SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 2815

AUTOMATSKA DETEKCIJA DEMENCIJE IZ GOVORA KORISTEĆI TRANSFORMERSKE MODELE

Lovro Matošević

Zagreb, lipanj 2022.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 2815

AUTOMATSKA DETEKCIJA DEMENCIJE IZ GOVORA KORISTEĆI TRANSFORMERSKE MODELE

Lovro Matošević

Zagreb, lipanj 2022.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

Zagreb, 11. ožujka 2022.

DIPLOMSKI ZADATAK br. 2815

Pristupnik:	**Lovro Matošević (0036507745)**
Studij:	Računarstvo
Profil:	Računarska znanost
Mentor:	izv. prof. dr. sc. Alan Jović

Zadatak: **Automatska detekcija demencije iz govora koristeći transformerske modele**

Opis zadatka:

Demencija, bolest koja je najčešće kronične ili progresivne prirode, jedan je od glavnih uzroka invalidnosti među starijom populacijom. Rana, automatska detekcija demencije težak je zadatak i može uključivati analizu lingvističkih značajki tekstualnih transkripti ili akustičkih značajki govora. U ovom diplomskom radu istražit će se mogućnosti i ograničenja rane detekcije demencije koristeći duboko učenje. Koristit će se skup podataka Pitt korpus koji se nalazu u okviru dijeljene baze podataka DementiaBank, koja je napravljena u svrhu istraživanja kognitivnih poremećaja, a posebice demencije. Podatke koji su dostupni u audio i tekstnom formatu potrebno je najprije predobraditi. S obzirom da je motivacija ovog rada izrada računalnog alata za ranu detekciju demencije koji bi medicinskim stručnjacima služio kao pomoć pri dijagnozi, ispitat će se kvaliteta dostupnih modela za automatsko prepoznavanje govora koji se koriste za transkripciju govora. Predobrađeni podaci koristit će se kao ulaz u odabrane modele dubokog učenja. Prvenstveno, ispititat će se kvaliteta korištenja aktualnih transformerskih modela, kao što je RoBERTa te će se rezultati usporediti sa srodnim istraživanjima. Konačno, u radu će se napraviti analiza čestih pogrešaka modela.

Rok za predaju rada: 27. lipnja 2022.

Za početak, želim se od srca zahvaliti svom mentoru, izv. prof. dr. sc. Alanu Joviću prije svega za njegovo stručno vodstvo prilikom pisanja ovog diplomskog rada, a zatim i za ukazano povjerenje te za to što je imao strpljenja čitati moje podugačke mailove. Od srca se zahvaljujem i svojoj obitelji, prije svega majci Višnji i ocu Antunu, za bezuvjetnu podršku koju su mi pružali, kako tijekom faksa, tako i tijekom života.

SADRŽAJ

1. Uvod

Demencija je krovni pojam za skupinu simptoma uzrokovanih poremećajem rada mozga. Posljedica demencije gubitak je, odnosno slabljenje kognitivnih moždanih funkcija – razmišljanja, pamćenja i rasuđivanja – do te mjere da utječe na svakodnevni život i aktivnosti oboljele osobe [8]. Ozbiljna je bolest koja je najčešće kronične ili progresivne naravi te je jedan od glavnih uzroka invalidnosti među starijom svjetskom populacijom. Predstavlja značajnu prepreku ne samo oboljelima, već i njihovim obiteljima i karijerama. Više od 50 milijuna ljudi u svijetu pati od demencije, a procijenjeno je da će se broj slučajeva utrostručiti do 2050. godine [32].

Osim što je svjetski zdravstveni problem od kritične važnosti, demencija je i značajan teret za svjetsku ekonomiju. Prema nekim procjenama, ekonomski učinak demencije iznosi više od 1 bilijuna dolara, što iznosi nešto više od 1% ukupnog svjetskog BDP-a. Također, neke procjene ukazuju na to da se za medicinsku i socijalnu skrb za demenciju troši više novaca nego na srčane bolesti, moždane udare i bolesti raka zajedno.

Trenutno ne postoji lijek za demenciju kao ni standardizirani test za detekciju demencije. Medicinskim stručnjacima diljem svijeta glavni je cilj dijagnosticirati demenciju što ranije kako bi se oboljelima mogla pružiti pomoć na vrijeme i tako im što značajnije ublažiti simptome i usporiti propadanje kognitivnih funkcija. Na taj način oboljelima se omogućuje kvalitetniji život i produljuje im se životni vijek. Sve od navedenog predstavlja motivaciju za razvoj učinkovitog alata za detekciju demencije koji bi se, u idealnom slučaju, mogao koristiti zajedno sa nekom vrstom standardiziranog testa.

U ovom radu istražit će se mogućnosti i ograničenja rane detekcije demencije koristeći duboko učenje, konkretnije koristeći modele koji koriste mehanizam pozornosti, kao što su BERT, RoBERTa, XLNet te Audio Spectrogram Transformer. Koristit će se Pittov korpus [5], skup podataka dostupan u okviru dijeljene baze podataka DementiaBank, napravljene u svrhu istraživanja kognitivnih poremećaja, a posebice demencije. S obzirom na to da je jedna od važnijih motivacija ovog rada izrada računalnog alata

za ranu detekciju demencije koja bi medicinskim stručnjacima služila kao pomoć pri dijagnozi, ispitat će se i kvaliteta dostupnih modela za automatsko prepoznavanje govora te će se tako transkribirani govor koristiti kao ulaz u odabrane modele dubokog učenja.

2. Demencija

Kao što je već spomenuto u uvodu, demencija je krovni pojam, odnosno zajednički naziv za skupinu simptoma uzrokovanih poremećajem moždane aktivnosti [1]. Vrlo je važno za naglasiti da se ne radi o jednoj, specifičnoj bolesti. Demencija prije svega utječe na ljudsko razmišljanje, ponašanje i znatno umanjuje sposobnost obavljanja svakodnevnih zadataka. Kod demencije, rad mozga oslabljen je do te razine da utječe na intelektualni i društveni život oboljele osobe.

Liječnici najčešće oboljelima dijagnosticiraju demenciju ukoliko su značajno umanjene dvije ili više kognitivnih funkcija [1]. Kognitivne funkcije koje su oštećene mogu uključivati pamćenje, razumijevanje informacija, govorne sposobnosti, snalaženje u prostoru, koncentraciju i prosuđivanje. Osim što oboljeli od demencije mogu imati poteškoća u rješavanju problema, nerijetko se događaju i promjene vlastite osobnosti. Konkretni simptomi kod onih koji boluju od demencije ovise o tome u kojim dijelovima mozga se javljaju oštećenja uzrokovana jednim od mogućih oboljenja koje uzrokuju demenciju, a o kojima je nešto više rečeno u nastavku ovog poglavlja. Različite vrste živčanih stanica, ovisno o kojem se obliku demencije radi, prestaju funkcionirati, gube vezu s drugim srodnim stanicama i odumiru. Demencija je najčešće progresivne prirode, što znači da simptomi demencije variraju u ozbiljnosti od najblažeg stadija bolesti gdje su simptomi takvi da tek blago počinju utjecati na život bolesnika, pa sve do najozbiljnijeg stadija kada bolesnici potpuno ovise o drugima za sve osnovne životne aktivnosti. Progresivna priroda demencije posljedica je širenja oboljenja na druge dijelove mozga.

Osim same demencije, za kontekst ovog rada važno je spomenuti i blaga kognitivna oštećenja (engl. *mild cognitive impairment, MCI*) [46]. Blago kognitivno oštećenje definirano je lošijim učinkom od normalnog na objektivnih neuropsihološkim kognitivnim testovima, ali u slučaju kada dijagnosticirana osoba može održavati normalne dnevne funkcije, primjerice održavanje mogućnosti za izršavanjem društvenih zadaćao kao što su dnevne aktivnosti na poslu ili u vlastitome domu [1, 46]. Blago kognitivno oštećenje može se svrstati kao "amnestičko", u kojem slučaju je smanjen učinak neke

od funkcija iz domene pamćenja, ili se može svrstati kao "neamnestičko", u kojem slučaju je smanjen učinak u nekoj domeni koja nije pamćenje, primjerice govor.

2.1. Faktori nastupa demencije

Za početak važno je napomenuti da od demencije može oboljeti bilo tko. Rizik za oboljenje od demencije se povećava s godinama. Iako su većina oboljelih od demencije starije životne dobi, većina starijih osoba od nje ne obolijeva. Također, demencija nije normalan dio starenja, već je ona posljedica oboljenja mozga. Događa se i da osobe mlađe od 65 godina obolijevaju od demencije, a to se naziva ranim nastupom demencije.

Rizični faktori za nastup demencije različiti su za svaku osobu. Unatoč tomu, postoji nekoliko ključnih faktora koji se dijele u dvije kategorije: faktori na koje se ne može utjecati i faktori na koje se može utjecati.

Faktori na koje se ne može utjecati su godine, genetika i obiteljska povijest bolesti [1]. Starenjem, kao što je već spomenuto, rizik za nastup demencije se povećava. Također, genetika je jedan od faktora, jer nekoliko je rijetkih vrsta demencije povezano sa specifičnim genima. Nadalje, postojanje demencije u obiteljskoj povijesti bolesti povećava rizik za nastup demencije.

S druge strane, faktori na koje se može utjecati su srčano, tjelesno i mentalno zdravlje [1]. Sukladno tomu, određeni zdravstveni čimbenici i životni stil utječu na rizik nastupanja demencije. Osobe s neliječenim krvožilnim problemima, primjerice osobe s visokim krvnim tlakom, pod većim su rizikom. Bavljenje fizičkim aktivnostima uvelike smanjuje rizik od nastupanja demencije. Bavljenjem fizičkom aktivnosti povećava se dotok krvi u mozak te se stimulira rast moždanih stanica i veza između njih. Pravilna prehrana i pravilni ritam spavanja također smanjuju rizik. Konačno, nezanemariv je utjecaj mentalnog zdravlja. Mentalna aktivnost, kao i fizička, potiče nastanak novih moždanih stanica i jača veze između njih. Depresija je, između ostalog, još jedan od čimbenika povezanih s demencijom. Određena istraživanja pokazala su pozitivnu korelaciju između depresije i demencije.

2.2. Uzroci demencije

Nastup demencije povezan je s mnoštvom različitih oboljenja. Razlozi za obolijevanje od demencije, u većini slučajeva, nisu poznati. U nastavku je predstavljeno nekoliko

najčešćih oblika demencije.

Najčešći oblik demencije koji čini čak dvije trećine svih slučajeva demencije je demencija uzrokovana Alzheimerovom bolesti, kraće nazivano Alzheimerova demencija. Osim što je najčešći oblik demencije, Alzheimerova demencija najrelevenantnija je vrsta demencije za ovaj rad. Naime, gubitak govornih sposobnosti nije karakterističan za sve vrste demencija. Dakle, upravo zbog toga što je gubitak kognitivnih funkcija koje utječu na govor jedna od karakterističnih stvari kod Alzheimerove demencije, ta vrsta demencije izrazito je važna za ovaj rad. Alzheimerova bolest uzrokuje postupno gubljenje kognitivnih funkcija. Karakteriziraju je moždana atrofija uzrokavana dvjema abnormalnostima u mozgu: amiloidnim pločicama i neurofibrilarnim zapetljanjima [1]. Nastup Alzheimerove demencije je spor i postupan. Bolest postupno napreduje tijekom nekoliko mjeseci ili godina.

Krvožilna demencija drugi je najčešći oblik demencije. To je kognitivno oštećenje uzrokovano oštećenim krvnim žilama u mozgu. Potencijalnih uzroka oštećenja krvnih žila u mozgu ima mnogo: mali, često cistični kronični moždani udari, višestruki mikro moždani udari, veliki moždani udari u kojima je uključeno unutarmoždano krvarenje, ateroskleroza, glioza mozga, fokalna atrofija mozga i još neki rijeđi [1]. Simptomi ove vrste demencije mogu se pojaviti iznenadno, nakon moždanog udara. Osim iznenadno, mogu se i pojavljivati postupno. Simptomi se uvelike razilikuju, ovisno o veličini, vrsti i mjestu moždanog oštećenja. Krvožilna demencija u nekim slučajevima može biti slična Alzheimerovoj demencije, a nažalost je prilično česta i kombinacija Alzheimerove bolesti i krvožilne demencije.

Oboljenje Lewyjevim tjelešcima još je jedan od oblika demencija. Karakterizira ga atrofija mozga, često generaliziranog oblika, a koja je uzrokovana prisutnošću Lewyjevih tjelešaca [1]. Lewyjeva tjelešca abnormalne su nakupine bjelančevine alfa-sinukleina. Lewyjeva tjelešca stvaraju se u živčanim stanicama, a uzrokuju promjene u kretanju, razmišljanju i ponašanju. Oboljeli mogu iskusiti jake fluktuacije u razmišljanju i koncentraciji koje unutar kratkih razdoblja variraju od normalne sposobnosti funkcioniranja do ozbiljne smetenosti. Također, oboljeli mogu doživjeti vizualne halucinacije. Postoje tri poremećaja koji se međusobno preklapaju, a mogu spadati u oboljenje s Lewyjevim tjelešcima. Ta tri poremećaja su demencija s Lewyjevim tjelešcima, Parkinsonova bolest te demencija uzrokovana Parkinsonovom bolesti. Nastup oboljenja Lewyjevim tjelešcima često je spor i postupan, a traje mjesecima ili godinama. Također, ova bolest ponekad se pojavljuje zajedno s Alzheimerovom bolesti i sa krvožilnom demencijom.

Frontotemporalna demencija vrsta je demencije koja se obično pojavljuje u pede-

setim ili šezdesetim godinama. Urokovano je fokalnom atrofijom mozga koja utječe na frontalni i/ili temporalni režanj mozga [50, 1]. Frontotemporalna demencija javlja se u dva glavna oblika, a to su frontalni, čiji su simptomi promjene u ponašanju i osobnosti, te temporalni, čiji su simptomi najčešće problemi u izražavanju. Nažalost, spomenuta dva oblika često se preklapaju. Frontalni režnjevi mozga odgovorni su za sposobnost zaključivanja i ponašanja u društvenim situacijama. Sukladno tomu, oboljeli često imaju problema s održavanjem određene razine društveno prihvatljivog ponašanja, a to se očituje u nepristojnom ponašanju, zanemarivanju odgovornosti, agresivnosti, impulzivnosti i u manjku inhibicije. Temporalna demencija dijeli se u dva glavna oblika, a to su semantička demencija, koja uključuje postupni gubitak razumijevanja riječi, pamćenja imena, probleme u pronalaženju riječi i opće probleme razumijevanja jezika, te na fluentnu afaziju koja utječe na sposobnost tečnog govora i tečnog izražavanja [50].

Demencija također može biti mješovita. Mješovita demencija vrsta je demencije u kojoj se mozak mijenja pod utjecajem više od jednog uzročnika demencije. Iako su u ovom potpoglavlju predstavljeni najčešći oblici demencije, odnosno oboljenja koje uzrokuju demenciju, postoje i drugi, rijeđi oblici koji neće biti spomenuti. Znatiželjnog čitatelja upućuje se da istraži dostupnu literaturu na internetu [1, 50] koje, srećom, na temu demencije ima sve više i više.

2.3. Utjecaj demencije na govor

Postoje značajni dokazi da demencija, specifično demencija uzrokovana Alzheimerovom bolesti, utječe na govor oboljelih. Pacijenti koji boluju od Alzheimerove bolesti imaju značajno manje rezultate u područjima verbalnog izražavanja, slušnog razumijevanja, ponavljanja izrečenog, čitanja i pisanja u odnosu na kontrolne grupe [15]. Szatloczki et al. [41] povezuju tempo govora, stanke u govoru i količinu izrečenog sa ranim stadijima demencije. Iz svega od navedenog može se zaključiti da je govor jedna od stvari čijom se analizom može jasno dijagnosticirati demencija.

Oklijevanja u govoru, periodi tišine i poštapalice (engl. *filler words*) kao što su „uh" i „um" češće se pojavljuju u govoru osoba koje boluju od demencije. Prema Vrljić [43], poštapalice su značenjski prazne riječi koje se u jeziku rabe bez stvarne potrebe i jedina im je uloga pružiti govorniku koji slobodni trenutak za pronalaženje odgovarajuće riječi ili misli kojom će nastaviti svoj govor. Dakle, ima smisla da osobe oboljele od demencije češće koriste poštapalice upravo zbog toga što su im oštećene kognitivne funkcije pa često zaboravljaju detalje i kontekst razgovora. Khodabakhsh et al. [26] su evaluirali lingvističke i prozodijske značajke za detekciju demencije iz

govora. Zaključili su da su prozodijske značajke superiornije nad lingvističkima za slučaj detekcije. Konkretnije, njihov zaključak bio je da su značajke poput omjera tišine i govora (engl. silence to utterance ratio), prosječni broj riječi, stopa izgovaranja riječi (engl. word rate) i stopa izgovaranja poštapalica (engl. filler word rate) vrlo korisne za klasifikaciju.

Dakle, analiza govora pacijenata potencijalno bi mogla dovesti do stvaranja moćnog, efektivnog dijagnostičkog alata. Jedan od testova kojem je cilj iz govora pacijenata detektirati simptome demencije uzrokovane Alzheimerovom bolesti je bostonski test slike krađe kolačića, detaljnije opisan u sljedećem poglavlju. Zaključci predstavljeni u ovom potpoglavlju predstavljaju jednu od temeljnih motivacija za odabir pristupa detekciji demencije koji je odabran u ovom radu.

3. Skup podataka

Jedan od najvećih izazova pri učenju modela koji bi mogao rano detektirati demenciju iz govora je nedostatak velikog, adekvatnog skupa podataka. U trenutku pisanja ovog rada, najveći dostupni skup podataka je Pittov korpus, dostupan u sklopu dijeljenje baze podataka DementiaBank [5].

3.1. Pittov korpus

Pittov korpus najčešće je korišten skup podataka za detekciju demencije i slične zadatke vezane za demenciju. Uzorci sadržani u Pittovom korpusu dobiveni su u sklopu velikog longitudinalnog istraživanja Alzheimerove demencije koje je održano između 1983. i 1988. godine. Kako bi sudjelovali u istraživanju, sudionici su morali imati ispunjene određene kriterije. Nisu smjeli imati nikakve prethodne kognitivne bolesti, odnosno kognitivna oštećenja. Također, nisu smjeli uzimati nikakve lijekove koji mogu utjecati na središnji živčani sustav.

Ukupno je 292 sudionika sudjelovalo u istraživanju. Od tih 292 sudionika, 98 sudionika je svrstano u kontrolnu grupu, a ostalih 194 sudionika svrstano je u dementnu grupu. Konkretnije, svakome od spomenutih 194 ispitanika iz dementne grupe klasificirano je postojanje jednog od sljedećeg: blago kognitivno oštećenje, vjerojatna demencija ili definitivna demencija.

Svakom je od ispitanika dan test slike krađe kolačića. Test slike krađe kolačića svaki ispitanik proveo je između jednog i tri puta. Spomenuti test detaljnije je objašnjen u sljedećem potpoglavlju. Skup podataka Pittov korpus sadrži zvučne zapise svakog od prethodno spomenutih ispitivanja. Svaki zvučni zapis, osim govora ispitanika, sadrži i govor ispitivača, odnosno medicinskog stručnjaka koji je obavljao ispitivanje. Osim zvučnih zapisa govora, u skupu podataka su sadržani i tekstni zapisi izvršenih ispitivanja koje su napravili stručni lingvisti. Ti tekstni zapisi, u daljnjem tekstu transkripti, zapisani su u posebnom formatu koji se zove CHAT format [31]. CHAT je format korišten u svim skupovima podataka dijeljene baze podataka DementiaBank.

CHAT format ističe se po tome što, osim transkribiranog govora sudionika, sadrži i niz dodatnih informacija kao što su gramatičke značajke riječi i gramatički odnosi između riječi te određene morfofonološke informacije.

Pittov korpus sastoji se od audiozapisa ispitanika na testu slike krađe kolačića i pripadnih transkripata napravljenih od strane ekspertnih lingvista. Kao što je i česta praksa u radovima na temu detekcije demencije, u ovom radu odlučeno je ukloniti sve ispitanike kojima su dijagnosticirani blagi kognitivni poremećaji, odnosno MCI. Nakon uklanjanja ispitanika s blagim kognitivnim poremećajima, u skupu podataka ostalo je ukupno 509 audiozapisa i njihovih pripadnih lingvističkih transkripata. Detaljnija statistika o skupu podataka, nakon uklanjanja ispitanika s blagim kognitivnim poremećajima, prikazana je u tablici 3.1.

Tablica 3.1: Statistika skupa podataka

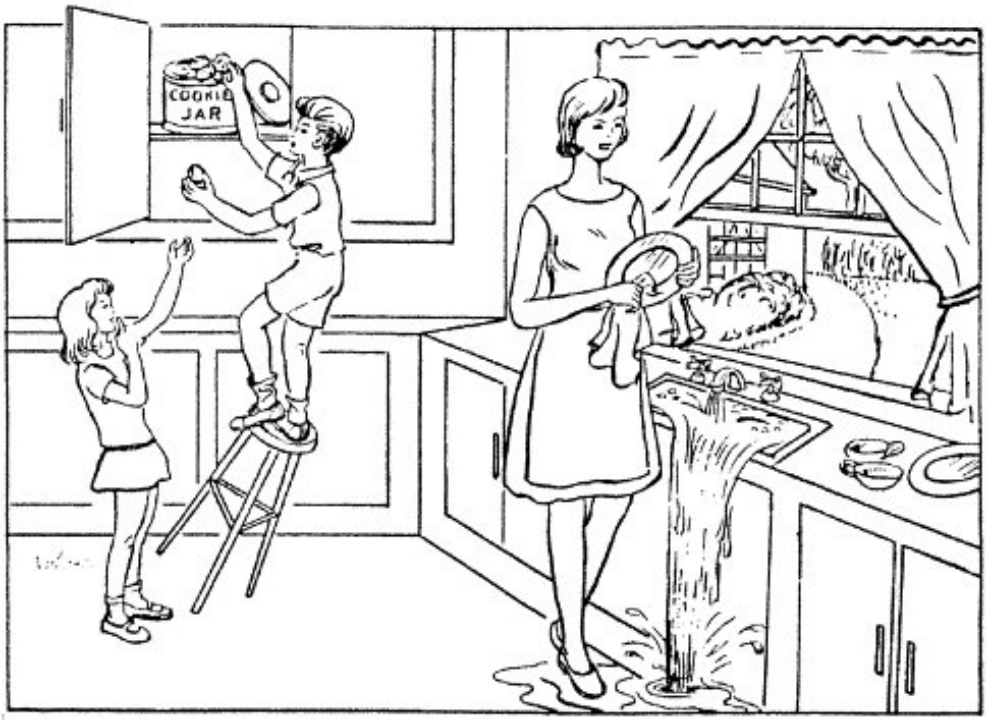
Broj ispitanika	Broj transkripata	Prosječni broj transkripata po pacijentu	Prosječni broj znakova po transkriptu	Prosječni broj riječi po transkriptu
275	509	1.85	521.42	107.09

3.2. Test slike krađe kolačića

Test slike kolačića predstavlja jedan od pokušaja da se napravi standardizirani test koji bi mogao otkriti rane indikatore nastupanja demencije. Konkretnije, cilj testa slike krađe kolačića jest procijeniti govorne i kognitivne sposobnosti ispitanika. Ispitanici koji su podvrgnuti ovom testu imaju zadatak opisati sve što vide na slici, koja se može vidjeti na slici 3.1. Slika prikazuje majku koja pere suđe dok dvoje djece pokušava ukrasti kolačiće iz tegle s kolačićima, shodno čemu je i slika dobila svoj upečatljivi naziv.

Slika također sadrži elemente i informacije iz različitih semantičkih kategorija. Zdravi ispitanici mogu percipirati i opaziti svaki aspekt slike te ju isto tako logički, konzistentno i koherentno opisati stručnjaku koji provodi ispitivanje.. Ispitanici s neurološkim oštećenjima mogu pokazati manjkavosti u izvršnim neurološkim funkcijama, odnosno ispitivač može iz njihovih opisa slike primijetiti ugrožene kognitivne vještine koje uključuju pozornost, pamćenje i planiranje. Ovakvi ispitanici se možda

neće sjetiti da su već opisali određeni dio scene prikazan na slici pa će taj dio opisati po drugi ili treći put. Kao posljedica, njihov opis sadrži ponovljeni govor koji ne pridonosi nove informacije. Ako su ugrožene kognitivne vještine kao što su planiranje i organizacija, ispitanici s demencijom često neće moći prenijeti informacije u smislenom poretku ili dati koherentni opis scene prikazane na slici. Kao posljedica toga, njihov opis može izgledati fragmentirano i neorganizirano [10].



Slika 3.1: Cookie Theft picture, preuzeto iz Goodglass et al. [21]

4. Srodni radovi

Postoje brojni radovi na temu detekcije demencije. Iako većina njih ima određene sličnosti, primjerice većina ih koristi podatke s testa slike krađe kolačića iz Pittova korpusa, svi se razlikuju na neki način. U nastavku ovog poglavlja predstavljeni su dosadašnji različiti pristupi detekciji demencije i neki od njihovih rezultata.

4.1. Pristupi temeljeni na značajkama

Prvi pristupi detekciji demencije iz govora bili su temeljeni na značajkama [7]. Klasifikator je učen koristeći brojne lingvističke značajke kao što su stopa korištenja glagola, imenica, zamjenica i slične. Daljnja poboljšanja ovih pristupa temeljenih na značajkama temeljena su na povećanju broja ekspertno definiranih značajki [33], dodavanjem akustičkih značajki [27], povezivanjem lingvističkih značajki s neuropsihološkim testovima [16] i tako dalje. Umjesto da koriste značajke uvedene od strane ljudskih eksperata, neka istraživanja za detekciju demencije fokusirala su se na grupiranje (engl. *clustering*) prednaučenih vektorskih reprezentacija (engl. *embedding*) riječi ispitanika metodom GloVe [48].

4.2. Pristupi temeljeni na dubokom učenju

U zadnjih nekoliko godina povećao se broj istraživačkih radova koji koriste pristupe temeljene na dubokom učenju umjesto klasičnih, onih temeljenih na značajkama. Prvi od radova u kojem je korišteno duboko učenje je onaj od Orimaye et al. [34], u kojem je duboka neuronska mreža učena za zadatak predviđanja blagih kognitivnih poremećaja iz govora.

U nekolicini radova koji su uslijedili, uglavnom su korištene konvolucijske neuronske mreže (engl. convolutional neural network, CNN) [25] i povratne neuronske mreže s dugim kratkoročnim pamćenjem (engl. *long short-term memory, LSTM*) [17]. Prateći nedavni uspjeh transformerskih modela, objavljeno je nekoliko istraživanja koji koriste

transformere za detekciju demencije. Većina ih koristi modele BERT ili RoBERTa. Transformerska arhitektura, uključujući BERT i RoBERTu je detaljnije objašnjena u poglavlju o modelima.

U tablici 4.1 prikazana je usporedba nekih od važnijih srodnih radova i njihovih rezultata. Također, u tablici su sadržane informacije o tome koji je skup podataka korišten, koja je metoda modeliranja i koja je evaluacijska metoda korištena u svakom od radova. Konačno, točnost i mjera F1 su prikazani za svaki od radova kod kojih je ta informacija bila dostupna.

Važno je za napomenuti da bi se svi rezultati prikazani u tablici trebali uzeti sa zrncem soli. Naime, teško je izvršiti izravnu usporedbu rezultata radova prikazanih u tablici. Prije svega, gotovo ni u jednom od spomenutih radova nije opisan način na koji su transkripti dostupni u Pittovom korpusu predobrađeni, iako postoje neke iznimke. Preciznije, Yancheva i Rudzicz [48] jasno su naznačili da su iz transkripata uklonili sva pojavljivanja poštapalica i ponovljenih dijelova govora. Karlekar et al. [25] podijelili su sve transkripte u pojedinačne iskaze te su uklonili sve iskaze bez oznaka o rečeničnim dijelovima (engl. part of speech, POS). Ilias i Askounis uklonili su sve fonološke fragmente i nestandardne oblike iz transkripata, uključujući i poštapalice i ponovljene dijelove govora [24]. U svim ostalim radovima prikazanima u tablici ne spominje se izričito na koji način su podaci pripremljeni.

Nadalje, u većini srodnih radova nije pobliže objašnjeno je li skup podataka podijeljen prema pacijentima ili prema uzorcima. Podjelom prema uzorcima događa se curenje podataka zato što je u tom slučaju moguće da isti ispitanik završi i u skupu za učenje i u skupu za testiranje. Dodatno, kao što je očevidno, nije korištena ista evaluacijska metoda u svim radovima što znatno može utjecati na vjerodostojnost rezultata. Konačno, u nekim radovima izbačeni su uzorci s ispitanicima koji su imali blage kognitivne poremećaje, dok su u ostatku radova takvi uzorci uključeni.

Sve navedeno sugerira nemogućnost izravne usporedbe rezultata srodnih radova

Tablica 4.1: Usporedba srodnih radova za detekciju demencije

Skup podataka	Model	Validacija	Točnost, %	F1, %	Referenca
ADReSS Challenge [30]	BERT	Stratified 10-fold CV	85.56%	85.43%	Ilias & Askounis [24]
Pittov korpus	CNN-RNN	-	84.9%	-	Karlekar et.al. [25]
Pittov korpus	LSTM	-	83.7%	-	Karlekar et.al. [25]
Pittov korpus	CNN-LSTM	Leave-One-Out CV	85.6%	-	Fritsch et.al. [17]
Pittov korpus	SVM	10-fold CV	78%	82%	Hernandez-Dominguez et.al. [23]
Pittov korpus	RoBERTa (512 tokena)	10-fold CV	86.72%	-	Jonasson, Wahlforss [2]
Pittov korpus	S-BERT Large LR	10-fold CV	88.08%	87.23%	Roshanzamir et.al. [38]
ADReSS Challenge [30]	BERT	5-fold CV	80%	74%	Saltz et.al. [39]
Pittov korpus	BERT	5-fold CV	80%	74%	Saltz et.al. [39]

5. Modeli

U ovom poglavlju pobliže su objašnjeni modeli dubokog učenja korišteni u ovom radu. Prvi dio poglavlja odnosi se na modele koji izvode klasifikaciju na tekstnim podacima. U drugom dijelu poglavlja objašnjen je odabrani model za automatsko prepoznavanje govora, odnosno model korišten za transkripciju govora u tekst. Konačno, u trećem dijelu ovog poglavlja predstavljen je odabrani model koji izvodi klasifikaciju izravno iz govora.

5.1. Modeli za klasifikaciju teksta

Prije pojave mehanizma pozornosti (engl. *attention*) i transformerske arhitekture neuronske mreže, za zadatke strojnog prevođenja korištena je struktura koder-dekoder temeljena na povratnim neuronskim mrežama (u daljnjem tekstu RNN) i mrežama s dugim kratkoročnim pamćenjem (u daljnjem tekstu LSTM). Takve mreže radile su na sljedeći način: LSTM u funkciji kodera bi prošao kroz cijeli ulazni niz i kodirao ga u kontekstni vektor, koji je zadnje skriveno stanje dijela LSTM/RNN. Sva srednja stanja kodera su zanemarena, odnosno samo je zadnje stanje korišteno kao početno skriveno stanje dekodera. Zatim, dekoderski dio, odnosno LSTM ili RNN bi stvarao riječi u rečenici, jednu za drugom. Glavni nedostatak ovog pristupa je ovisnost o koderu. Ako kkoder proizvede loš sažetak, odnosno nedovoljno dobar izlazni vektor, prijevod će također biti loš. Uistinu, u praksi je pokazano da su takvi koderi stvarali loše prijevode kad bi im se kao ulaz dale dugačke rečenice. Opisani problem naziva se problem dalekosežne ovisnosti (engl. *long-range dependency problem*) RNN-ova i LSTM-ova.

RNN-ovi ne mogu pamtiti dulje rečenice i nizove zbog problema nestajanja gradijenta (engl. *vanishing gradient*) i problema eksplozije gradijenata (engl. *exploding gradient*). Čak su i autori opisane mreže koder-dekoder demonstrirali da se učinak takve mreže znatno smanjuje ako se ulazni niz povećava [9]. Iako je upravo zbog opisanih problema RNN-ova razvijen LSTM, kod LSTM-a se javljaju novi problemi. Osim što je LSTM u određenim slučajevima prilično zaboravan, kao veći problem

se pokazala nemogućnost davanja većeg značaja određenim riječima u ulaznom nizu. Upravo zbog toga, razvijen je novi mehanizam koji se naziva pozornost i zajedno s njime je predstavljena nova arhitektura zvana transformer.

5.1.1. Pozornost

Mehanizam pozornosti stvoren je ponajprije da riješi problem uskog grla koji nastaje pri korištenju kodiranog vektora fiksne duljine, u kojem slučaju dekoder ima ograničeni pristup informacijama iz ulaza. Kao što je već spomenuto, taj problem posebno je problematičan za dulje ulazne nizove jer dimenzionalnost njihova prikaza jednaka je kao i za kraće, jednostavnije nizove [4]. Osnovni mehanizam pozornosti koji su uveli Bahdanau et al. [4] podijeljen je u tri osnovna koraka. Prvi korak je računanje ocjene poravnanja (engl. *alignment score*). Model poravnanja uzima kodirana skrivena stanja, h_i, i prethodni izlaz dekodera, s_{t-1}, kako bi izračunao ocjenu, $e_{t,i}$, koja pokazuje koliko su dobro elementi ulaznog niza poravnati s trenutnim izlazom na poziciji t. Model poravnanja predstavljen je funkcijom, $a(.)$, koju može implementirati unaprijedna neuronska mreža: $e_{t,i} = a(s_{t-1}, h_i)$. Drugi korak je računanje težina, $\alpha_{t,i}$, primjenjući funkciju softmax na prethodno izračunate ocjene poravnanja: $\alpha_{t,i} = softmax(e_{t,i})$. Treći, posljednji korak je stvaranje jedinstvenog kontekstnog vektora, c_t, koji se u svakom vremenskom koraku šalje dekoderu. Kontekstni vektor se računa kao težinska suma svih skrivenih stanja kodera: $c_t = \sum_{i=1}^{T} \alpha_{t,i} h_i$.

Bahdanau et al. su koristili RNN za koder i dekoder. Međutim, mehanizam pozornosti može se reformulirati u poopćeni oblik koji se može koristiti za bilo koji niz-u-niz zadatak.

Poopćeni mehanizam pozornosti sastoji se od tri glavne komponente, upiti (engl. *queries*), Q, ključevi (engl. *keys*), K, i vrijednosti (engl. *values*), V. U analogiji s mehanizmom pozornosti predložen od Bahdanau et al., upit je analogan s prethodnim izlazom dekodera, s_{t-1}, dok vrijednosti V odgovaraju kodiranim ulazima, h_i. U Bahdanauovoj pozornosti, ključevi K i vrijednosti V su isti vektor. Poopćeni mehanizam pozornosti izvodi se u tri koraka. U prvom koraku svaki se od vektora upita q uspoređuje s bazom ključeva i računa se izlazna vrijednost e_{q,k_i}. Operacija usporedbe računa se kao skalarni produkt pojedinog upita sa svakim vektorom ključa, k_i: $e_{q,k_i} = q \cdot k_i$. Izlazne vrijednosti zatim se predaju funkciji softmax kako bi se generirale težine: $\alpha_{q,k_i} = softmax(e_{q,k_i})$. U zadnjem koraku poopćena pozornost računa se kao težinska suma vektora vrijednosti, v_{k_i}, u kojem je svaki vektor uparen s odgovarajućim

ključem: $attention(q, K, V) = \sum_i \alpha_{q,k_i} v_{k_i}$. U kontekstu strojnog prevođenja, svakoj riječi u ulaznom nizu pridružuju se zasebni vektori upita, ključa i vrijednosti. Ti vektori generiraju se množenjem kodiranog prikaza specifične riječi koju se razmatra sa tri različite težinske matrice koje se stvaraju tijekom učenja.

Naposlijetku, nakon što su objašnjeni Bahdanauov mehanizam pozornosti i poopćeni mehanizam pozornosti, još je potrebno uvesti mehanizam pozornosti koji koriste transformerske arhitekture, mehanizam samopozornosti (engl. *self-attention*). U svojem revolucionarnom članku "Attention Is All You Need", Vaswani et al. uvode mehanizam samopozornosti u kojem se prikaz niza, odnosno rečenice, računa stavljajući u odnos različite riječi u istom nizu [42].

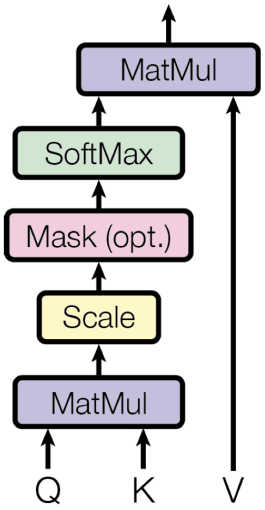
Glavne komponente ovog mehanizma pozornosti su:

- q i k, vektori dimenzija, d_k, koji sadržavaju upite i ključeve
- v, vektor dimenzija, d_v, koji sadrži vrijednosti
- Q, K i V, matrice koje u sebi sadrže odgovarajuće skupove upita, ključeva i vrijednosti
- W^Q, W^K i W^V, projekcijske matrice koje se koriste u generiranju različitih potprostora prikaza matrica upita, vrijednosti i ključeva
- W^O, projekcijska matrica za izlaz s više glava (engl. *multi-head*)

U svojoj srži, o funkciji pozornosti može se razmišljati kao o preslikavanju upita i skupa parova ključ-vrijednost u izlaz. Vaswani et al. predlažu skaliranu pozornost skalarnog produkta (engl. *scaled dot-product attention*), a potom nadograđuju tu ideju i predlažu pozornost s više glavi (engl. *multi-head attention*). Za početak, potrebno je prvo opisati skaliranu pozornost skalarnog produkta.

Skalirana pozornost skalarnog produkta prati recept poopćenog mehanizma pozornosti koji je prethodno opisan. Kao što i samo ime govori, prvo se računa skalarni produkt za svaki upit, q, sa svakim od ključeva, k. Nakon toga, svaki od rezultata dijeli se sa $\sqrt{d_k}$, te se dobiveni rezultat šalje u funkciju softmax. Na taj način dobivaju se težine koje se koriste za skaliranje vrijednosti, v. Opisani mehanizam može se vidjeti na slici 5.1. U praksi, izračuni se mogu efikasno primijeniti na cijeli skup upita istovremeno. To se postiže ubacivanjem matrica, Q, K i V u funkciju pozornosti: $attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$. Faktor skaliranja ubačen je kako bi se donekle neutralizirao rast skalarnih produkata za velike vrijednosti d_k, u kojem slučaju bi se primjenom softmaxa dobili iznimno mali gradijenti koji vode do već spomenutog

problema nestajanja gradijenata.



Slika 5.1: Skalirana pozornost skalarnog produkta, preuzeto iz "Attention Is All You Need" [42]

.

U nastavku slijedi cijeli postupak izračuna skalirane pozornosti skalarnog produkta, korak po korak:

1. Računanje ocjene poravnanja množenjem skupa upita, skupljenih u matrici Q, sa ključevima matrice K. Ako je matrica Q dimenzija $m \times d_k$, a matrica K dimenzija $n \times d_k$, tada je rezultantna matrica dimenzija $m \times n$:

$$QK^T = \begin{bmatrix} e_{11} & e_{12} & \cdots & e_{1n} \\ e_{21} & e_{12} & \cdots & e_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ e_{m1} & e_{m2} & \cdots & e_{mn} \end{bmatrix}$$

2. Skaliranje svake ocjene poravnanja sa $\frac{1}{\sqrt{d_k}}$:

$$\frac{QK^T}{\sqrt{d_k}} = \begin{bmatrix} \frac{e_{11}}{\sqrt{d_k}} & \frac{e_{12}}{\sqrt{d_k}} & \cdots & \frac{e_{1n}}{\sqrt{d_k}} \\ \frac{e_{21}}{\sqrt{d_k}} & \frac{e_{22}}{\sqrt{d_k}} & \cdots & \frac{e_{2n}}{\sqrt{d_k}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{e_{m1}}{\sqrt{d_k}} & \frac{e_{m2}}{\sqrt{d_k}} & \cdots & \frac{e_{mn}}{\sqrt{d_k}} \end{bmatrix}$$

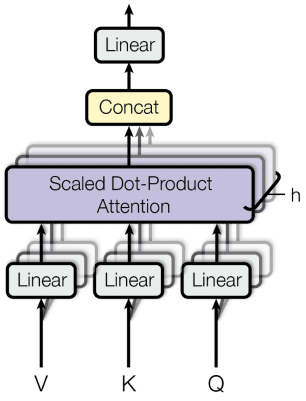
3. Primjena funkcije softmax kako bi se dobio skup težina:

$$softmax(\frac{QK^T}{\sqrt{d_k}}) = \begin{bmatrix} softmax(\frac{e_{11}}{\sqrt{d_k}} & \frac{e_{12}}{\sqrt{d_k}} & \cdots & \frac{e_{1n}}{\sqrt{d_k}}) \\ softmax(\frac{e_{21}}{\sqrt{d_k}} & \frac{e_{22}}{\sqrt{d_k}} & \cdots & \frac{e_{2n}}{\sqrt{d_k}}) \\ \vdots & \vdots & \ddots & \vdots \\ softmax(\frac{e_{m1}}{\sqrt{d_k}} & \frac{e_{m2}}{\sqrt{d_k}} & \cdots & \frac{e_{mn}}{\sqrt{d_k}}) \end{bmatrix}$$

4. Konačno, množenje rezultantnih težina iz prethodnog koraka sa matricom vrijednosti V, dimenzija $n \times d_v$:

$$softmax(\frac{QK^T}{\sqrt{d_k}}) \cdot V = \begin{bmatrix} softmax(\frac{e_{11}}{\sqrt{d_k}} & \frac{e_{12}}{\sqrt{d_k}} & \cdots & \frac{e_{1n}}{\sqrt{d_k}}) \\ softmax(\frac{e_{21}}{\sqrt{d_k}} & \frac{e_{22}}{\sqrt{d_k}} & \cdots & \frac{e_{2n}}{\sqrt{d_k}}) \\ \vdots & \vdots & \ddots & \vdots \\ softmax(\frac{e_{m1}}{\sqrt{d_k}} & \frac{e_{m2}}{\sqrt{d_k}} & \cdots & \frac{e_{mn}}{\sqrt{d_k}}) \end{bmatrix} \cdot \begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1d_v} \\ v_{21} & v_{12} & \cdots & v_{2d_v} \\ \vdots & \vdots & \ddots & \vdots \\ v_{n1} & v_{n2} & \cdots & v_{nd_v} \end{bmatrix}$$

Nadogradnja na opisani mehanizam skalirane pozornosti skalarnog produkta je mehanizam pozornosti s više glava, također predstavljena u "Attention Is All You Need". Predloženi mehanizam linearno projicira upite, ključeve i vrijednosti h puta, svaki puta koristeći različitu naučenu projekciju. Zatim se na svaku od h projekcija paralelno primjenjuje mehanizam skalirane pozornosti skalarnog produktra te se h dobivenih izlaza konkatenira i ponovno linearno projicira kako bi se dobio konačni rezultat. Opisani mehanizam može se vidjeti na slici 5.2.



Slika 5.2: Pozornost s više glavi, preuzeto iz "Attention Is All You Need" [42]

.

Osnovna ideja pozornosti s više glava je omogućiti funkciji pozornosti dohvat informacija iz različitih prikaza potprostora, što ne bi bilo moguće kada bi se koristila samo jedna "glava" pozornosti. Funkcija pozornosti s više glava može se prikazati na sljedeći način: $multihead(Q, K, V) = concat(head_1, ..., head_h)W^O$. U toj funkciji,

svaka glava $head_i$, $i = 1, ..., h$ koristi jednostruku funkciju pozornosti koja je definirana vlastitim naučenim projekcijskim matricama: $head_i = attention(QW_i^Q, KW_i^K, VW_i^V)$.

Za lakše shvaćanje kompleksnog postupka izračunavanja pozornosti s više glava, u nastavku je, korak po korak, prikazan cijeli postupak:

1. Prvi korak je računanje linearno projiciranih verzija upita, ključeva i vrijednosti kroz množenje s pripadnim težinskim matricama, W_i^Q, W_i^K i W_i^V za svaku glavu $head_i$

2. Drugi korak je primjena jednostruke funkcije pozornosti za svaku glavu. Dakle, prvo se množe matrice upita i ključeva. Zatim se primjenjuju operacije skaliranja i softmaxa. Konačno, korištenjem matrice vrijednosti, generira se izlaz za svaku glavu.

3. Treći korak je konkateniranje svih izlaza glava, $head_i$, $i = 1, ..., h$.

4. Četvrti, posljednji korak je primjena linearne projekcije nad konkateniranim izlazom kroz množenje sa matricom težina, W^O, a time se dobiva konačni rezultat.

5.1.2. Transformerska arhitektura

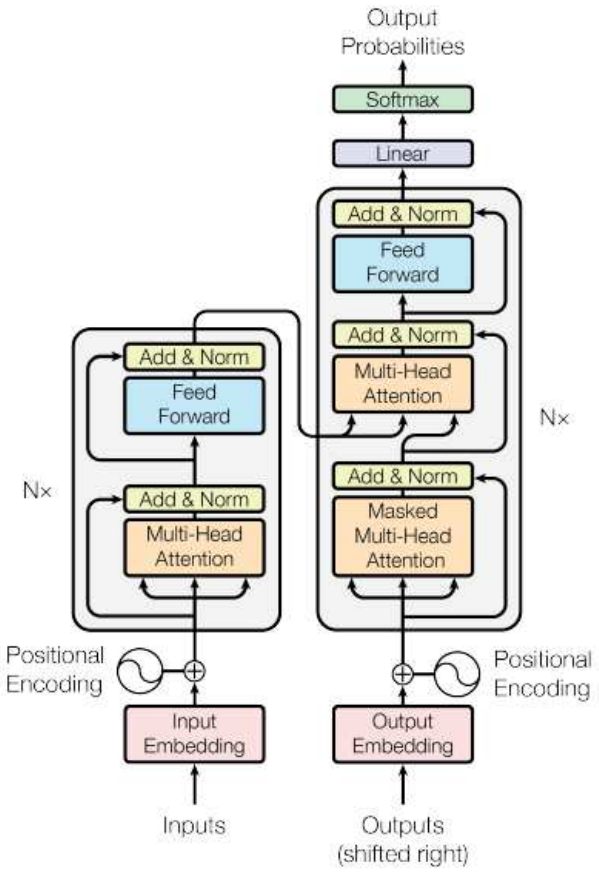
Nakon što je detaljno objašnjen mehanizam pozornosti, za razumijevanje korištenih modela u ovom radu potrebno je još opisati klasičnu arhitekturu transformera.

Mehanizam samopozornosti koji koriste transformeri invarijantan je na permutaciju. Dakle, ako se promijeni poredak riječi u ulaznoj rečenici, odnosno ulaznom nizu, izlazi će biti isti. To predstavlja problem. Kako bi se taj problem riješio, potrebno je stvoriti određeni prikaz pozicije riječi u rečenici i dodati ga u vektorske reprezentacije za riječi. Vaswani et al. uvode pozicijske vektorske reprezentacije koji su istih dimenzija kao vektorske reprezentacije za riječi, kako bi se to dvoje moglo sumirati. Dakle, primjenjuje se funkcija koja preslikava pozicije riječi u rečenici u brojčane vektore s realnim vrijednostima. U "Attention Is All You Need", Vaswani et al. koriste funkcije sinusa i kosinusa:

$$PE_{pos,2i} = sin(pos/10000^{2i/d_{model}})$$
$$PE_{pos,2i+1} = cos(pos/10000^{2i/d_{model}})$$

U prikazanim formulama pos označava poziciju riječi, d_{model} označava broj dimenzija vektorske reprezentacije, a i označava pojedinu dimenziju [42].

Arhitektura transformera prikazana je na slici 5.3. Može se primijetiti da je na lijevoj strani slike koder, dok je na desnoj strani slike dekoder.



Slika 5.3: Arhitektura transformera, preuzeto iz "Attention Is All You Need" [42]

.

Sad kada su objašnjeni glavni dijelovi modela, moguće je opisati strukturu kodera i dekodera. Osnovni zadatak kodera je preslikati ulazni niz u n-dimenzionalni vektor realnih brojeva. Prije ulaza u koder, vektorskim reprezentacijma dodaje se prethodno opisana pozicijska vektorska reprezentacija kako bi se dodala informacija o poziciji. Koder opisan u originalnom članku sastoji se od $N = 6$ identičnih slojeva, a svaki sloj sastoji se od dva podsloja, što se može vidjeti na slici 5.3 [42]. Prvi podsloj implementira već opisani mehanizam samopozornosti s više glava. Drugi podsloj je potpuno povezana unaprijedna mreža, koja se sastoji od dvije linearne transformacije s aktivacijom ReLU (engl. *Rectified Linear Unit*) između:

$$FFN(x) = ReLU(W_1 x + b_1)W_2 + b_2$$

Svaki od šest slojeva kodera primjenjuje iste linearne transformacije na sve riječi ulaznog niza, ali svaki sloj ima različite težine (W_1, W_2) i pomake (b_1, b_2). Nadalje, postoji rezidualna poveznica između svakog podsloja, odnosno između podsloja koji koristi pozornost i podsloja koji je potpuno povezana unaprijedna mreža. Spomenuta

rezidualna poveznica zbraja izlaz svakog od podslojeva s njegovim ulazom. Također, svaki podsloj popraćen je normalizacijskim slojem, $layernorm(.)$, koji normalizira izračunatu sumu ulaza u podsloj i izlaza koji je generirao podsloj, $sublayer(x)$:

$$layernorm(x + sublayer(x))$$

Izlaz zadnjeg sloja kodera šalje se u dekoder, prikazan na desnom dijelu slike 5.3. Glavni zadatak dekodera je generiranje izlaznog niza na temelju ulaza koji se prima od kodera. Dekoder ima nekoliko sličnosti s koderom. Dekoder se, kao i koder, u originalnom članku sastoji od $N = 6$ identičnih slojeva, a svaki od tih slojeva sastoji se od tri podsloja. Prvi podsloj prima prethodni izlaz iz dekodera s dodanim pozicijskim vektorskim reprezentacijama i nad njim primjenjuje samopozornost s više glava. Za razliku od kodera koji je namijenjen da radi nad svim riječima u ulaznom nizu, neovisno o njihovoj poziciji u nizu, dekoder je dizajniran na način da se posveti samo prethodnim riječima. Dakle, predviđanje za riječ na poziciji i može ovisiti samo o poznatim izlazima za riječi koje riječi na poziciji i prethode u ulaznom nizu. Za mehanizam pozornosti s više glava to se ostvaruje uvođenjem maske preko vrijednosti koje su rezultat skaliranog množenja matrica Q i K. Spomenuto maskiranje radi na način da "uguši" vrijednosti u matricama koje bi inače rezultirale nedopuštenim vezama:

$$mask(QK^T) = mask(\left(\begin{bmatrix} e_{11} & e_{12} & \cdots & e_{1n} \\ e_{21} & e_{12} & \cdots & e_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ e_{m1} & e_{m2} & \cdots & e_{mn} \end{bmatrix}\right) = \begin{bmatrix} e_{11} & -\infty & \cdots & -\infty \\ e_{21} & e_{12} & \cdots & -\infty \\ \vdots & \vdots & \ddots & \vdots \\ e_{m1} & e_{m2} & \cdots & e_{mn} \end{bmatrix}$$

Zbog maskiranja, dekoder je jednosmjeran, za razliku od kodera koji je dvosmjeran. Drugi podsloj dekodera sadrži mehanizam samopozornosti s više glava koji je sličan onome u prvom podsloju kodera. S dekoderske strane, ovaj mehanizam prima upite iz prethodnog dekoderskog podsloja, a ključeve i vrijednosti iz izlaza kodera. To dozvoljava dekoderu da se pobrine za sve riječi ulaznog niza. Treći, konačni podsloj sastoji se od potpuno povezane unaprijedne mreže koja je slična onoj s koderske strane. Također, svaki od tri podsloja dekodera ima rezidualne veze i popraćene su normalizacijskim slojem, baš kao i kod kodera.

Dakle, zaključno, transformerski model radi na sljedeći način:

1. Svaka riječ ulaznog niza pretvara se u d_{model}-dimenzionalnu vektorsku reprezentaciju riječi.

2. Svakom od ulaznih vektorskih reprezentacija dodaje se vektor koji kodira pozicije, a koji je isto dimenzije d_{model}, kako bi se u ulaz dodala informacija o poziciji.

3. Dobivene vektorske reprezenttacije šalju se u koderski blok.

4. Dekoder kao ulaz prima svoj izlaz iz prethodnog vremenskog koraka, $t - 1$.

5. Ulazu u dekoder iz prošlog koraka također se, kao kod ulaza u koder, dodaje pozicijsko kodiranje na isti način kao i kod kodera.

6. Ulaz u dekoder, poboljšan pozicijskim kodiranjem šalje se u prethodno objašnjena tri podsloja dekoderskog bloka. Maskiranje se primjenjuje u prvom podsloju kako bi se spriječilo dekoder da gleda riječi koje slijede. U drugom podsloju, dekoder prima i izlaz zadnjeg sloja kodera.

7. Naposlijetku, izlaz dekodera prolazi kroz potpuno povezani sloj, nakon kojeg slijedi sloj softmax koji služi za generiranje predikcije za sljedeću riječ u izlaznom nizu.

5.1.3. BERT

Jedan od najvećih izazova u području obrade prirodnog jezika manjak je podataka za učenje. Sveukupno, dostupno je ekstremno puno tekstnih podataka, ali za stvaranje skupova podataka za pojedini zadatak potrebno je uložiti znatnu količinu resursa za njihovo označavanje. Nažalost, kako bi dali dobre rezultate, modeli dubokog učenja zahtijevaju dostupnost ogromnih označenih skupova podataka, reda veličine u milijunima ili milijardama, ovisno o modelu i zadatku. Kako bi se taj problem riješio, znanstvenici su razvili razne tehnike za učenje modela opće namjene koristeći neoznačene tekstne podatke s interneta, kojih srećom ima vrtoglavo puno. Proces učenja takvog modela opće namjene zove se predučenje (engl. *pre-training*). Takvi prednaučeni modeli opće namjene zatim se mogu ugađati (engl. *fine-tuning*) na manjim skupovima podataka specifičnim za određeni zadatak. Najpoznatiji od takvih jezičnih modela opće namjene je BERT, koji je postigao *state-of-the-art* rezultate na nizu zadataka iz područja obrade prirodnog jezika [14].

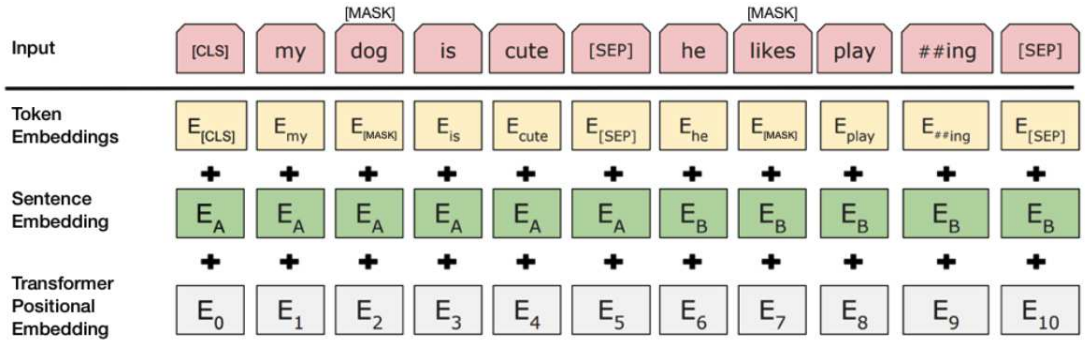
BERT, odnosno *"Bidirectional Encoder Representations from Transformers"*, uvodi nekoliko tehničkih inovacija. Jedna od ključnih tehničkih inovacija je uvođenje dvosmjernog učenja transformera u modeliranje jezika. Rezultati u članku u kojem je uveden BERT pokazuju da jezični model koji je učen dvosmjerno može imati dublji

jezični kontekst u odnosu na jezične modele učene u jednom smjeru. Pri učenju jezičnih modela, jedan od problema je definiranje zadatka predviđanja. Mnogi modeli predviđaju koja je sljedeća riječ u nizu, primjerice: *"Pas se vratio u svoju ___"*. Takav "usmjereni" pristup implicitno ograničava kontekst učenja. Kako bi riješio taj problem, BERT koristi dvije strategije učenja.

Prva od strategija je maskirano jezično modeliranje (engl. *masked language modelling*). Prije nego se nizove riječi pusti u BERT, 15% riječi u svakom nisu zamjenjuje se s tokenom [MASK]. Model zatim pokušava pogoditi koja je originalna riječ maskirana na temelju konteksta koje pružaju ostale, nemaskirane riječi u nizu. S tehničke strane, predviđanje izlaznih riječi zahtijeva dodavanje klasifikacijskog sloja nakon izlaza kodera, množenje izlaznih vektora s ugradbenom matricom čime se oni prebacuju u dimenziju vokabulara te konačno računanje vjerojatnosti za svaku riječ iz vokabulara koristeći softmax.

Druga strategija je predviđanje sljedeće rečenice (engl. *next sentence prediction*). Tijekom učenja BERT-a, model prima parove rečenica kao ulaz i uči predvidjeti je li druga od uparenih rečenica sljedbenik prve u originalnom tekstu. Naime, učenje je postavljeno tako da je kod 50% ulaznih parova druga rečenica uistinu sljedbenik prve, dok je kod ostalih 50% nasumična rečenica iz korpusa ubačena kao druga u paru. Kako bi se pomoglo modelu da razlikuje dvije rečenice u procesu učenja, ulazni nizovi su prije ulaza u model obrađeni na sljedeći način: token [CLS] je ubačen na početak prve rečenice, a token [SEP] je ubačen na kraju svake rečenice, vektorska reprezentacija rečenice koji označava radi li se o rečenici A ili o rečenici B dodaje se svakom tokenu te se konačno pozicijska vektorska reprezentacija dodaje svakom tokenu s ciljem da se jasno naznači na kojem mjestu u nizu se nalazi. Prethodno opisani ulaz u BERT može se vidjeti na slici 5.4. Model BERT se na zadacima maskiranog jezičnog modeliranja i predviđanja sljedeće rečenice uči zajedno, istovremeno, a cilj je smanjiti kombiniranu funkciju gubitka spomenuta dva zadatka.

Arhitekturno, BERT se razlikuje od običnog transformera opisanog u prethodnom potpoglavlju. BERT koristi samo koderski dio transformera. Dakle, nije dizajniran za zadatke kao što su generiranje teksta i prevođenje. U originalnom članku, autori predstavljaju dvije verzije BERT-a, $BERT_{BASE}$, sa 12 slojeva, 768 skrivenih čvorova, 12 glavi pozornosti i 110 milijuna parametara, te $BERT_{LARGE}$ sa 24 sloja, 1024 skrivenih čvorova, 16 glavi pozornosti i 340 milijuna parametara. BERT je učen na BooksCorpusu, koji sadrži oko 800 milijuna riječi, te na čitavoj engleskoj Wikipediji, koja sadrži više od 2500 milijuna riječi.



Slika 5.4: Prikaz ulaza u BERT. Ulazne vektorske reprezentacije su suma vektorske reprezentacije tokena, rečenice i pozicije. Preuzeto iz "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" [14]

.

5.1.4. RoBERTa

U svom članku, *"RoBERTa: A Robustly Optimized BERT Pretraining Approach"*, Liu et al. pokazali su da je BERT značajno podnaučen te su predložili poboljšani recept za učenje modela BERT, koji su nazvali RoBERTa [29]. U proces učenja BERT-a uvode nekoliko jednostavnih modifikacija.

Prvo, uče svoj model duže, s većim podskupovima primjeraka (engl. *batch*) za učenje i s više podataka. Skup podataka korišten za učenje RoBERTe veličine je oko 160 GB, što je znatno više od 16 GB podataka na kojima je BERT učen. Pri učenju, korištena je veličina podskupa za učenje od 8000 - u usporedbi s veličinom podskupa za učenje od 256 kod učenja BERT-a opet se radi o znatnom povećanju.

Drugo, zadatak predviđanja sljedeće rečenice u potpunosti je uklonjen kod RoBERTe, dok je zadatak maskiranog jezičnog modeliranja izmijenjen. Proces maskiranja kod RoBERTae odvija se dinamički, tijekom učenja, dok se kod BERT-a maskiranje riječi događa prije samog procesa učenja. Razlog tomu je sljedeći: svaki put kad se rečenica ubaci u podskup za učenje, ona se maskira, dakle potencijalni broj različitih maskiranih verzija iste rečenice nije ograničen kao kod BERT-a.

Konačno, autori RoBERTe odlučili su se učenje na dužim nizovima, ali važno je za napomenuti da je ograničenje od 512 tokena ostavljeno. Također, RoBERTa koristi istu arhitekturu kao i BERT. Na slici 5.5 prikazani su rezultati modela RoBERTa na raznim zadacima mjerila GLUE (engl. *GLUE benchmark*) [44]. Iz prikazanih rezultata jasno je vidljivo da je model RoBERTa postigao bolje rezultate na nizu zadataka iz područja obrade prirodnog jezika.

	MNLI	QNLI	QQP	RTE	SST	MRPC	CoLA	STS	WNLI	Avg
Single-task single models on dev										
$\text{BERT}_{\text{LARGE}}$	86.6/-	92.3	91.3	70.4	93.2	88.0	60.6	90.0	-	-
$\text{XLNet}_{\text{LARGE}}$	89.8/-	93.9	91.8	83.8	95.6	89.2	63.6	91.8	-	-
RoBERTa	**90.2/90.2**	**94.7**	**92.2**	**86.6**	**96.4**	**90.9**	**68.0**	**92.4**	**91.3**	-
Ensembles on test (from leaderboard as of July 25, 2019)										
ALICE	88.2/87.9	95.7	**90.7**	83.5	95.2	92.6	**68.6**	91.1	80.8	86.3
MT-DNN	87.9/87.4	96.0	89.9	86.3	96.5	92.7	68.4	91.1	89.0	87.6
XLNet	90.2/89.8	98.6	90.3	86.3	**96.8**	**93.0**	67.8	91.6	**90.4**	88.4
RoBERTa	**90.8/90.2**	**98.9**	90.2	**88.2**	96.7	92.3	67.8	**92.2**	89.0	**88.5**

Slika 5.5: Usporedba rezultata RoBERTe na GLUE mjerilu. Preuzeto iz "RoBERTa: A Robustly Optimized BERT Pretraining Approach" [29]

.

5.1.5. XLNet

Klasični transformerski modeli imaju jedan zamjetan nedostatak, a to je da rade sa nizovima fiksne duljine. Primjerice, što ako je modelu za predviđanje određene maskirane riječi u rečenici potrebno neko znanje, odnosno kontekst iz nekog od prethodnih nizova? Arhitektura transformer-XL riješava ovaj problem tako što dozvoljava trenutnom nizu da vidi informacije iz prethodnih nizova [11]. Upravo se na toj arhitekturi temelji XLNet [49].

Glavni doprinos XLNeta je modificirani cilj učenja jezičnog modela koji uči uvjetne razdiobe svih permutacija tokena u nizu. XLNet je autoregresijski model. Za neki niz x, autoregresijski model je onaj koji računa vjerojatnost $Pr(x_i|x_{<i})$. U kontekstu jezičnog modeliranja, to je vjerojatnost tokena x_i u rečenici, uvjetovanog tokenima $x_{<i}$ koji mu prethode. Tokene $x_{<i}$ u tom slučaju zovemo *kontekst*. Nadalje, autoregresijski modeli također mogu učiti iz odnosa pojedinog tokena i onih tokena koji slijede nakon njega. U tom slučaju cilj se može gledati kao izračun $Pr(x_i) = Pr(x_i|x_{>i})$. Ali to nisu jedini mogući ciljevi. Naime, model bi mogao naučiti nešto novo i ako se gleda odnos dva tokena najbliža onom kojeg promatramo: $Pr(x_i) = Pr(x_i|xi-1, x_{i+1})$. Na isti način može se uzeti bilo koja kombinacija tokena iz ulaznog niza. Autori XLNeta predlažu korištenje funkcije cilja koja je očekivanje svih takvih permutacija. Primjerice, rečenica $x = \{Ovo, je, recenica\}$ sa $T = 3$ tokena ima 3! mogućih permutacija: $\mathbb{Z} = [1,2,3],[1,3,2]...[3,2,1]$. Model XLNet je autoregresijski za sve takve permutacije, odnosno može izračunati vjerojatnost tokena x_i za dane tokene $x_{<i}$ za bilo koji poredak iz permutacijskog skupa. Spomenute ideje utjelovljene su u jednadžbi iz originalnog članka [49]:

$$\hat{\theta} = \underset{\theta}{argmax}[\mathbb{E}_{z\sim\mathbb{Z}}[\sum_{t=1}^{T} log[Pr(x_{z[t]}|x_{z[<t]})]]$$

Postoji još jedna stvar na koju se treba osvrnuti. Ciljana vjerojatnost ne bi trebala biti uvjetovana samo indeksima kontekstnih tokena, već i indeksom tokena čija se vjerojatnost računa. Dakle, u slučaju rečenice *"Ovo je rečenica."* ako se razmatra prva riječ, traži se $Pr(Ovo|1, je + 2)$; vjerojatnost za riječ *"Ovo"* ako je ona prvi token i ako je riječ *"je"* drugi token. Naime, arhitektura transformera implicitno kodira pozicijsku informaciju u vektorske reprezentacije riječi *"Ovo"* i *"je"*, što bi bilo prikazano kao: $(Ovo|Ovo + 1, je + 2)$. Nažalost, modelu je na takav način trivijalno znati da je riječ *"Ovo"* dio rečenice i da bi vjerojatnost trebala biti velika. Riješenje ovog problema je mehanizam samopozornosti u dva toka (engl. *two-stream self-attention*). Svaka pozicija tokena i ima dva pridružena vektora u svakom sloju m samopozornosti, vektore h_i^m i g_i^m. Vektori h pripadaju toku sadržaja (engl. *content stream*), dok vektori g pripadaju toku upita (engl. *query stream*). Pozornost toka sadržaja odgovara klasičnoj samopozornosti iz transformera, dok je pozornost toka upita uvedena da zamijeni BERT-ov token [MASK]. Vektori toka sadržaja inicijaliziraju se tako što se vektorske reprezentacije tokena dodaju pozicijskim vektorskim reprezentacijama. Vektori toka upita inicijaliziraju se na način da se generična vektorska reprezentacija w dodaje pozicijskim vektorskim reprezentacijama. Važno je za primijetiti da je vektor w isti neovisno o tokenu, te se na taj način ne može koristiti za razlikovati tokene. U svakom sloju, svaki vektor sadržaja, h_i, ažurira se koristeći sebe i ostale vektore h koji su ostali nemaskirani. Shodno tomu, u svakom sloju svaki se vektor upita g_i ažurira koristeći nemaskirane vektore sadržaja i sebe. Ažuriranje koristi g_i kao upit, a h_j kao ključeve i vrijednosti, gdje j predstavlja indeks nemaskiranog tokena u kontekstu i.

5.2. Model za automatsko prepoznavanje govora

Prije objašnjenja samog modela za automatsko prepoznavanje govora, potrebno je objasniti što je to uopće automatsko prepoznavanje govora. Automatsko prepoznavanje govora (engl. *automatic speech recognition, ASR*) jedno je od područja računske lingvistike koje se bavi prepoznavanjem i transkripcijom izgovorenog jezika u tekst. Ponekad se može čuti da se takve sustave zove "govor-u-tekst" (engl. *speech-to-text, STT*). U nastavku neće biti daljnje, dublje rasprave o samom području automatskog prepoznavanja govora, već je ideja da čitatelj bitnije koncepte shvati iz objašnjenja modela Wav2Vec2, koje slijedi u nastavku ovog potpoglavlja.

5.2.1. Wav2Vec2

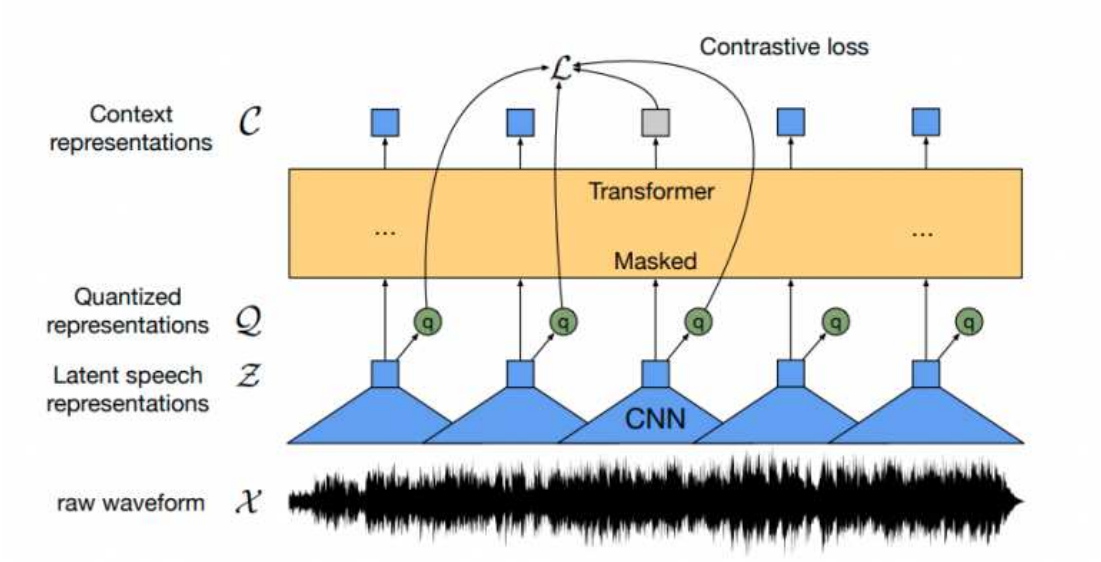
Kao što je već spomenuto u prethodnom potpoglavlju, najveći problem pri učenju modela dubokog učenja su velike količine označenih podataka koje su potrebne. Označeni podaci pogotovo su skupi u domeni prepoznavanja govora. Do nedavno je bilo potrebno imati tisuće sati transkribiranog govora kako bi se naučio model za automatsko prepoznavanje govora sa prihvatljivim performansama. Pojavom prednaučenih modela kao što je Wav2Vec2 uspjeh koji su prednaučeni modeli doživjeli u području obrade prirodnog jezika zaživio je u i u području prepoznavanja govora. Wav2Vec2, u trenutku pisanja ovog rada, jedan je od *state-of-the-art* modela za automatsko prepoznavanje govora, a svoj uspjeh duguje samonadziranom (engl. *self-supervised*) učenju [3].

Model Wav2Vec2 učen je u dvije faze. Prva faza je učenje u samonadziranom načinu koje je izvršeno koristeći neoznačene podatke. Cilj prve faze je naći najbolji mogući prikaz govora, na sličan način na koji se traže vektorske reprezentacije riječi kod modela u području obrade prirodnog jezika. Druga faza učenja je nadzirano ugađanje (engl. *supervised fine-tuning*), tijekom kojeg se koriste označeni podaci kako bi se model naučio predviđanju određenih riječi ili fonema.

Arhitektura modela prikazana je na slici 5.6. Na početku se nalazi višeslojni konvolucijski koder za značajke (engl. *multi-layer convolutional feature encoder*) $f : \mathcal{X} \mapsto \mathcal{Z}$ koji kao ulaz prima zvuk $\mathcal{X}$ i van daje latentni govorni prikaz $\mathbf{z}_1, ...\mathbf{z}_T$ za T vremenskih koraka. Ti izlazi šalju se transformeru, $g : \mathcal{Z} \mapsto \mathcal{C}$ kako bi se napravili prikazi $\mathbf{c}_1, ..., \mathbf{c}_T$ koji u sebi sadrže informacije o cijelom nizu. Izlaz kodera značajki diskretizira se u $\mathbf{q_t}$, sa kvantizacijskim modulom $\mathcal{Z} \mapsto \mathcal{Q}$ za prikaz izlaza u samonadziranom cilju.

Konkretnije, glavna ideja predučenja je slična kao kod BERT-a; dio ulaza u transformer se maskira i cilj je pogoditi maskirani latentni vektor značajki ‡␣. Međutim, autori članka poboljšali su tu jednostavnu ideju sa metodom koja se zove kontrastno učenje (engl. *contrastive learning*). Kontrastno učenje koncept je u kojem se ulaz transformira na dva različita načina. Poslije toga, model se uči za prepoznavanje pripadaju li dvije transformacije ulaza istom objektu. Za Wav2Vec2, transformerski slojevi predstavljaju prvu transformaciju, dok je druga transformacija ona dobivena kvantizacijom. Konkretnije, za maskirani latentni prikaz $\mathbf{z_t}$, cilj je dobiti takav kontekstni prikaz $\mathbf{c_t}$ za koji se točno može predvidjeti kvantizirani prikaz $\mathbf{q_t}$ između ostalih kvantiziranih prikaza.

Kvantizacija je proces pretvorbe vrijednosti iz kontinuiranog prostora u konačni



Slika 5.6: Ilustracija Wav2Vec2 koji zajednički uči kontekstualizirani prikaz govora i niz diskretiziranih govornih jedinica. Preuzeto iz "Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations" [3]

.

skup vrijednosti u diskretnom prostoru. No, kako se točno to implementira u automatskom prepoznavanju govora? Neka je $\mathbf{z_t}$ jedan latentni vektorski prikaz govora koji pokriva dva fonema. Broj fonema u bilo kojem jeziku je konačan. Štoviše, broj svih mogućih parova fonema je konačan. To znači da se oni mogu savršeno prikazati s istim prikazom. Nadalje, s obzirom da je njihov broj konačan, može se stvoriti kodni zapis koji sadrži sve moguće parove fonema. Kvantizacija se tada svodi na biranje pravog zapisa. Međutim, broj svih mogućih zvuka je ogroman. Kako bi doskočili tomu, autori članka Wav2Vec2 stvorili su G kodnih zapisa, od kojih se svaki sastoji od V kodnih riječi. Kako bi se stvorio kvantizirani prikaz, bira se najbolja riječ iz svakog od spomenutih kodnih zapisa. Zatim, odabrani vektori se konkateniraju i prolaze kroz linearnu transformaciju, a rezultat je kvantizirani prikaz. Za odabir najbolje kodne riječi iz svakog kodnog zapisa, koristi se Gumbelova varijanta funkcije softmax:

$$p_{g,v} = \frac{exp(sim(l_{g,v}+n_v)/\tau}{\sum_{k=1}^{V} exp((l_{g,k}+n_k)/\tau)}$$

gdje sim označava kosinusnu sličnost, l označava logite izračunate iz $\mathbf{z}$, n_k označava operaciju $-log(-log(u_k))$, gdje je u_k uzorak iz uniformne distribucije $U(0,1)$. Logiti su vektor nenormaliziranih predviđanja koje generira model za klasifikaciju, a koji se najčešće na kraju predaju u normalizacijsku funkciju kao što je funkcija softmax. Konačno, τ označava temperaturu. Temperatura je jedan od hiperparametara modela koji se koristi za podešavanje nasumičnosti predviđanja modela izmjenom,

odnosno skaliranjem logita prije primjene funkcije softmax. Ako je temperatura postavljena na 1, softmax se računa izravno nad neskaliranim logitima. Ako je temperatura manja od 1, primjerice 0.5, softmax se računa nad $\frac{logiti}{0.5}$, čime se dobivaju veće vrijednosti ulaza u softmax. Veće ulazne vrijednosti u funkciju softmax znače da će model biti sigurniji u svoje predviđanje, dok manje ulazne vrijednosti, konkretnije one koje se dobiju postavljanjem temperature na broj veći od 1, čine model manje sigurnim u njegova predviđanja.

Funkcija cilja je zbroj dvije funkcije gubitka: kontrastnog gubitka i gubitka različitosti (engl. *diversity loss*):

$$L = L_m + \alpha L_d$$

L_m, odnosno kontrastni gubitak je odgovoran za učenje modela za predviđanje ispravnih kvantiziranih prikaza q_t između $K + 1$ kandidata kvantiziranih prikaza $q' \in Q_t$. Spomenuti skup Q_t sastoji se od ciljanog prikaza q_t i K distraktora. Kontrastni gubitak, dakle, izgleda ovako:

$$L_m = -log \frac{exp(sim(c_t, q_t)/\kappa)}{\sum_{\widetilde{q} \in Q_t} exp(sim(c_t, \widetilde{q})/\kappa)}$$

gdje sim označava kosinusnu sličnost, a κ temperaturu koja je konstantna tijekom učenja.

Gubitak različitosti predstavlja tehniku regularizacije. Autori članka odlučili su se za $G = 2$ kodna zapisa, svaki sa $V = 320$ kodnih riječi. Dakle, to je ukupno $320 * 320 = 102400$ ukupno mogućih kvantiziranih prikaza. Međutim, moguće je da model neće iskoristiti sve te mogućnosti, već će, primjerice, koristiti samo 100 riječi iz svakog kodnog zapisa. To pobija cijeli potencijal kodnog zapisa, a upravo iz tog razloga autori su se odlučili za korištenje gubitka različitosti. Gubitak različitosti temeljen je na entropiji, čija je formula:

$$H(X) = -\sum_{x} P(x) log(P(x))$$

gdje x predstavlja mogući ishod nasumične diskretne varijable $\mathcal{X}$, a $P(x)$ vjerojatnost događaja x. Entropija je maksimalne vrijednosti kada je razdioba podataka uniformna, a u ovom slučaju to je kada su sve kodne riječi iz kodnog zapisa birane istom frekvencijom. Maksimizacijom entropije, dakle, potiče se model da iskoristi sve kodne riječi. Konačno, gubitak različitosti je:

$$L_d = \frac{1}{GV} * (-H(\overline{p}_g)) = \frac{1}{GB} \sum_{g=1}^{G} \sum_{v=1}^{V} \overline{p}_{g,v} log(\overline{p}_{g,v})$$

S obzirom da druga, odnosno faza ugađanja Wav2Vec2 ne sadrži nikakve novitete, autori nisu posvetili puno pažnje tom dijelu članka. Tijekom ove faze učenja ne koristi

se kvantizacije, već se koristi nasumično inicijalizirani linearni projekcijski sloj koji je dodan na vrh kontekstne mreže. Model se ugađa koristeći konektivističku temporalnu klasifikaciju (engl. *connectionist temporal classification, CTC*).

Rezultati prikazani u članku Wav2Vec2 su izvrsni. Model je postigao *state-of-the-art* rezultate na nizu skupova podataka. Najvažniji od tih rezultata je onaj na skupu podataka Librispeech [35]. Model Wav2Vec2 učen na svih 960 sati označenih podataka dostupnih u Librispeech korpusu postigao je stopu pogrešnih riječi (engl. *word error rate, WER*) od 1.8 na testnom skupu. Upravo taj ugođeni model korišten je za transkripciju u ovom radu. WER je najčešće korištena metrika za mjerenje točnosti modela automatskog prepoznavanja govora. Računa se na sljedeći način:

$$WER = \frac{S+D+I}{N} = \frac{S+D+I}{S+D+C}$$

, gdje S označava broj zamijenjenih riječi, D broj izbrisanih riječi, I broj umetnutih riječi i C broj točnih riječi. Dakle, neka je primjer zapis *"Ovo je diplomski rad."*, a njegova pripadna transkripcija *"Ov deep lomski rad"*. Primjer zamijenjene riječi je prva riječ, odnosno "Ovo" koje prelazi u "Ov". Primjer izbrisane riječi je riječ "je", koja se pojavljuje u originalnom zapisu, ali ne i u transkripciji. Primjer umetnutih riječi je prijelaz "diplomski" u "deep lomski".

5.3. Model za klasifikaciju govora

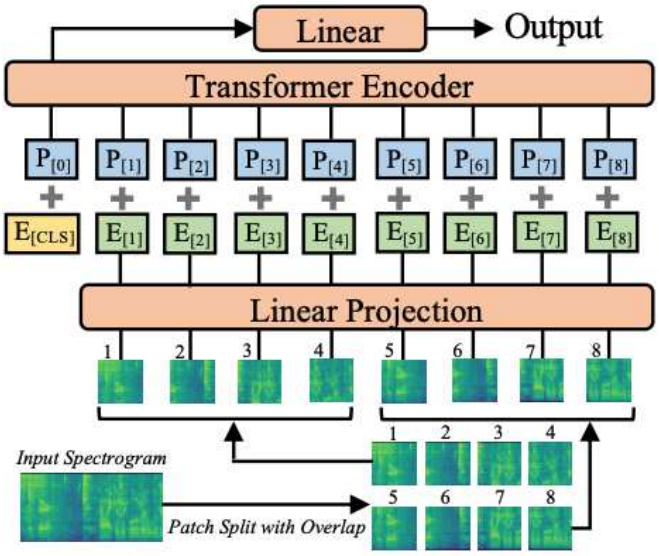
Druga grana modela korištena u ovom radu su modeli koji izravno rade klasifikaciju iz signala govora. U prošlom desetljeću, konvolucijske neuronske mreže prihvaćene su diljem cijele zajednice podatkovnih znanstvenika i inženjera kao glavni blok za izgradnju zvukovnih klasifikacijskih modela koji rade na principu "s kraja na kraj" (engl. *end-to-end*) kako bi naučili izravno preslikavanje zvukovnih spektrograma u odgovarajuće krajnje oznake. U nastavku je predstavljen model koji je korišten u ovom radu.

5.3.1. Transformer Audio Spektrograma

Kao što je već spomenuto, u prošlom desetljeću za zadatak klasifikacije zvuka ponajviše su korištene konvolucijske neuronske mreže. Kako bi se bolje uhvatio dulji globalni kontekst, jedan od novih trendova je dodavanje mehanizma samopozornosti na konvolucijsku neuronsku mrežu, čime nastaju hibridni modeli CNN-pozornost. Međutim, Yuan et al., autori članka "AST: Audio Spectrogram Transformer" [20], zapitali

su se je li uistinu potrebno oslanjanje na konvolucijske neuronske mreže ili bi možda neuronska mreža temeljena isključivo na pozornosti bila dovoljno dobra na zadacima klasifikacije zvuka.

Potaknuti znatiželjom, osmislili su transformer audio spektrograma (engl. *Audio Spectrogram Transformer*, dalje: AST), model bez konvolucija, isključivo temeljen na pozornosti koji se primjenjuje izravno na zvukovne spektrograme. Arhitektura AST-a prikazana je na slici 5.7. Na početku, ulazni zvukovni valni oblik (engl. *waveform*) od t sekundi pretvara se u niz 128-dimenzionalnih značajki "log Mel-filterbank" koji se računa sa Hammingovim prozorom širine 25 milisekundi, svakih 10 milisekundi. Time se dobiva $128 \times 100t$ spektrogram koji služi kao ulaz u AST. Nakon toga, ulazni spektrogram dijeli se u niz od N 16×16 izreza sa preklapanjem od 6 i u vremenskoj i u frekvencijskoj dimenziji. N predstavlja broj izreza, odnosno $N = 12\lceil(100t - 16)/10\rceil$. Koristeći linearni projekcijski sloj, svaki od 16×16 izreza spljošti se u 768-dimenzionalnu vektorsku reprezentaciju izreza. S obzirom na to da transformerska arhitektura ne pamti informaciju o poretku ulaza, te na to da niz izreza nije vremenski poredan, autori dodaju pozicijska vektorska reprezentacija veličine 768 svakoj vektorskoj reprezentaciji izreza kako bi omogućili modelu da nauči prostornu strukturu 2D zvukovnog spektrograma.



Slika 5.7: Arhitektura AST-a. Preuzeto iz "AST: Audio Spectrogram Transformer" [20]

.

Nadalje, slično kao kod BERT-a, na početak niza dodaje se token [CLS]. Dobiveni niz izreza šalje se transformeru. AST je dizajniran za zadatke klasifikacije, stoga autori koriste samo koderski dio transformera. Korištena je originalna arhitektura ko-

dera transformera, opisana u jednom od prethodnih potpoglavlja ovog rada. Autori kao prednosti ovakvog pristupa navode jednostavnost implementacije i reprodukcije, te lakša mogućnost za prijenosno učenje (engl. *transfer learning*) [20]. Konačno, token [CLS] nakon prolaza kroz koder transformera služi kao prikaz zvukovnog spektrograma. Taj prikaz provlači se kroz linearni sloj sa sigmoidom kao aktivacijskom funkcijom te se tako preslikava u krajnje klasifikacijske oznake.

Motivirani nedostatkom adekvatnih zvukovnih skupova podataka, te vođeni pretpostavkom da slike i zvukovni spektrogrami imaju slične formate, autori AST-a odlučili su se za prijenos znanja modela prednaučenog na ImageNetu [13]. U nastavku neće biti objašnjen način na koji su to izveli, a znatiželjnog čitatelja upućuje se da pročita originalni članak u kojem je to detaljno opisano [20].

Prema autorima, tri su glavne prednosti AST-a nad prethodnim modelima za klasifikaciju zvuka. Prvo, AST ima superiornije rezultate. Na svakom od skupova podataka na kojima je testiran, odnosno na skupu podataka AudioSet [18], skupu podataka ESC-50 [37] te skupu podataka Speech Commands [45], AST postiže *state-of-the-art* rezultate. Drugo, AST podržava ulaze varijabilne duljine i može se primijeniti na različite zadatke bez ikakve promjene u arhitekturi. Treće, u usporedbi s najboljim hibridnim modelima CNN-pozornost, AST ima jednostavniju arhitekturu s manje parametara te brže konvergira tijekom učenja.

6. Eksperimenti, rezultati i diskusija

Ovo poglavlje posvećeno je provedenim eksperimentima i predstavlja okosnicu ovog rada. Za početak, opisan je način na koji je sam eksperiment postavljen i zašto su odabrani modeli koji su odabrani. Zatim, predstavljeni su važniji dijelovi programskog koda korištenog za provedbu eksperimenata. Konačno, predstavljeni su rezultati, kratka analiza čestih pogrešaka modela i diskusija dobivenih rezultata.

S obzirom na to da je glavna motivacija ovog rada istraživanje o kvaliteti potencijalnog alata za detekciju demencije iz govora, izabrano je nekoliko različitih pristupa eksperimentima kako bi se istražio što veći broj slučajeva. eksperimenti su podijeljeni u dvije različite grupe: eksperimenti s modelima koji rade klasifikaciju demencije iz teksta te eksperimenti s modelima koji rade klasifikaciju demencije izravno iz govora. U nastavku slijedi detaljan opis poduzetih koraka za predobradu podataka i učenje modela u nadi da će zainteresirani čitatelj moći vjerno reproducirati dobivene rezultate.

6.1. Postavka eksperimenta - tekstni pristup

Transkripte dostupne u skupu podataka Pittov korpus napravili su profesionalni lingvisti. No, ako se razmotri realni slučaj u kojem bi se alat za automatsku detekciju demencije koristio, jasno je da je neisplativo koristiti profesionalne lingviste za transkripciju govora u tekst. U realnom slučaju, pacijent bi došao na ispitivanje kod doktora te bi pričao u mikrofon, u slučaju testa slike krađe kolačića opisao bi što vidi na slici, a zatim bi program za automatsko prepoznavanje govora njegov govor transkribirao i slao ga dalje na obradu modelu za klasifikaciju demencije iz teksta. Sukladno upravo takvim, realnim slučajem, eksperiment s tekstom podijeljen je u dva slučaja. U prvom slučaju modelima su dani transkripti dostupni u Pittovom korpusu. S obzirom na to da su ti transkripti doista precizno transkribirani, pretpostavka je da će dobiveni rezultati takvog eksperimenta predstavljati svojevrsnu gornju granicu za točnost modela. U drugom slučaju, zvučni zapisi razgovora ispitanika i doktora dostupni u Pittovom korpusu prvo su bili pušteni kroz odabrani model za automatsko prepoznavanje govora, a zatim

se tako transkribirani tekst koristio za učenje modela za detekciju demencije iz teksta.

6.1.1. Predobrada podataka i postavka učenja za eksperimente s lingvističkim transkriptima

Kao što je već spomenuto u poglavlju o skupu podataka, transkripti dostupni u Pittovom korpusu napisani su u CHAT formatu [31]. Naime, s obzirom na to da transkripti sadrže i govor ispitanika i govor doktora koji je vodio ispitivanje, prvi korak predobrade bio je izvlačenje samo izgovorenih riječi ispitanika iz transkripata. Nadalje, sve morfofonološke informacije i informacije o gramatičkim odnosima koje su dostupne u transkriptima su odbačene.

Zatim, bilo je potrebno odlučiti koje informacije iz transkribiranog govora ispitanika odbaciti, a koje zadržati. Kako bi se pobliže shvatila motivacija za odbacivanjem, koja na prvi pogled može izgledati neutemeljeno, potrebno je detaljnije objasniti sadržaj transkripata. Transkripti ne sadrže samo izgovorene riječi ispitanika, već i dodatne informacije kao što su informacije o jednostavnim događajima koje je osoba izvršila pri svom govoru. Takvi jednostavni događaji u transkriptima su prethođeni parom znakova: *&=*. Primjerice, ako se u transkriptu pojavljuje *"&=coughs"*, to znači da se ispitanik točno u tom trenutku zakašljao. Nadalje, transkripti sadrže niz različitih poštapalica (na engleskom) kao što su : *"uh", "um", "er"* i tako dalje. Spomenute poštapalice u transkriptima prethođene su parom znakova: *&-*. Konačno, transkripti također sadrže i dijelove govora koje je ispitanik ponovio, prethođene znakovima *[/]*), dijelove govora na koje se ispitanik vratio (engl. *retraced*), prethođene znakovima *[///]* te dijelove govora koje je ispitanik reformulirao, označene znakovima *[///]*. U daljnjem tekstu na sve takve odsječke govora referirat će se pod nazivom ponovljeni govor.

Početna pretpostavka bila je da poštapalice i ponovljeni govor mogu znatno utjecati na detekciju demencije. Stoga, napravljena su tri različita eksperimenta. Prvi eksperiment sadržavao je transkripte u kojima su uklonjene sve informacije spomenute u prethodnom odlomku. Drugi eksperiment sadržavao je transkripte u kojima je uklonjeno sve osim ponovljenih dijelova govora, odnosno sve osim dijelova transkripata koji su prethođeni znakovima *[/], [///]* i *[///]*. Treći i konačni eksperiment ovog dijela uključivao je korištenje transkripata u kojima je uklonjeno sve osim poštapalica i ponovljenih dijelova govora. Dakle, u trećem su eksperimentu iz transkripata uklonjeni sve dodatne informacije, osim onih koji su označeni znakovima *[/], [///], [///]* i *&-*.

Primjer dva predobrađena transkripta može se vidjeti u tablici 6.1. Crveni tekst, koji je korišten samo u trećem eksperimentu, označava poštapalice. Plavi tekst, kori-

šten u drugom i trećem eksperimentu, predstavlja dijelove govora koji su ponovljeni, dijelove govora na koje se ispitanik vratio i dijelove govora koje je ispitanik reformulirao. Prvi eksperiment ne sadrži obojani tekst prikazan u tablici 6.1.

S obzirom na to da je, kao što je spomenuto u poglavlju o skupu podataka, svaki ispitanik sudjelovao u testu slike krađe kolačića između jedan i tri puta, uzorke se grupiralo prema pacijentima. To je napravljeno kako bi se izbjeglo učenje i testiranje modela na istim pacijentima. Dakle, uz pretpostavku da je pacijent Ivan Ivić sudjelovao u ispitivanju tri puta, pri učenju i testiranju modela osigurano je da se svaki uzorak ispitivanja Ivana Ivića pojavi samo u skupu za učenje ili samo u skupu za testiranje.

Nadalje, skup podataka nije savršeno ujednačen. Kontrolna grupa sadrži 242 uzorka, što čini 47.54% cjelokupnog skupa podataka. Naime, grupiranje uzoraka po pacijentu dodalo je neujednačenosti. Razlog tomu je to što kontrolna grupa sadrži samo 98 ispitanika, što sačinjava samo 35.64% ukupnog broja ispitanika. Kako bi se pristranost smanjila što je više moguće, u svim eksperimentima koristila se stratificirana deseterostruka unakrsna validacija (engl. *stratified 10-fold cross-validation*). Dakle, 275 ispitanika podijeljeno je u 10 grupa, u prosjeku 27 pacijenata po grupi.

Što se tiče modeliranja, osnovne verzije BERT-a, RoBERTe i XLNETa, dostupne u biblioteci HuggingFace Transformers [47] su korištene. eksperimentirano je sa maksimalnim duljinama od 256 i 512 tokena za BERT, RoBERTa i XLNet. Veličina podskupa za učenje od 16 korištena je u svim eksperimentima. Implementacija optimizatora AdamW iz HuggingFace transformer biblioteke korištene je zajedno sa sljedećim parametrima: $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e-8$ i L2 propadanje težina od 0.01. Stopa učenja postavljena je na $5e-6$. Također je eksperimentirano s korištenjem optimizatora MADGRAD [12]. Iako je on pokazao obećavajuće rezultate u obliku brže konvergencije, odlučeno je da će se koristiti AdamW jer postoji neusporedivo više dostupno literature za AdamW u odnosu na MADGRAD. Za svaki preklop (engl. *fold*) novi model učen je 30 epoha. Nadalje, točnost, preciznost, odziv i mjera F1 zabilježeni su za svaki model i kasnije uprosječeni sa rezultatima svih modela deseterostruke unakrsne validacije.

6.1.2. Predobrada podataka i postavka učenja za eksperimente s transkriptima modela Wav2Vec2

Prvi korak predobrade podataka u ovom slučaju bio je obrada samih zvukovnih zapisa ispitivanja. Bilo je potrebno izrezati sve isječke ispitivanja u kojima se pojavljuje govor ispitanika. Srećom, CHAT format [31] za svaki izgovoreni dio sadrži i vremenske

Tablica 6.1: Primjeri transkripta za različite korake predobrade za svaki od tri eksperimenta s učenjem s lingvističkim transkriptima. Plavi tekst, koji označava ponovljene dijelove govora, korišten je u drugom i trećem eksperimentu. Crveni tekst, koji označava poštapalice, korišten je samo u trećem eksperimentu.

Transkript	Oznaka
mhm oh I see a part of the whole kitchen is that all the kitchen or isn't it uh oh I can't read a lady a mother were in her kitchen in her kitchen doing some work I suppose and the uh there's another woman there sharing their pleasures or whatever oh have you have you checked heard of that new game that they started to play after christmas did you is a well it looks like I'd say this is well let's see it looks like oh my mother will beat me by my wife will beat me by a couple rows of this that's that's like the washing would say washing machine or let me see I can't oh that's the son come out of from school maybe or something that's a youngster there well that's just as though they getting ready to go to school or they're just coming out from school and right there he's uh same as back there except for down there in the bottom I think it's uh that's a little	Dementia
okay uh the child's falling off the chair he's taking cookies out of the jar the girl is standing on the floor uh asking for a cookie the door to the cabinet door is open mother is washing dishes the sink is overflowing the water's running uh I don't know if she's dryin em or washin em anyway and the kitchen window has curtains the window's open um it looks like a view of the back there are three dishes on the uh counter	Control

oznake u kojima se pojavljuje govor stoga nije bilo potrebno koristiti alate za detekciju glasovne aktivnosti. U nekim slučajevima događalo se da se govor ispitanika i medicinskog stručnjaka koji obavlja ispitivanje preklapaju, odnosno radi se o trenucima kad bi oboje govorili u isto vrijeme. Umjesto rezanja takvih dijelova u kojima se događa preklapanje, oni su ostavljeni kakvi jesu. Sljedeći korak predobrade podataka bio je ponovno uzorkovanje (engl. *resampling*) zvučnih zapisa sa stopom uzorkovanja (engl. *sample rate*) od 16 kHz zbog zahtjeva modela Wav2Vec2 da su svi ulazni zapisi te frekvencije. Konačno, svi zapisi pretvoreni su iz .mp3 formata u .wav format, također zbog zahtjeva modela.

Korištena je verzija *wav2vec2-base-960h* modela Wav2Vec2. Spomenuta verzija ugođena je na čitavom skupu podataka Librispeech [35] koji se sastoji od 960 sati govora. Već u ovom trenutku pretpostavka je bila da bi model Wav2Vec2 ugođen na Librispeechu mogao imati problema s transkripcijom zvučnih zapisa ispitivanja. Naime, skup podataka Librispeech sastoji se od audiozapisa koji su, velikom većinom, vrlo čiste kvalitete. To je i za očekivati jer je sastavljen od audio knjiga, a audio knjige su pretežito napravljene tako da kvaliteta i čistoća govora čitača bude visoka. Međutim, kod audiozapisa Pittova korpusa to nije slučaj. Radi se o snimkama iz 80-ih godina prošlog stoljeća, a s obzirom na tadašnje stanje tehnologije nije ni za očekivati vrhunsku kvalitetu. Nadalje, zbog resursnih ograničenja nije bilo moguće učitati čitave audiozapise u model odjednom, već je bilo potrebno učitavati komadiće od 30 sekundi. Izlaz modela za svaki od spomenutih komadića potom se spajao kako bi se dobio jedan, cjeloviti izlaz. Pri dekodiranju korišten je pohlepni dekoder CTC (engl. *greedy CTC decoder*) [22].

S obzirom na to da je dekoder jedna od stvari koje nisu same po sebi implementirane u modelu Wav2Vec2, te s obzirom na to da je jedna od želja autora ovog rada da rezultati budu što reproducibilniji, u nastavku je prikazan isječak koda u kojem je prikazana korištena implementacija pohlepnog dekodera CTC. Kod 6.1. prikazuje klasu *"GreedyCTCDecoder"*, koja pri inicijalizaciji prima popis svih oznaka, odnosno popis svih slova abecede i ostalih znakova koje sustav koristi kroz varijablu *"labels"*, te indeks oznake koja predstavlja prazninu kroz varijablu *"blank"*. Metoda *"forward"* prima logite tenzora koji su izlaz iz modela Wav2Vec2 kroz varijablu *"emission"* te kao izlaz vraća dekodirani transkript.

```
class GreedyCTCDecoder(torch.nn.Module):
    def __init__(self, labels, blank=0):
        super().__init__()
        self.labels = labels
        self.blank = blank

    def forward(self, emission: torch.Tensor) -> str:

        indices = torch.argmax(emission, dim=-1)  # [num_seq,]
        indices = torch.unique_consecutive(indices, dim=-1)
        indices = [i for i in indices if i != self.blank]
        return "".join([self.labels[i] for i in indices])
```

Kod 6.1.: Klasa GreedyCTCDecoder

S obzirom na to da proces transkribiranja audiozapisa idejno ne spada u poglavlje o rezultatima, on će biti objašnjen u ovom potpoglavlju. Prije nego se predstave rezultati transkribiranja, potrebno je objasniti inicijalne pretpostavke autora. Naime, audiozapisi govora ispitanika nastali su, kao što je već objašnjeno u jednom od početnih poglavlja, u longitudinalnom istraživanju Alzheimerove demencije između 1983. i 1988. godine. Sukladno tomu, spomenuti audiozapisi izrazito su loše kvalitete, barem u usporedbi sa modernim standardima. Također, znatno variraju u kvaliteti. Određene snimke su dobre kvalitete, dok postoji nekoliko snimaka na kojima se gotovo ni ne čuje govor ispitanika. Dakle, pretpostavka je bila da će kvaliteta transkripata generiranih modelom Wav2Vec2 znatno varirati. Pokazalo se da je ta pretpostavka bila točna.

Kao mjera točnosti modela Wav2Vec2 uzet će se mjera stopa pogrešnih riječi, odnosno u daljnjem tekstu WER, prema kratici s engleskog jezika.

U tablici 6.2 prikazana je kratka statistika vrijednosti WER-ova dobivenih usporedbom generiranih transkripata i pripadnih lingvističkih transkripata. Vrijednosti najboljeg, odnosno najmanjeg WER-a i najlošijeg, odnosno najvećeg WER-a znatno se razlikuju. To potvrđuje pretpostavku da će kvaliteta generiranih skripti znatno varirati. Osim kvalitete audiozapisa, uzročnik lošeg WER-a na nekim primjerima može biti i naglasak ispitanika. Naime, model Wav2Vec2 prednaučen je na LibriSpeechu koji se uglavnom sastoji od audiozapisa ljudi koji govore standardnim, čistim engleskim naglaskom, dok u određenim audiozapisima Pittova korpusa ispitanici pričaju jakim američkim južnjačkim naglaskom i pri tome neke riječi spajaju ili ih krate.

Tablica 6.2: Statistike WER-ova izračunatih na temelju generiranih transkripata i pripadnih lingvističkih transkripata. Manji WER označava bolji rezultat.

Broj transkripcija	Prosječni WER	Standardna devijacija WER-a	Najbolji WER	Najlošiji WER
509	1.16	5.98	0.09	95.00

U tablici 6.3 prikazana su dva primjera. Prvi prikazani primjer je usporedba najboljeg generiranog transkripta, odnosno onog s najmanjom vrijednosti WER, sa odgovarajućim lingvističkim transkriptom. Drugi prikazani primjer je usporedba najlošijeg generiranog transkripta, odnosno onog s najvećom vrijednosti WER, sa odgovarajućim lingvističkim. Iz te tablice može se uočiti nekoliko stvari. Prva stvar je to da je najbolja transkripcija zaista dobra. Vrlo je slična pripadnom lingvističkom transkriptu. Druga, očitija stvar je katastrofalno loša transkripcija prikazana u drugom primjeru. Prikazani primjer ujedno je i najgore transkribiran tekst. Razlog tomu je prilično loša kvaliteta odgovarajućeg audiozapisa. Pri lošoj kvaliteti audiozapisa misli se na lošu kvalitetu snimke, koja se najčešće manifestira kao pretih, gotovo nečujan govor ispitanika, potencijalno zbog lošeg mikrofona, te na pozadinski šum prisutan u audiozapisima.

Iako je postojala mogućnost da se uklone sve stršeće vrijednosti, odnosno katastrofalno loše transkribirani zapisi, primjerice zapis prikazan u tablici 6.3, odlučeno je da će svi transkripti biti uključeni u eksperiment. Tako se omogućuje kvalitetnija usporedba između eksperimenata s lingvističkim transkriptima i eksperimenata s generiranima.

Konačno, nakon što su svi zapisi prošli kroz model Wav2Vec2, odnosno nakon što su stvoreni svi transkripti, izvršeno je učenje modelima BERT, RoBERTa i XLNet. Korištene su iste postavke kao i one iz prethodnog potpoglavlja.

6.2. Postavka eksperimenta - govorni pristup

Za eksperimente u kojima se demencija klasificirala izravno iz govora korišten je model AST [20] koji je detaljnije opisan u poglavlju o modelima. Koraci predobrade podataka u ovom slučaju bili su isti kao i oni kod predobrade podataka za ulaz u model Wav2Vec2. Dakle, ulaz u model bili su izrezani audiozapisi koji su sadržavali samo go-

Tablica 6.3: Usporedba najbolje i najlošije transkripcije sa odgovarajućim lingvističkim transkriptima.

Generirani transkript	Lingvistički transkript	WER
the little boy is in the cookie jar and standing on a stool that's falling out from under him and the mother has his washing dishes and the kitchen sunk but she doesn't have the plug gan and she's having a flood that's it	the little boy is in the cookie jar and standing on a stool that's falling out from under him and the mother has is washing dishes in the kitchen sink but she doesn't have the plug in and she's having a flood that's it	0.091
an	okay hm let me see the boy is getting cookies out of the cookie jar and the stool is just about to fall over the little girl is reaching up for a cookie and the mother is drying dishes the water is running in to the sink and the sink is running over onto the floor you want action you want all the action there is anyhow I think that's all the action there is and that little girl is laughing that little girl did I say the mother was drying dishes drying a dish	95.0

vor ispitanika s testa slike krađe kolačića, ponovno uzorkovani za stopom uzorkovanja od 16 kHz, prebačeni u .wav format.

Korištena je implementacija AST-a preuzeta s Githuba autora [19]. Bilo je potrebno načiniti niz promjena u dostupnom kodu. Za početak, stvorene su .json datoteke za skup za učenje i skup za testiranje. Svaka od json datoteka sadržavala je popis putanja do audiozapisa Pittova korpusa te njihovu pripadnu oznaku, dakle demencija ili kontrolna grupa. Osim toga, bilo je potrebno odrediti parametre za spektralnu augmentaciju "SpecAug" [36] koja se koristi pri učenju. Također, bilo je potrebno dodati normalizacijsku statistiku za skup podataka. Normalizacijska statistika sastoji se od prosjeka i standardne devijacije spektrograma korištenog skupa podataka. Konačno, bilo je potrebno podesiti neke od parametara za učenje kao što su stopa učenja i raspoređivač za stopu učenja, postavke zagrijavanja, optimizator i slične. U nastavku slijedi pregled svih korištenih parametara.

U tablici 6.4 prikazane su konačne vrijednosti parametara korištene pri učenju modela AST. Parametri koji nisu spomenuti postavljeni su na pretpostavljene vrijednosti. Može se primijetiti da su neki od parametara postavljeni na neočekivane vrijednosti, primjerice veličina podskupa za učenje od 2. Također, jedan od važnijih parametara, *"–target_length"*, koji označava koliko će se sekundi ulaznog audiozapisa uzimati pri klasifikaciji, postavljen je na 2450, što odgovara otprilike 24 sekunde. Ulazni audiozapisi u nekim slučajevima kraći su od tog, ali postoje i oni osjetno dulji od 24 sekunde. Spomenuta dva parametra nisu mogla biti veća zbog resursnih ograničenja. Za ilustraciju, model inicijaliziran s opisanim parametrima u potpunosti zauzima resurse grafičke kartice Tesla K80. Osim sa maksimalnom duljinom od 24 sekunde, provedena su i dva eksperimenta u kojima je maksimalna duljina ulaznog audiozapisa postavljena na 5 sekundi i na 20 sekundi, dok su svi ostali parametri jednakih vrijednosti kao i parametri u tablici 6.4.

6.3. Rezultati

U ovom potpoglavlju predstavljeni su rezultati svih provedenih eksperimenata. Diskusija o dobivenim rezultatima nalazi se u sljedećem potpoglavlju.

U eksperimentima s lingvističkim transkriptima i eksperimentima s transkriptima napravljenima modelom Wav2Vec2 korištena je stratificirana unakrsna deseterostruka validacija. Rezultati prikazani u tablicama predstavljaju prosjek svih konačnih modela na svakom preklopu. U eksperimentima s modelom AST, zbog vremenskih ograničenja prilikom izrade ovog rada, nije korištena unakrsna validacija. Dakle, rezultati

Tablica 6.4: Popis korištenih parametara i njihovih vrijednosti za učenje modela AST.

Parametar	Vrijednost	Opis
–lr	1e-5	Stopa učenja
–optim	Adam	Odabrani optimizator
–batch-size	2	Veličina podskupa za učenje
–n-epochs	20	Broj epoha učenja
–lr_patience	2	Koliko epoha čekati za smanjivanje stope učenja ako se točnost ne poveća
–freqm	48	Maksimalna duljina maske za frekvencije
–timem	128	Maksimalna duljina maske za vrijeme
–bal	None	Korištenje balansiranog uzorkovanja
–fstride	10	Pomak u frekvencijskoj domeni
–tstride	10	Pomak u vremenskoj domeni
–imagenet_pretrain	True	Korištenje modela AST prednaučenog na ImageNetu
–audioset_pretrain	True	Korištenje modela AST prednaučenog na Audiosetu
–target_length	2450	Duljina snimke koja se klasificira
–num_mel_bins	128	Broj particija Mel za particioniranje frekvencija

eksperimenata s modelom AST, za razliku od prethodnih eksperimenata, ne prikazuju uprosječeni rezultat nekoliko modela na različitim preklopima, već predstavljaju konačni rezultat jednog modela.

U tablici 6.5 mogu se vidjeti rezultati eksperimenata na testnom skupu u kojima su se koristili lingvistički transkripti i modeli BERT. S obzirom da je korišteni skup podataka donekle neujednačen, kao glavnu mjeru kvalitete modela u svim eksperimentima uzimat će se mjera F1. Sukladno tomu, u ovom slučaju najbolje rezultate postiže inačica modela BERT sa korištenjem maksimalne duljine tokenizatora od 256 na drugom eksperimentu. Spomenuti model ima mjeru F1 jednaku 86.89%, te točnost od 86.42%. Nadalje, može se primijetiti da je preciznost svih modela veća od odgovarajućih odziva. U prvom i u trećem eksperimentu bolje su se pokazale inačice BERT-a u kojima se koristila duljina tokenizatora od 512, u odnosu na drugi eksperiment u kojem je bolje rezultate postigla inačica BERT-a u kombinaciji s maksimalnom duljinom tokenizatora od 256.

Tablica 6.5: Rezultati eksperimenata s lingvističkim transkriptima i modelom BERT na testnom skupu. BERT256 označava model koji koristi maksimalnu duljinu tokenizra od 256, dok BERT512 označava model koji koristi maksimalnu duljinu tokenizatora od 512.

	Prvi eksperiment	Prvi eksperiment	Drugi eksperiment	Drugi eksperiment	Treći eksperiment	Treći eksperiment
Model	BERT256	BERT512	BERT256	BERT512	BERT256	BERT512
Preciznost	**91.86%**	90.17%	88.78%	90.55%	89.84%	90.76%
Odziv	80.08%	84.44%	**85.58%**	81.42%	83.27%	82.89%
Mjera F1	84.99%	86.61%	**86.89%**	85.34%	85.76%	86.36%
Točnost	85.29%	86.26%	**86.42%**	85.03%	85.22%	86.29%

U tablici 6.6 prikazani su rezultati eksperimenata u kojima su korišteni lingvistički transkripti i modeli RoBERTa na testnom skupu. U ovom slučaju najbolje rezultate postiže inačica modela RoBERTa sa korištenjem maksimalne duljine tokenizatora od 512 na drugom eksperimentu. Spomenuti model ima mjeru F1 jednaku 90.28%, te točnost od 90.16%. U svakom od tri eksperimenta s lingvističkim transkriptima gdje je korištena RoBERTa, inačica RoBERTe u kojima se koristila duljina tokenizatora od 512 postiže bolje rezultate od odgovarajuće inačice s duljinom tokenizatora od 256. Opet, kao i u slučaju eksperimenata s BERT-om, preciznosti u svakom eksperimentu veće su od odgovarajućih vrijednosti odziva.

U tablici 6.7 prikazani su rezultati eksperimenata u kojima su korišteni lingvistički transkripti i modeli XLNet. U ovom slučaju najbolje rezultate postiže inačica modela XLNet koja koristi maksimalnu duljinu tokenizatora od 256, na trećem eksperimentu.

Tablica 6.6: Rezultati eksperimenata s lingvističkim transkriptima i modelom RoBERTa. RoBERTa256 označava model koji koristi maksimalnu duljinu tokenizra od 256, dok RoBERTa512 označava model koji koristi maksimalnu duljinu tokenizatora od 512.

	Prvi eksperiment	Prvi eksperiment	Drugi eksperiment	Drugi eksperiment	Treći eksperiment	Treći eksperiment
Model	ROBERTA256	ROBERTA512	ROBERTA256	ROBERTA512	ROBERTA256	ROBERTA512
Preciznost	**94.26%**	93.46%	90.21%	92.81%	92.88%	91.87%
Odziv	80.31%	83.30%	87.27%	**88.60%**	83.46%	86.09%
Mjera F1	86.31%	87.75%	88.27%	**90.28%**	87.27%	88.49%
Točnost	86.93%	87.76%	87.74%	**90.16%**	87.22%	88.16%

Spomenuti model ima mjeru F1 jednaku 88.59%, te točnost od 88.37%. U prvom i u trećem eksperimentu bolje rezultate postižu inačice XLNeta u kojima je korištena maksimalna duljina tokenizatora od 256, dok je u drugom eksperimentu bolji rezultat postigla inačica sa maksimalnom duljinom tokenizatora od 512. I u ovom slučaju, kao i s prethodna dva modela, preciznost je veća od odziva u svakom eksperimentu za svaki model.

Tablica 6.7: Rezultati eksperimenata s lingvističkim transkripiama i modelom XLNet. XLNet256 označava model koji koristi maksimalnu duljinu tokenizatora od 256, dok XLNet512 označava model koji koristi maksimalnu duljinu tokenizatora od 512.

	Prvi eksperiment	Prvi eksperiment	Drugi eksperiment	Drugi eksperiment	Treći eksperiment	Treći eksperiment
Model	XLNet256	XLNet512	XLNet256	XLNet512	XLNet256	XLNet512
Preciznost	89.79%	86.36%	**90.77%**	90.21%	91.42%	90.70%
Odziv	83.39%	83.14%	83.78%	84.37%	**86.25%**	79.63%
Mjera F1	86.27%	83.77%	85.49%	86.93%	**88.59%**	83.98%
Točnost	85.92%	84.17%	86.11%	86.44%	**88.37%**	84.42%

U tablici 6.8 prikazani su rezultati eksperimenata s transkriptima koje je transkribirao model Wav2Vec2 za automatsko prepoznavanje govore. Najbolje rezultate postiže model XLNet sa maksimalnom duljinom tokenizatora od 512. Konkretnije, postigao je mjeru F1 od 82.77%, te točnost od 81.22%. Modeli BERT i RoBERTa s duljinama tokenizatora od 256 postigli su bolje rezultate od modela BERT i RoBERTa s duljinama tokenizatora od 512, dok je za XLNet situacija obratna. Zanimljivo je za primijetiti da, za razliku od eksperimenata s lingvističkim transkriptima, u ovim eksperimentima preciznost nije veća od odziva u svakom slučaju. Naime, odziv je veći u svim slučajevima osim u slučaju eksperimenta s modelom RoBERTa s maksimalnom duljinom tokenizatora od 256.

U tablici 6.9 prikazani su rezultati eksperimenata s modelom AST. Najbolji od tri

Tablica 6.8: Rezultati eksperimenata s transkriptima napravljenima modelom Wav2Vec2 za automatsko prepoznavanje govora. Sufiks 256 uz ime modela označava da je korištena maksimalna duljina tokenizatora od 256, dok sufiks 512 uz ime modela označava da je korištena maksimalna duljina tokenizatora od 512.

Model	BERT256	BERT512	RoBERTa256	RoBERTa512	XLNet256	XLNet512
Preciznost	80.81%	79.44%	**82.84%**	79.28%	80.59%	80.74%
Odziv	84.79%	84.40%	80.39%	82.49%	85.96%	**86.12%**
Mjera F1	82.31%	81.60%	81.04%	79.62%	82.51%	**82.77%**
Točnost	81.07%	79.95%	80.21%	77.80%	80.94%	**81.22%**

modela u eksperimentima pokazao se model AST sa postavljenim parametrom maksimalne duljine ulaznog audiozapisa od 24 sekunde.

Tablica 6.9: Rezultati eksperimenata s modelom AST. Provedena su tri eksperimenta u kojima je varirana maksimalna duljina ulaznog audiozapisa.

Model	AST-5-sec	AST-10-sec	AST-24-sec
Preciznost	56.45%	73.55%	**80.83%**
Odziv	72.30%	80.87%	**84.40%**
Mjera AUC	73.55%	77.71%	**82.61%**
Točnost	70.91%	76.43%	**81.82%**

6.4. Diskusija

Kao što je i očekivano, u sva tri eksperimenta s lingvističkim transkriptima, RoBERTa, čiji su rezultati prikazani u tablici 6.6, postiže bolje rezultate od modela BERT, čiji su rezultati prikazani u tablici 6.5. To je očekivano jer RoBERTa predstavlja svojevrsnu nadogradnju na model BERT. Također, na istim eksperimentima, RoBERTa postiže bolje rezultate od XLNeta, čiji su rezultati prikazani u tablici 6.7, osim u slučaju trećeg eksperimenta i korištenja maksimalne duljine tokenizatora od 256. Nadalje, u slučaju istih eksperimenata s RoBERTom, svi modeli s maksimalnom duljinom tokenizatora postavljenom na 512 postižu bolje rezultate od onih modela gdje je korištena maksimalna duljina tokenizatora od 256. Iz toga se može zaključiti da modeli RoBERTa imaju koristi od većeg ulaznog teksta. Kod istih eksperimenata s modelima BERT i

XLNet, situacija je nešto drugačija. Naime, u slučaju eksperimenata s BERT-om, najbolji rezultat postignut je u drugom eksperimentu koji koristi maksimalnu duljinu tokenizatora od 256. U ostala dva eksperimenta bolje rezultate postigli su modeli BERT s većom duljinom tokenizatora. Kod eksperimenata s XLNetom situacija je potpuno suprotna. U prvom i trećem eksperimentu bolje rezultate postigle su inačice XLNeta u kombinaciji s kraćim tokenizatorima.

Jedan zanimljiv detalj je to da u slučaju eksperimenata s lingvističkim transkriptima svi naučeni modeli imaju veću sklonost prema lažno negativnim primjercima nego prema lažno pozitivnim primjercima, što pokazuje činjenica da svi modeli imaju veću preciznost nego odziv. Vrlo vjerojatno, glavni uzrok tomu je djelomična neujednačenost skupa podataka. Pri učenju modela za detekciju demencije, trebalo bi se težiti većem odzivu. Razlog tomu je potreba za smanjenjem broja lažno negativnih primjeraka, kao što je i slučaj u većini medicinskih zadataka. Veći je problem ako se pacijentu s demencijom kaže da nema demenciju, jer će ostati bez prijeko potrebne medicinske pomoći, nego ako se zdravom pacijentu dijagnosticira demencija, jer će se daljnjim ispitivanjem vrlo vjerojatno ustvrditi nepostojanje demencije.

Uvjerljivo najgore rezultate u slučaju eksperimenata s lingvističkim transkriptima postižu modeli na prvom eksperimentu, odnosno na eksperimentu u kojem nisu uključene poštapalice i ponovljeni dijelovi govora. Modeli BERT i RoBERTa postigli su najbolje rezultate na drugom eksperimentu, a modeli XLNet najbolje rezultate postigli su na trećem eksperimentu. Iz svega spomenutog, može se donijeti nekoliko zaključaka. Prvo i najvažnije, ponovljeni govor karakterističan je za demenciju, stoga uključivanje ponovljenog govora u transkripte pospješuje rezultate modela. To potvrđuje učinak svih modela na prvom eksperimentu, koji je osjetno lošiji nego na drugom i trećem eksperimentu. Drugo, uključivanje poštapalica u skup podataka može pridonijeti boljem učinku modela za detekciju demencije. U slučaju modela BERT i RoBERTa to nije slučaj jer su mjera F1 i točnost bili manji na trećem eksperimentu nego na drugom. Međutim, u slučaju modela XLNet situacija je drukčija. Najbolji model XLNet upravo je onaj naučen na trećem eksperimentu iz čega se može zaključiti da se u slučaju modela XLNet uključivanje poštapalica u transkripte mogu poboljšati rezultati modela, no oni su i dalje lošiji od najboljih modela RoBERTe u drugom eksperimentu.

Gledajući presjek skupa neispravno klasificiranih uzoraka iz svakog eksperimenta, postoji 20 uzoraka koji svi modeli neispravno klasificiraju. Od tih 20 uzoraka, 13 ih ima oznaku demencije, a ostalih 7 pripada kontrolnoj grupi. U tablici 6.10 prikazana su dva uvijek neispravno klasificirana primjera, jedan iz kontrolne grupe, a drugi iz grupe s demencijom. Ponovno, plavi tekst označava dijelove transkripata koji su kori-

šteni u drugom i trećem eksperimentu, dok crveni tekst označava dijelove transkripata korištene samo u trećem eksperimentu.

Kao što je i pretpostavljeno, rezultati eksperimenata s transkriptima modela Wav2Vec2 znatno su lošiji od onih sa lingvističkim transkriptima. Kao najbolji model u ovim eksperimentima pokazao se XLNet sa maksimalnom duljinom tokenizatora postavljenom na 512. Obje inačice modela XLNet postižu bolje rezultate od ostala četiri naučena modela. Zanimljivo je za primijetiti da RoBERTa, model koji je postigao najbolje rezultate na eksperimentima s lingvističkim transkriptima, u ovom slučaju ima najlošije rezultate.

Kako bi se pokušalo objasniti zašto je tome tako, prvo je potrebno naglasiti neke stvari. Za početak, audiozapisi sa ispitivanja testa slike krađe kolačića dostupni u Pittovom korpusu daleko su od savršene kvalitete. Štoviše, kvaliteta značajno varira između pojedinih snimaka. Neki od audiozapisa gotovo su u potpunosti tihi ili imaju izrazito lošu kvalitetu snimljenog govora, dok neki od ostalih audiozapisa imaju sasvim zadovoljavajuću kvalitetu, kao što je već i spomenuto u opisu predobrade eksperimenta. Sukladno tomu, kao što je i pokazano u potpoglavlju o predobradi eksperimenta, kvaliteta stvorenih transkripata jako varira. Neke od audiozapisa, model Wav2Vec2 gotovo je idealno transkribirao, ali postoje i oni transkripti koji su praktički neupotrebljivi.

Česta pojava u dobivenim transkriptima je približno dobro transkribirana riječ, ali sa manjom pogreškom, primjerice *"cooky"* ili *"cokie"* umjesto *"cookie"*. Dakle, postoji potencijalno ogroman broj riječi izvan vokabulara modela. Nadalje, veliku ulogu u kvaliteti rezultata modela igra način tokenizacije ulaznih nizova. Modeli BERT koriste algoritam tokenizacije koji se zove kodiranje parova bajtova (engl. *byte pair encoding, BPE*) [40]. Modeli RoBERTa koriste vrlo sličan algoritam, kodiranje parova bajtova na razini bajtova (engl. *byte-level byte pair encoding, BBPE*), gdje je vokabular nešto manji nego kod originalnog kodiranja parova bajtova. Konačno, XLNet koristi algoritam tokenizacije *"SentencePiece"* [28]. Zaključno, XLNet se na zadatku detekcije demencije na transkriptima napravljenima korištenjem modela Wav2Vec2 pokazao kao najbolji.

Rezultati eksperimenata s AST-om pokazuju nekoliko stvari. Prvo i najvažnije, modeli za klasifikaciju zvuka imaju veliki potencijal za detekciju demencije. Kvaliteta takvih modela znatno ovisi o kvaliteti audiozapisa, čak i više nego kod modela automatskog prepoznavanja govora. To još više dolazi do izražaja kada se u obzir uzme činjenica da je korišteni model AST prednaučen na Audiosetu, skupu audiozapisa čija je kvaliteta vrlo dobra, neusporedivo bolja u odnosu na snimke iz Pittova korpusa. S obzirom na to, dobiveni rezultati i više su nego vrlo dobri. Drugo, model AST na za-

Tablica 6.10: Primjeri dva transkripta koje su svi modeli neispravno klasificirali u svim eksperimentima. Plavi tekst, koji označava ponovljene dijelove govora, korišten je u drugom i trećem eksperimentu. Crveni tekst, koji označava poštapalice, korišten je samo u trećem eksperimentu.

Transkript	Opis
okay I'll start the mother is drying dishes and the sink is over flowing the water is falling onto the floor uh the boy is on his stool uh taking cookies out of a cookie jar and he has one cookie two cookies one in each hand the uh girl is standing reaching up for a cookie with her uh finger over her mouth telling him to be quiet the stool is on one leg uh there's drapes on the window there's a path uh between the grass and the bushes and this little picture is a part of the house and part of the tree in the upper window there are uh uh doors on the uh cabinets in the sink and uh it's daylight um there's two cups and a dish on it on the sink should I describe the two faucets	Jedan od uzoraka iz demencijske grupe kojeg su svi modeli netočno klasificifali
well the boy's trying to get in this cookie jar and the stool overturns and uh the little girl is expecting to hand her a cookie uh the mother is her sink is running over and she's standing in some of the water and uh she's drying a dish or wiping a dish and uh you said everything is happening well the water is still runnin in the sink and I said it's it's overflowing and she's standing in the water and that's I guess look somebody laying in the lawn out there but I can't uh	Jedan od uzoraka iz kontrolne grupe kojeg su svi modeli netočno klasificirali

datku detekcije demencije ima koristi od duljih audiozapisa. Očevidno je poboljšanje točnosti, pogotovo nakon povećanja s 5 sekundi na 10 sekundi, vidljivo u tablici 6.9.

Učenje AST-a za detekciju demencije na Pittovom korpusu predstavlja prvi pokušaj primjene dubokog modela za klasifikaciju zvuka na zadatak detekcije demencije. Prema mišljenju autora ovog rada, ovakvi modeli za zadatak detekcije demencije imaju više potencijala od modela koji klasificiraju tekst, iako su tekstni modeli na eksperimentima postigli bolje rezultate. Nekoliko je razloga zašto bi modeli poput AST-a mogli postići bolje rezultate od tekstnih modela. Najvažniji razlog je to što modeli za klasifikaciju zvuka implicitno u stvorenim spektrogramima imaju dostupno znanje o nizu značajki govora, kao što su primjerice prozodijske značajke, a za koje je već pokazano [26] da su sasvim sigurno dobri pokazatelji nastupa demencije. Tekstni modeli sami po sebi ne mogu iskoristiti prozodijske značajke. Nadalje, dostupnost novih, kvalitetnih audiozapisa, primjerice postojanje moderne verzije Pittova korpusa gotovo sigurno omogućilo bi znatno bolje rezultate modela poput AST-a na zadatku detekcije demencije.

Za kraj, potrebno je dati još nekoliko općih komentara i dati kratku usporedbu rezultata sa srodnim radovima.

Odabrana podjela između skupa za učenje i testiranje mogla je jako utjecati na pouzdanost rezultata, s obzirom na to da je korišteni skup podataka vrlo malen i djelomično nebalansiran. Dva najčešće korištena skupa podataka u srodnim radovima su skupovi podataka Pittov korpus [5] i ADReSS Challenge [30]. Skup podataka ADReSS Challenge podskup je Pittova korpusa i savršeno je balansiran sa 78 ispitanika u demencijskoj skupini i 78 ispitanika u kontrolnoj skupini. Teško je izravno usporediti dobivene rezultate sa srodnim radovima jer, kao što je spomenuto ranije, neki od radova koriste skup podataka ADReSS Challenge, dok većina ostalih koji koriste Pittov korpus ne navode jesu li uključili ispitanike kojima je klasificiran MCI, odnosno blago kognitivno oštećenje, te na koji način su predobradili dostupne lingvističke transkripte.

Rezultati naučenog modela RoBERTa na lingvističkim transkriptima bolji su od onih koje su postigli Jonasson i Wahlforss [2] na istom zadatku koristeći lingvističke transkripte. Njihov najbolji model na lingvističkim transkriptima postiže točnost od 86.82% i preciznost od 90.69%, dok model RoBERTa naučen u sklopu ovog rada postiže točnost od 90.16% i preciznost od 92.81%. Razlika u rezultatima može biti zbog nekoliko stvari. Prvo, moguće je da su koristili druge, lošije parametre. Drugo, s obzirom na to da ne specificiraju na koji način su predobradili transkripte, njihova manja točnost može biti rezultat uklanjanja ponovljenih dijelova govora, ili poštapalica, za

koje je oboje pokazano da mogu utjecati na krajnje rezultate. Također, kao što je već rečeno, razlika u rezultatima moguća je zbog različite podjele između skupa za učenje i testiranje.

Nadalje, Jonasson i Wahlforss [2] također eksperimentiraju s transkriptima generiranima od strane modela za automatsko prepoznavanje govora. Njihov najbolji model na tom zadatku, RoBERTa s duljinom tokenizatora od 512 postiže točnost od 83.59%, dok najbolji model u ovom radu, XLNet sa duljinom tokenizatora od 512 postiže nešto lošiji rezultat sa točnošću od 81.22%. Razlog tomu vrlo je očit. Jonasson i Wahlforss u svom radu koriste Googleovu uslugu za automatsko prepoznavanje govora, dok je u ovom radu korišten javno dostupni model Wav2Vec2. Usluga koju pruža Google komercijalno je rješenje za automatsko prepoznavanje govora te zasigurno daje bolje rezultate od modela Wav2Vec2 korištenog u ovom radu. Nažalost, Jonasson i Wahlforss nisu pružili analizu WER-a stoga se ne mogu direktno usporediti Googleova usluga i Wav2Vec2 korišten u ovom radu. Daljnjim ugađanjem modela Wav2Vec2 , primjerice na snimkama lošije kvalitete kao što su i snimke Pittova korpusa, vrlo vjerojatno bi se mogli postići rezultati bliski Googleovoj usluzi za automatsko prepoznavanje govora.

7. Zaključak

Detekcija demencije izrazito je zahtjevan zadatak. Postoji niz prepreka, od kojih je vrlo vjerojatno najveća, nepostojanje adekvatnog, dovoljno velikog i ažurnog skupa podataka.

U ovom radu, pokazano je da prednaučeni transformerski modeli mogu ostvariti značajne rezultate na zadatku detekcije demencije. Najbolji model, RoBERTa učen na lingvističkim transkriptima postigao je točnost od 90.16%. Isto tako, pokazan je potencijal dubokih modela za klasifikaciju zvuka na zadatku detekcije demencije, konkretnije transformera audio spektrograma (AST). Glavnu prepreku predstavlja loša kvaliteta postojećih snimki dostupnih u Pittovom korpusu.

eksperimenti s transkriptima koje je iz audiozapisa Pittova korpusa generirao model za automatsko prepoznavanje govora Wav2Vec2, pokazuju da je moguće stvoriti jedinstveni alat za detekciju demencije koji bi davao zadovoljavajuće rezultate. Takav alat mogao bi se koristiti u kombinaciji sa standardiziranim testom za detekciju demencije te bi služio medicinskim stručnjacima kao pomoć pri dijagnozi. Autor ovog rada nada se da bi ovakvi radovi u kojima se pokazuje neizmjerni potencijal modela dubokog učenja mogli potaknuti stvaranje većeg, standardiziranog skupa podataka, a što bi moglo dovesti do razvoja uistinu moćnog alata za detekciju demencije koji bi pomogao odgoditi razvijanje težih stupnjeva demencije kod oboljelih, te samim time pomoći milijunima ljudi diljem svijeta i znatno im produžiti životni vijek.

Što se tiče potencijalnih nadogradnji i budućih radova, bilo bi dobro istražiti može li svojevrsno uključivanje prozodijskih značajki u kombinaciji sa tekstnim dubokim modelima pružiti dobre rezultate. Osim toga, moguće poboljšanje nalazi se u korištenju modela Longformer [6] koji bi dozvolio obradu jezičnih nizova duljih od 512 tokena.

LITERATURA

[1] Z. Arvanitakis, R. C. Shah, i D. A. Bennett. Diagnosis and Management of Dementia: Review. *JAMA*, 322(16):1589–1599, 10 2019. ISSN 0098-7484. doi: 10.1001/jama.2019.4782. URL https://doi.org/10.1001/jama.2019.4782.

[2] A. Aslaksen Jonasson i A. Wahlforss. Diagnosis of dementia using transformer models, 2020. URL http://kth.diva-portal.org/smash/record.jsf?dswid=-4861&pid=diva2%3A1458824.

[3] A. Baevski, Y. Zhou, A. Mohamed, i M. Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. U H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, i H. Lin, urednici, *Advances in Neural Information Processing Systems*, svezak 33, stranice 12449–12460. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf.

[4] D. Bahdanau, K. Cho, i Y. Bengio. Neural machine translation by jointly learning to align and translate. U *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1409.0473.

[5] J. T. Becker, F. Boller, O. L. Lopez, J. Saxton, i K. L. McGonigle. The natural history of Alzheimer's disease. Description of study cohort and accuracy of diagnosis. *Arch Neurol*, 51(6):585–594, Jun 1994.

[6] I. Beltagy, M. E. Peters, i A. Cohan. Longformer: The long-document transformer. *CoRR*, abs/2004.05150, 2020. URL https://arxiv.org/abs/2004.05150.

[7] R. S. Bucks, S. Singh, J. M. Cuerden, i G. K. Wilcock. Analysis of spontaneous, conversational speech in dementia of Alzheimer type: Evaluation of an

objective technique for analysing lexical performance. *Aphasiology*, 14(1):71–91, 2000. doi: 10.1080/026870300401603. URL https://doi.org/10.1080/026870300401603.

[8] H. Chertkow, H. H. Feldman, C. Jacova, i F. Massoud. Definitions of dementia and predementia states in Alzheimer's disease and vascular cognitive impairment: consensus from the Canadian conference on diagnosis of dementia. *Alzheimer's research & therapy*, 5(1):1–8, 2013.

[9] K. Cho, B. van Merrienboer, Ç. Gülçehre, F. Bougares, H. Schwenk, i Y. Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. U *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, stranice 1724–1734, Doha, Qatar, listopad 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1179. URL https://aclanthology.org/D14-1179.

[10] L. Cummings. Describing the Cookie Theft picture: Sources of breakdown in Alzheimer's dementia. *Pragmatics and Society*, 10:151–174, 03 2019. doi: 10.1075/ps.17011.cum.

[11] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. Le, i R. Salakhutdinov. Transformer-XL: Attentive language models beyond a fixed-length context. U *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, stranice 2978–2988, Florence, Italy, srpanj 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1285. URL https://aclanthology.org/P19-1285.

[12] A. Defazio i S. Jelassi. A momentumized, adaptive, dual averaged gradient method. *Journal of Machine Learning Research*, 23(144):1–34, 2022. URL http://jmlr.org/papers/v23/21-0226.html.

[13] J. Deng, W. Dong, R. Socher, L. Li, K. Li, i L. Fei-Fei. Imagenet: A large-scale hierarchical image database. U *2009 IEEE Conference on Computer Vision and Pattern Recognition*, stranice 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

[14] J. Devlin, M. Chang, K. Lee, i K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. U *Proceedings of the*

2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), stranice 4171–4186, Minneapolis, Minnesota, lipanj 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

[15] V. O. Emery. Language impairment in dementia of the Alzheimer type: a hierarchical decline? *Int J Psychiatry Med*, 30(2):145–164, 2000.

[16] E. Eyigoz, S. Mathur, M. Santamaria, G. Cecchi, i M. Naylor. Linguistic markers predict onset of Alzheimer's disease. *EClinicalMedicine*, 28(100583):100583, studeni 2020.

[17] J. Fritsch, S. Wankerl, i E. Nöth. Automatic Diagnosis of Alzheimer's Disease Using Neural Network Language Models. U *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, stranice 5841–5845, 2019. doi: 10.1109/ICASSP.2019.8682690.

[18] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, i M. Ritter. Audio set: An ontology and human-labeled dataset for audio events. U *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, stranice 776–780, 2017. doi: 10.1109/ICASSP.2017.7952261.

[19] Y. Gong. Audio spectrogram transformer. https://github.com/YuanGongND/ast, 2021.

[20] Y. Gong, Y. Chung, i J. R. Glass. AST: audio spectrogram transformer. *CoRR*, abs/2104.01778, 2021. URL https://arxiv.org/abs/2104.01778.

[21] H. Goodglass i E. Kaplan. *The assessment of aphasia and related disorders*. Lea & Febiger, 1972.

[22] A. Graves, S. Fernández, F. Gomez, i J. Schmidhuber. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. U *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, stranica 369–376, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595933832. doi: 10.1145/1143844.1143891. URL https://doi.org/10.1145/1143844.1143891.

[23] L. Hernandez-Dominguez, S. Ratte, G. Sierra-Martinez, i A. Roche-Bergua. Computer-based evaluation of Alzheimer's disease and mild cognitive impairment patients during a picture description task. *Alzheimers Dement (Amst)*, 10: 260–268, ožujak 2018.

[24] L. Ilias i D. Askounis. Explainable identification of dementia from transcripts using transformer networks. *IEEE Journal of Biomedical and Health Informatics*, stranice 1–1, 2022. doi: 10.1109/JBHI.2022.3172479.

[25] S. Karlekar, T. Niu, i M. Bansal. Detecting Linguistic Characteristics of Alzheimer's Dementia by Interpreting Neural Models. U *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, stranice 701–707, New Orleans, Louisiana, lipanj 2018. ACL. doi: 10.18653/v1/N18-2110. URL https://aclanthology.org/N18-2110.

[26] A. Khodabakhsh, F. Yesil, E. Guner, i C. Demiroglu. Evaluation of linguistic and prosodic features for detection of alzheimer's disease in turkish conversational speech. *EURASIP Journal on Audio, Speech, and Music Processing*, 2015, 12 2015. doi: 10.1186/s13636-015-0052-y.

[27] A. König, A. Satt, A. Sorin, R. Hoory, O. Toledo-Ronen, A. Derreumaux, V. Manera, F. Verhey, P. Aalten, P. H Robert, i R. David. Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease. *Alzheimers Dement. (Amst.)*, 1(1):112–124, ožujak 2015.

[28] T. Kudo i J. Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. U *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, stranice 66–71, Brussels, Belgium, studeni 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-2012. URL https://aclanthology.org/D18-2012.

[29] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, i V. Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. URL http://arxiv.org/abs/1907.11692.

[30] S. Luz, F. Haider, S. de la Fuente, D. Fromm, i B. MacWhinney. Alzheimer's Dementia Recognition through Spontaneous Speech: The ADReSS Challenge, 2020.

[31] B. MacWhinney. *The CHILDES project: Tools for analyzing talk: Volume I: Transcription format and programs, Volume II: The database*. Lawrence Erlbaum Associates Publishers, 3. izdanje, 2000.

[32] E. Nichols, J. D Steinmetz, S. E. Vollset, K. Fukutaki, J. Chalek, F. Abd-Allah, A. Abdoli, A. Abualhasan, E. Abu-Gharbieh, T. T. Akram, i et al. Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: An analysis for the global burden of disease study 2019. *The Lancet Public Health*, 7(2):E105–E125, Feb 2022.

[33] S. O. Orimaye, J. S. Wong, i K. J. Golden. Learning predictive linguistic features for Alzheimer's disease and related dementias using verbal utterances. U *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, stranice 78–87, Baltimore, Maryland, USA, lipanj 2014. ACL. doi: 10.3115/v1/W14-3210. URL https://aclanthology.org/W14-3210.

[34] S. O. Orimaye, J. S. Wong, i J. S. G. Fernandez. Deep-deep neural network language models for predicting mild cognitive impairment. U Abdoulaye Baniré Diallo, Engelbert Mephu Nguifo, i Mohammed Zaki, urednici, *Second international workshop on Advances in Bioinformatics and Artificial Intelligence: Bridging the Gap (BAI 2016)*, CEUR Workshop Proceedings, stranice 14–20. Rheinisch-Westfaelische Technische Hochschule Aachen, 2016. URL http://ceur-ws.org/Vol-1718/. BAI 2016, 11-07-2016.

[35] V. Panayotov, G. Chen, D. Povey, i S. Khudanpur. Librispeech: An asr corpus based on public domain audio books. U *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, stranice 5206–5210, 2015. doi: 10.1109/ICASSP.2015.7178964.

[36] D. S. Park, W. Chan, Y. Zhang, C. Chiu, B. Zoph, E. D. Cubuk, i Q. V. Le. SpecAugment: A simple data augmentation method for automatic speech recognition. U *Interspeech 2019*. ISCA, rujan 2019. doi: 10.21437/interspeech.2019-2680. URL https://doi.org/10.21437%2Finterspeech.2019-2680.

[37] K. J. Piczak. ESC: Dataset for Environmental Sound Classification. U *Proceedings of the 23rd Annual ACM Conference on Multimedia*, stranice 1015–1018. ACM Press. ISBN 978-1-4503-3459-4. doi: 10.1145/2733373.2806390. URL http://dl.acm.org/citation.cfm?doid=2733373.2806390.

[38] A. Roshanzamir, H. Aghajan, i M. Soleymani. Transformer-Based Deep Neural Network Language Models for Alzheimer's Disease Detection from Targeted Speech. *BMC Medical Informatics and Decision Making*, srpanj 2020. doi: 10.21203/rs.3.rs-49267/v1.

[39] P. Saltz, S. Lin, S. Cheng, i D. Si. Dementia detection using transformer-based deep learning and natural language processing models. U *2021 IEEE 9th International Conference on Healthcare Informatics (ICHI)*, stranice 509–510, Los Alamitos, CA, USA, aug 2021. IEEE Computer Society. doi: 10.1109/ICHI52183.2021.00094. URL https://doi.ieeecomputersociety.org/10.1109/ICHI52183.2021.00094.

[40] R. Sennrich, B. Haddow, i A. Birch. Neural machine translation of rare words with subword units. U *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, stranice 1715–1725, Berlin, Germany, kolovoz 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1162. URL https://aclanthology.org/P16-1162.

[41] G. Szatloczki, I. Hoffmann, V. Vincze, J. Kalman, i M. Pakaski. Speaking in Alzheimer's Disease, is That an Early Sign? Importance of Changes in Language Abilities in Alzheimer's Disease. *Front Aging Neurosci*, 7:195, 2015.

[42] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, i I. Polosukhin. Attention is all you need. U I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, i R. Garnett, urednici, *Advances in Neural Information Processing Systems*, svezak 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[43] S. Vrljić. Poštapalice u hrvatskom jeziku. *Jezik : časopis za kulturu hrvatskoga književnog jezika*, 54(2):60–64, 2007.

[44] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, i S. R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. U

Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, stranice 353–355, Brussels, Belgium, studeni 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5446. URL https://aclanthology.org/W18-5446.

[45] P. Warden. Speech commands: A dataset for limited-vocabulary speech recognition. *CoRR*, abs/1804.03209, 2018. URL http://arxiv.org/abs/1804.03209.

[46] T. A. Widiger, P. T. Costa, American Psychological Association, et al. *Personality disorders and the five-factor model of personality*. JSTOR, 2013.

[47] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, i J. Brew. Transformers: State-of-the-art natural language processing. U *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, stranice 38–45, Online, listopad 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6.

[48] M. Yancheva i F. Rudzicz. Vector-space topic models for detecting Alzheimer's disease. U *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, stranice 2337–2346, Berlin, Germany, kolovoz 2016. ACL. doi: 10.18653/v1/P16-1221. URL https://aclanthology.org/P16-1221.

[49] Z. Yang, Z. Dai, Y-Yang, J. G. Carbonell, R. Salakhutdinov, i Q. V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. U H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, i R. Garnett, urednici, *Advances in Neural Information Processing Systems*, svezak 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/dc6a7e655d7e5840e66733e9ee67cc69-Paper.pdf.

[50] J. J. Young, M. Lavakumar, D. Tampi, S. Balachandran, i R. R. Tampi. Frontotemporal dementia: latest evidence and clinical implications. *Therapeutic advances in psychopharmacology*, 8(1):33–48, 2018.

Automatska detekcija demencije iz govora koristeći transformerske modele

Sažetak

Demencija je ozbiljna bolest koja je najčešće kronične ili progresivne naravi te je jedan od glavnih uzroka invalidnosti među starijom svjetskom populacijom. Automatska detekcija demencije težak je zadatak. Uključuje analizu akustičkih značajki govora, lingvističkih značajki transkripata i slično. Ovaj rad bavi se istraživanjem mogućnosti modela dubokog učenja koji iz lingvističkih transkripata te direktno iz govora detektiraju potencijalnu demenciju. Korišten je skup podataka Pittov korpus, dio DementiaBank dijeljene baze podataka čija je glavna namjena proučavanje demencije i srodnih bolesti. eksperimentirano je s modelima kao što su BERT, RoBERTa, XLNet i Audio Spectrogram Transformer. Također, istražene su mogućnosti modela Wav2Vec2 za automatsko prepoznavanje govora kojim su pripremljeni transkripti koje su zatim poslužile kao ulaz u tekstne modele. Pokazano je da se korištenjem kvalitetnih transkripata mogu postići točnosti detekcije iznad 90%. Konačno, predstavljena je detaljna diskusija rezultata, u kojoj su između ostaloga uspoređeni rezultati korištenih pristupa detekciji demencije, te je predstavljena jezgrovita analiza određenih pogreški.

Ključne riječi: demencija, duboko učenje, Pittov korpus, transformer, RoBERTa, XLNet, Audio Spectrogram Transformer, automatsko prepoznavanje govora

Automatic detection of dementia from speech using transformer models

Abstract

Dementia is a serious disease that is usually chronic and progressive, while also being one of the leading causes of disability among the elderly population. Automatic detection of dementia is a difficult task. It includes the analysis of acoustic features of speech, linguistic features of transcripts and other similar features. This work explores the capabilities of deep learning models to accurately detect dementia from linguistic transcripts or directly from speech. The Pitt corpus dataset was used, which is a part of DementiaBank, a shared database intended for studying dementia and other similar diseases. Experiments were conducted with deep learning models such as BERT, RoBERTa, XLNet and Audio Spectrogram Transformer. Furthermore, Wav2Vec2 model was used for preparing transcripts that were later used as the input for text-based models, and its capabilities were assessed. It was shown that dementia detection using well-prepared speech transcripts can achieve accuracy rates of above 90%. Finally, a discussion on the results of the used approaches is given, and a brief analysis of common errors is presented.

Keywords: dementia, deep learning, Pitt corpus, transformer, RoBERTa, XLNet, Audio Spectrogram Transformer, automatic speech recognition