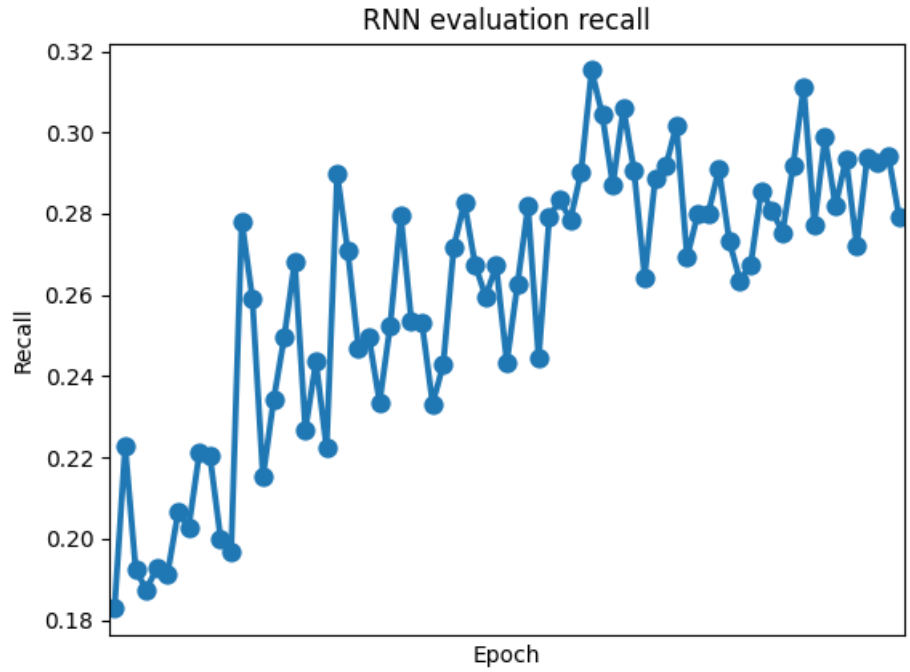
Prilog 3.2. *Grafikon točnosti klasifikatora s RNN slojevima*



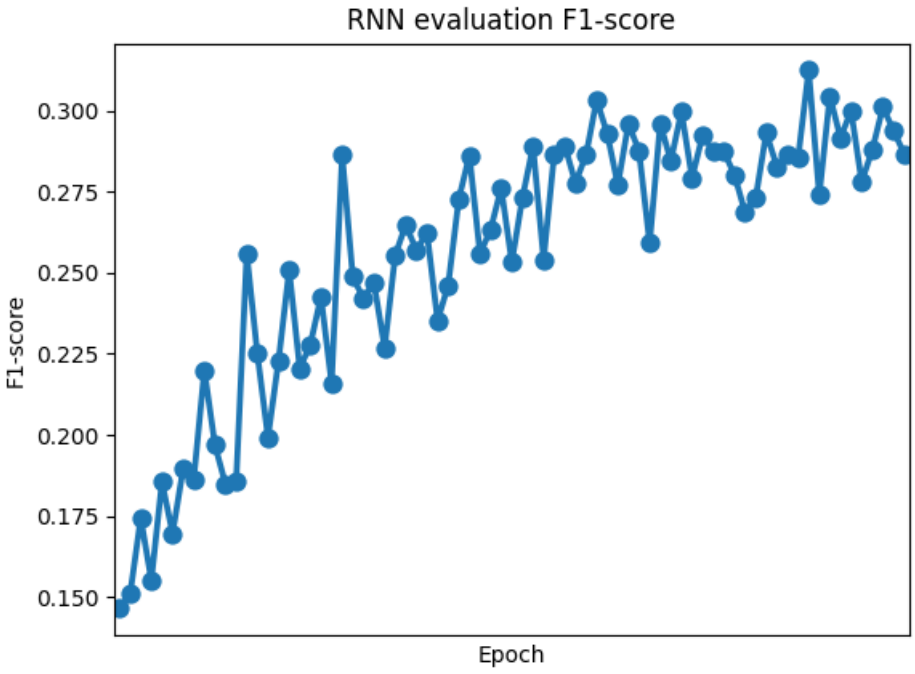
Prilog 3.3. *Grafikon funkcije gubitka klasifikatora s RNN slojevima*



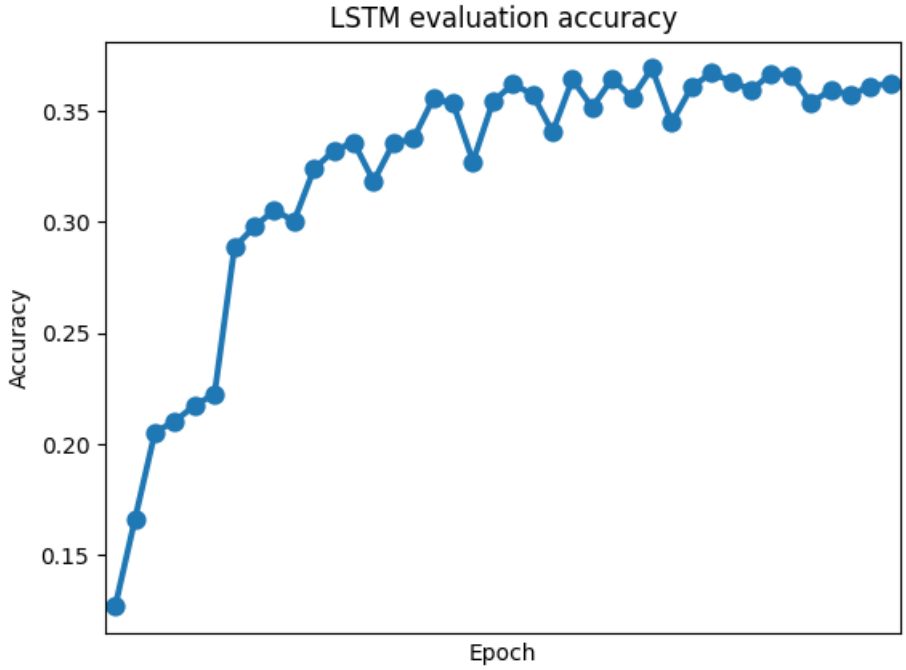
Prilog 3.4. *Grafikon preciznosti klasifikatora s RNN slojevima*



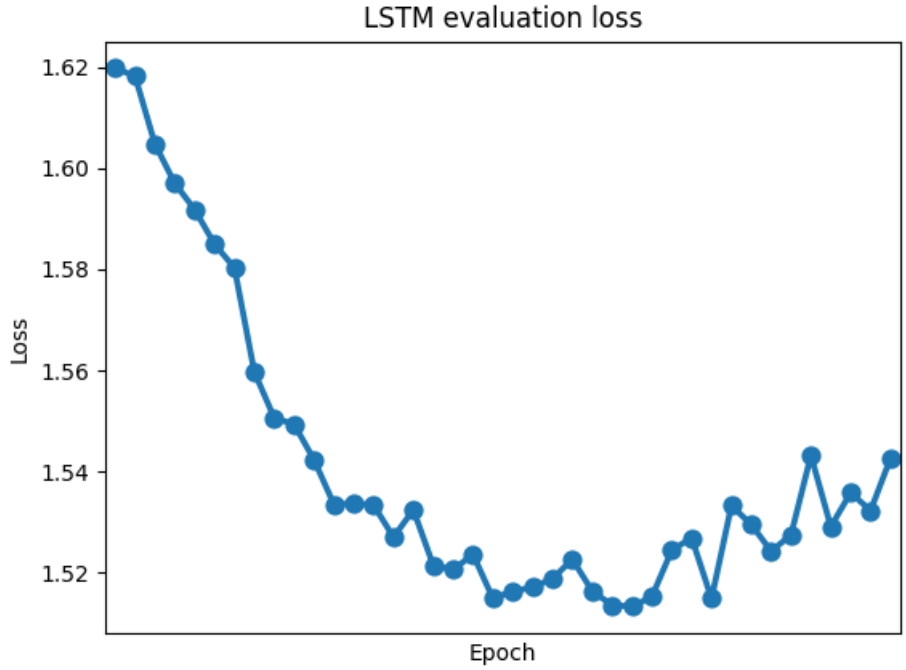
Prilog 3.5. *Grafikon odziva klasifikatora s RNN slojevima*



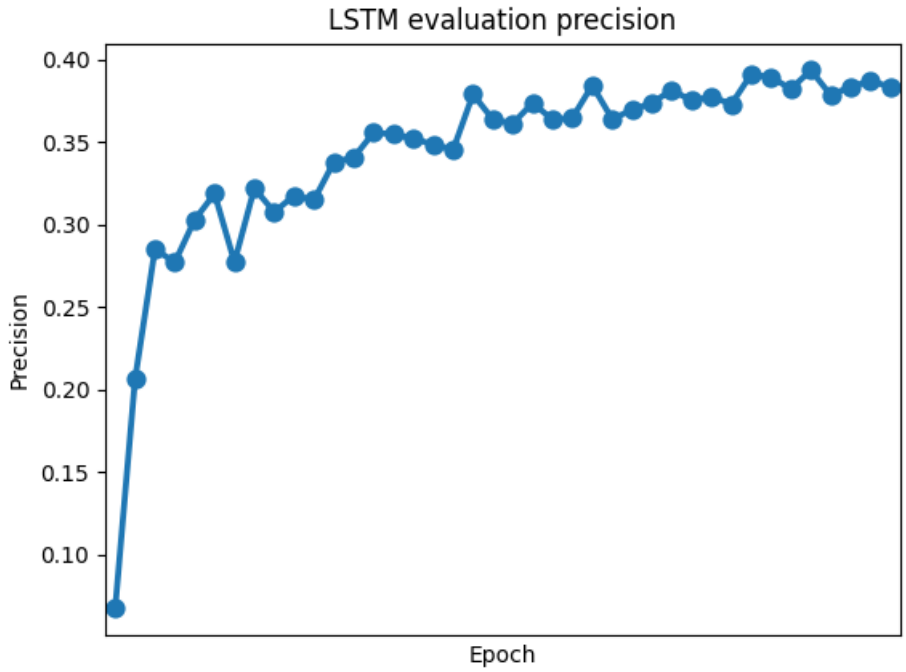
Prilog 3.6. *Grafikon F1-score klasifikatora s RNN slojevima*



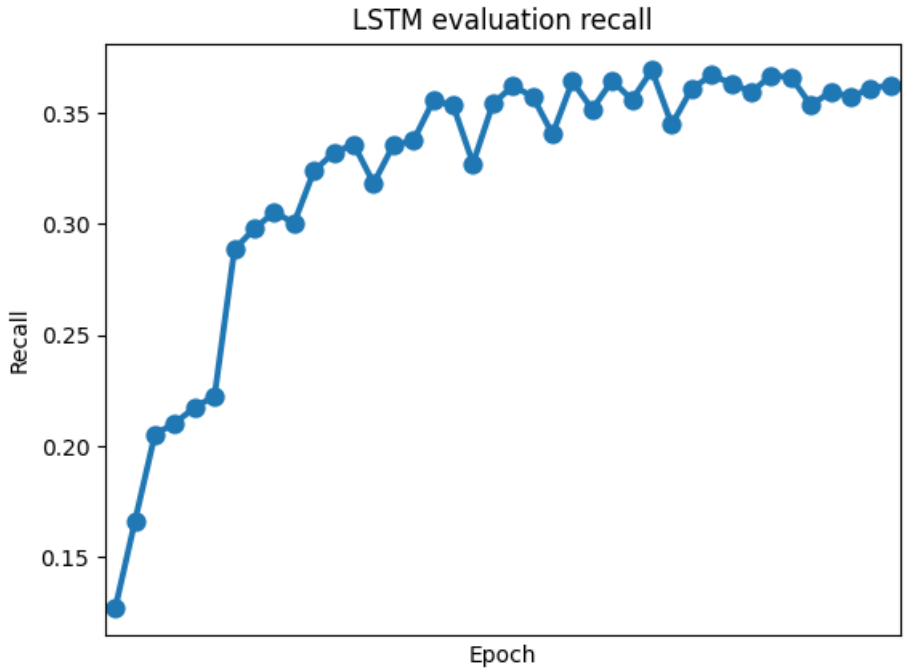
Prilog 3.7. *Grafikon točnosti klasifikatora s LSTM slojevima*



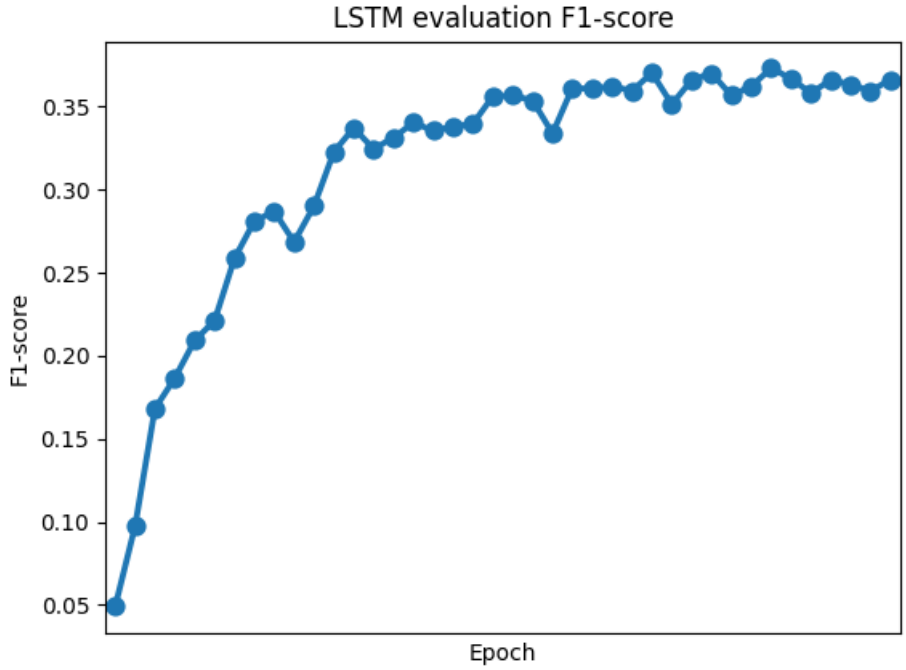
Prilog 3.8. *Grafikon funkcije gubitka klasifikatora s LSTM slojevima*



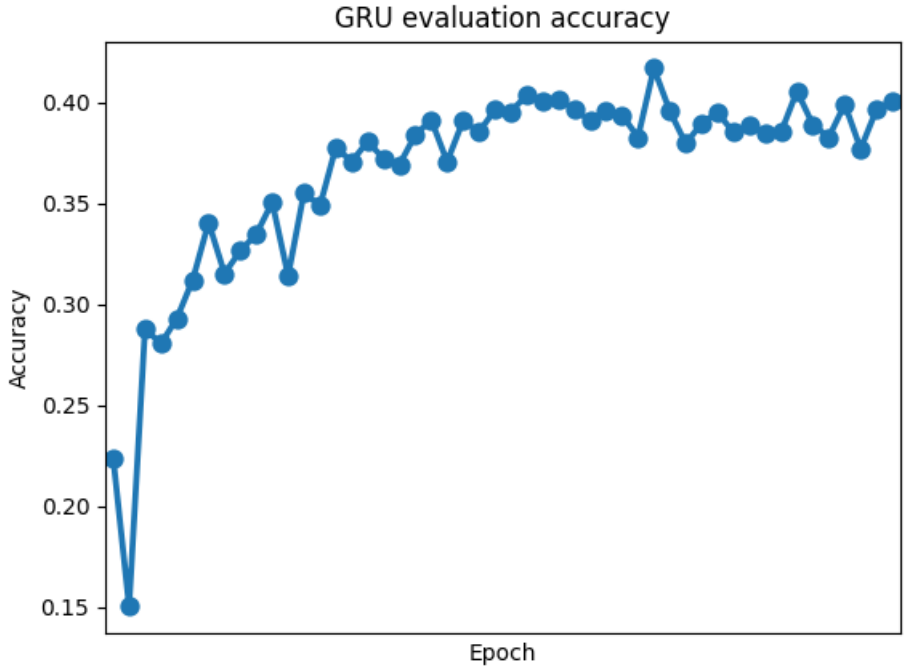
Prilog 3.9. *Grafikon preciznosti klasifikatora s LSTM slojevima*



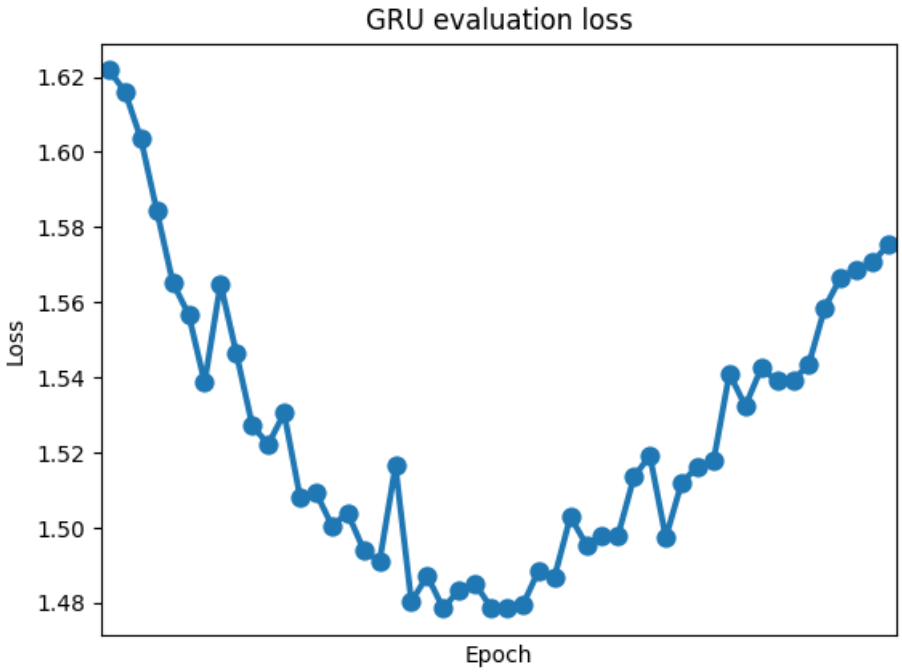
Prilog 3.10. *Grafikon odziva klasifikatora s LSTM slojevima*



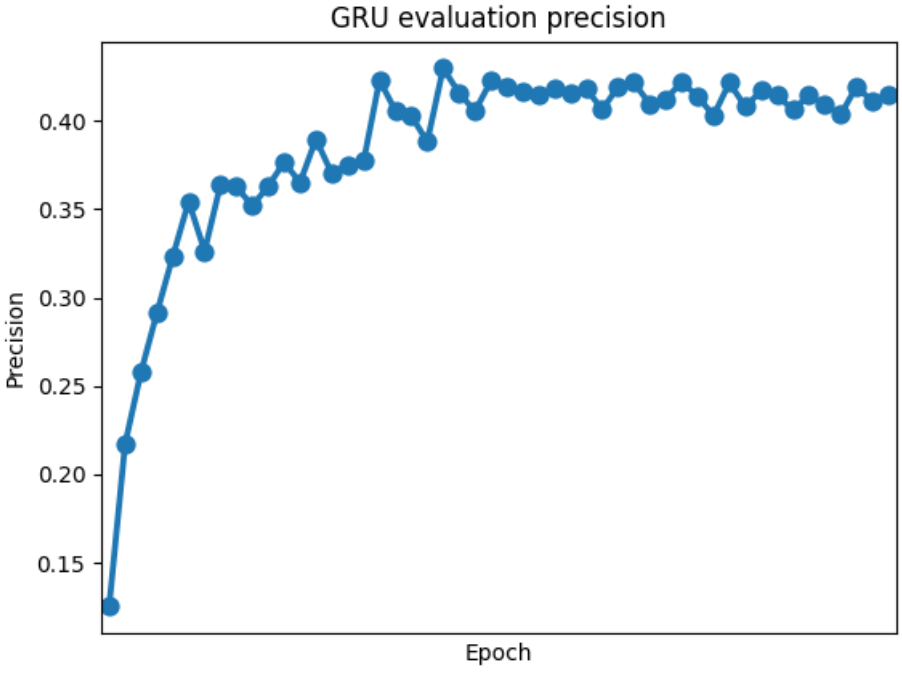
Prilog 3.11. *Grafikon F1-score klasifikatora s LSTM slojevima*



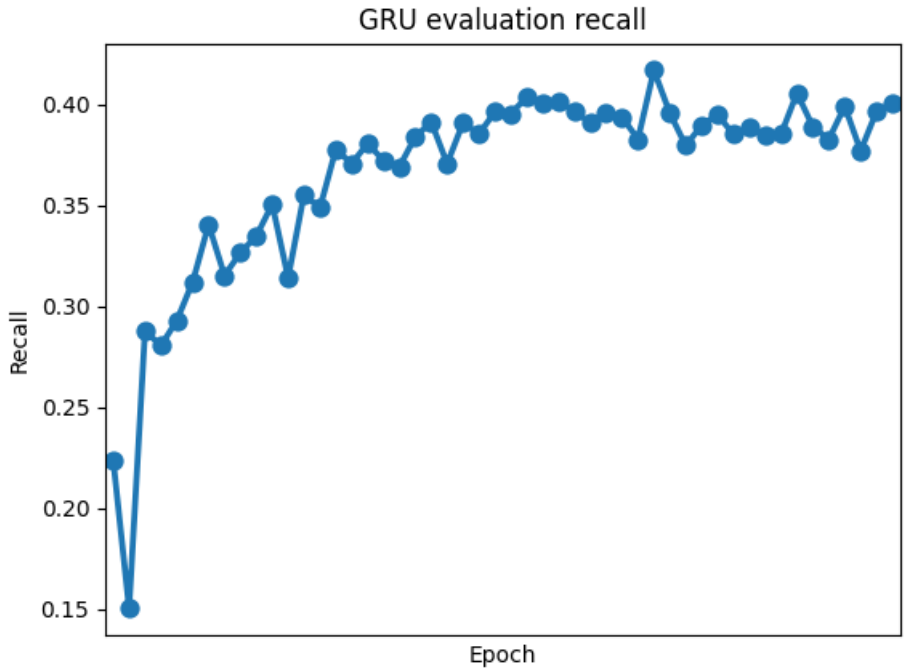
Prilog 3.12. *Grafikon točnosti klasifikatora s GRU slojevima*



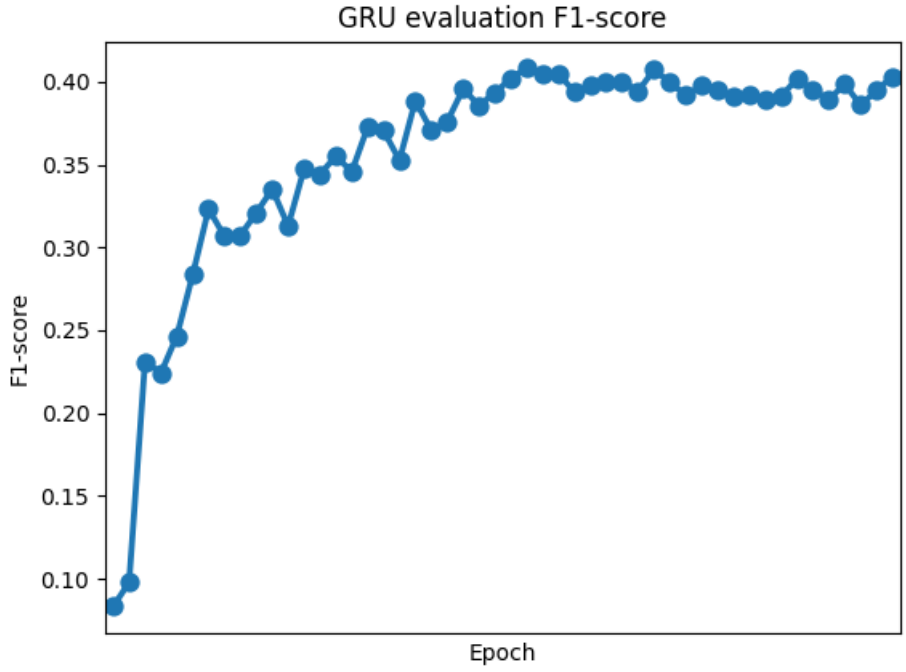
Prilog 3.13. *Grafikon funkcije gubitka klasifikatora s GRU slojevima*



Prilog 3.14. *Grafikon preciznosti klasifikatora s GRU slojevima*



Prilog 3.15. *Grafikon odziva klasifikatora s GRU slojevima*



Prilog 3.16. *Grafikon F1-score klasifikatora s GRU slojevima*