Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

Žan Pečovnik

Medjezikovni prenos napovednih modelov za sovražni govor

diplomsko delo

univerzitetni študijski program
prve stopnje
računalništvo in informatika

Mentor: prof. dr. Marko Robnik Šikonja
Somentor: doc. dr. Nikola Ljubešić

Ljubljana, 2020

Besedilo je oblikovano z urejevalnikom besedil LaTeX.

Fakulteta za računalništvo in informatiko izdaja naslednjo nalogo:

Tematika naloge:

Sovražen govor je pogosta težava družbenih omrežij, saj ovira normalno komuniciranje večine uporabnikov ter preprečuje demokratični družbeni dialog. Ročno moderiranje javnih spletnih mest je zamudno, zato se poskuša razviti avtomatsko prepoznavanje sovražnega govora z metodami strojnega učenja. Za razvoj napovednih modelov potrebujemo označene učne množice, ki za manjše jezike večinoma ne obstajajo ali pa so (pre)majhne. Medjezikovne vložitve in veliki vnaprej naučeni večjezikovni modeli omogočajo prenos naučenih napovednih modelov med jeziki. V zadnjem času so se uveljavili modeli tipa BERT, ki temeljijo na arhitekturi globokih nevronskih mrež transformer. Na podatkovnih množicah sovražnega govora, enotno označenega v treh jezikih, preizkusite medjezikovni prenos napovednih modelov temelječih na modelu BERT. Za učenje uporabite različne jezike in različno velike dele učne množice v ciljnem jeziku. Pristop statistično ovrednotite.

Zahvaljujem se mentorju prof. dr. Marku Robniku Šikonji in somentorju dr. Nikoli Ljubešiću za hitro odzivnost, nasvete in pomoč pri izdelavi diplomske naloge.

Rad bi se zahvalil še svoji družini, dekletu in prijateljem, ki so mi pomagali ter me spodbujali skozi celoten študij.

Kazalo

Seznam uporabljenih kratic

kratica	angleško	slovensko
GPU	Graphical Processing Unit	Grafična procesna enota
CPU	Central Processing Unit	Centralna procesna enota
LGBT	Lesbian, gay, bisexual, transsexual people	Lezbijke, geji, biseksualne in transspolne osebe
BERT	Bidirectional Encoder Representations from Transformers	Predstavitve z dvosmernimi kodirniki transformerjev - model BERT
NLP	Natural Language Processing	Obdelava naravnega jezika
RNN	Recurrent Neural Netowrk	Rekurenčna nevronska mreža
CNN	Convolutional Neural Network	Konvolucijska nevronska mreža
SVM	Support Vector Machines	Metoda podpornih vektorjev

Povzetek

Naslov: Medjezikovni prenos napovednih modelov za sovražni govor

Avtor: Žan Pečovnik

Z razvojem družbenih omrežij je narasla pogostost sovražnega govora v uporabniških vsebinah. Osredotočili se bomo na dve trenutno najbolj aktualni temi, LGBT in migrante. Za napovedovanje sovražnega govora bomo uporabili nevronsko mrežo BERT in naredili primerjavo med večjezikovnim modelom, ki je naučen na 104 različnih jezikih ter trojezikovnim modelom, ki je naučen na slovenščini, hrvaščini in angleščini. Ugotovili smo, da trojezikovni model za približno 5% natančneje napoveduje sovražni govor na jeziku, na katerem je bil model tudi naučen. Večjezikovni model, brez ali z dodatnim učenjem, natančneje kot trojezikovni model napoveduje sovražni govor na jezikih, na katerem prvotno model ni bil naučen. To kaže na boljši medjezikovni prenos večjezikovnega napovednega modela.

Ključne besede: sovražni govor, model BERT, nevronske mreže, medjezikovni prenos, strojno učenje, obdelava naravnega jezika.

Abstract

Title: Cross-lingual transfer of hate speech prediction models

Author: Žan Pečovnik

With the development of social networks, there has been a significant increase of hate speech in user generated contents. We focus on two most discussed topics, LGBT and migrants. We use the BERT neural network for prediction of hate speech and make a comparison between the multilingual model, trained on 104 different languages, and a trilingual model, trained on Slovene, Croatian and English. Results show that the trilingual model is approximately 5% more accurate predicting hate speech on a language that it was trained on. The multilingual model with or without additional training is more accurate on languages that it was not trained on. This indicates a better cross-lingual transfer of multilingual model.

Keywords: hate speech, BERT model, neural networks, cross-lingual transfer, machine learning, natural language processing.

Poglavje 1

Uvod

Z razvojem družbenih omrežij je narasla pogostost sovražnega govora v uporabniških komentarjih. Osredotočili se bomo na analizo in prepoznavanje sovražnih komentarjev pri dveh aktualnih temah. Glede na trenutno situacijo na Bližnjem Vzhodu, ko veliko ljudi beži pred nevarnimi razmerami in išče zatočišče v evropskih državah, je ena najbolj razširjenih tem migranti. Že več stoletij je tema LGBT tabu in tako ostaja še danes. Pri strojni analizi sovražnega govora največkrat poskušamo razločiti med nesovražnim in sovražnim govorom [13] ter med različnimi vrstami sovražnega govora [23, 24].

V zadnjem času se za klasifikacijo besedil pretežno uporabljajo nevronske mreže [5]. Za njihovo uporabo je besedila potrebno pretvoriti v vektorsko obliko t.i. vektorsko vložitev [14]. Ključna značilnost vektorskih vložitev je, da so si besede blizu v vektorskem prostoru, če so si tudi pomensko blizu. Za napovedovanje sovražnega govora bomo uporabili najsodobnejši napovedni model na področju procesiranja naravnega jezika BERT [6].

1.1 Sorodna dela

Na kratko predstavljamo raziskave iz dveh ključnih tem našega dela, detekcije sovražnega govora in medjezikovnega prenosa.

V zadnjem času se raziskovalci lotevajo razločevanja sovražnega govora s pomočjo globokih nevronskih mrež [15], največkrat konvolucijskih (CNN)

in rekurenčnih nevronskih mreže (RNN) [10]. Druga pogosto uporabljena tehnika je metoda podpornih vektorjev (Support Vector Machines - SVM) [3].

Ob pretvorbi besed nekega jezika v vektorsko obliko dobimo visokodimenzionalen vektorski prostor. Vektorski prostor nekega jezika se razlikuje od prostora drugega jezika. Če nam uspe oba vektorska prostora poravnati, lahko napovedni model napoveduje tudi v drugem jeziku; če sta si jezika podobna (npr. hrvaščina in slovenščina) to velja še v večji meri [21, 2, 9].

V našem delu napovedne modele, ki smo jih naučili na nekem jeziku, uporabimo za napovedovanje sovražnega govora v drugem jeziku, pri tem pa sistematično analiziramo različne količine potrebnih učnih podatkov iz obeh jezikov.

1.2 Napoved vsebine po poglavjih

V 2. poglavju bomo opisali, kakšna je arhitektura napovednega modela BERT, kako poteka predhodno učenje, kako v našem primeru poteka dodatno učenje in na kratko opisali oba uporabljena modela, večjezikovnega in trojezikovnega. V 3. poglavju bomo opisali kakšne so strukture podatkovnih množic, kako so bile pridobljene, kakšna je pravilna oblika vhodnih in izhodnih podatkov modela BERT ter kako smo pripravili podatke za uporabo. V 4. poglavju bomo opisali, kakšne poskuse smo opravili in kakšni so bili rezultati poskusov. V zadnjem, 5. poglavju, bomo povzeli rezultate poskusov in podali nekaj predlogov za možne izboljšave.

Poglavje 2

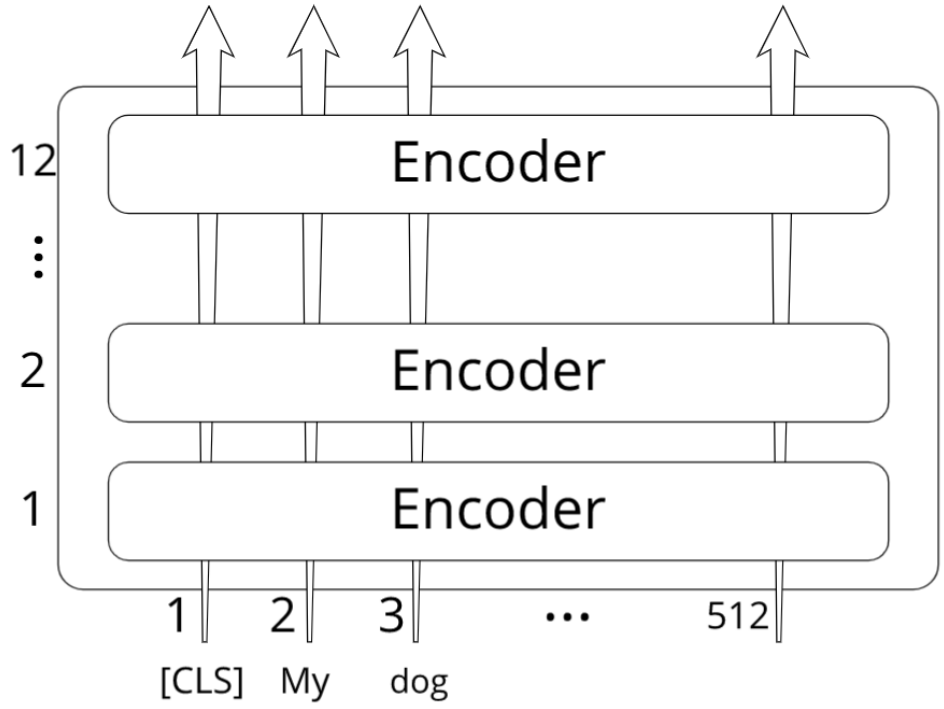
Metodologija

Poglavje vsebuje razlago metod, ki smo jih uporabili za odkrivanje in analizo sovražnega govora.

2.1 Napovedni modeli

V raziskavi smo uporabili dva različna modela globokih nevronskih mrež tipa BERT [20, 6] in naredili primerjavo njune klasifikacijske točnosti.

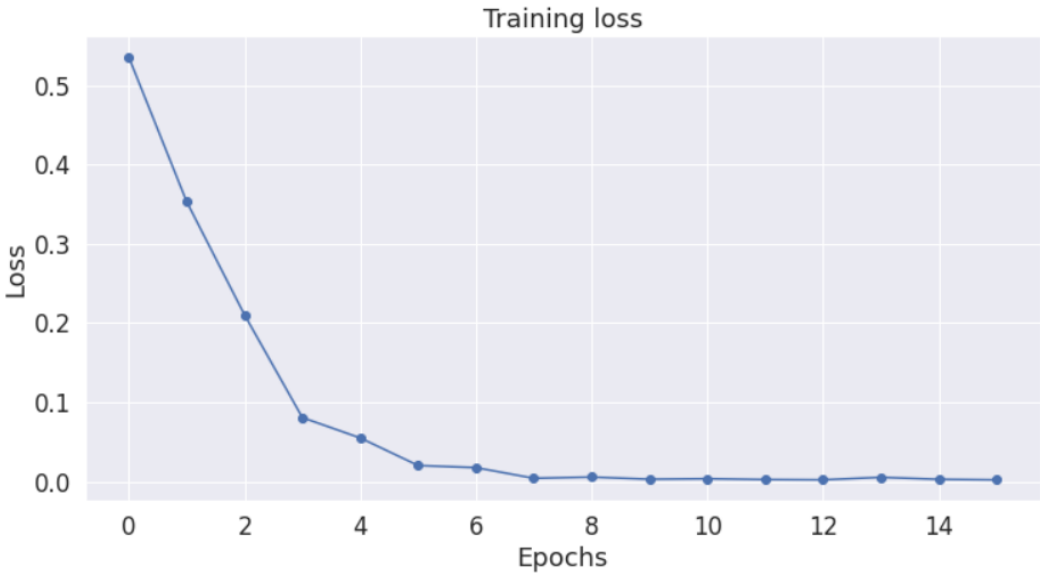
Model BERT je zgrajen iz večih plasti nevronskih celic tipa transformer in je natančno opisan v članku Vaswani in sod. [18]. Ker je arhitektura obsežna in zapletena, se ne bomo spuščali v njene podrobnosti. Originalno sta obstajali dve implementaciji arhitekture, prva je $\text{BERT}_{\text{BASE}}$ (slika 2.1), ki vsebuje 12 plasti oz. blokov transformerjev, 768 skritih nevronov v eni plasti, 12 glav mehanizma samopozornosti in skupaj 110 milijonov parametrov. Druga implementacija je $\text{BERT}_{\text{LARGE}}$, ki vsebuje 24 plasti, 1024 skritih nevronov v eni plasti, 16 glav mehanizma samopozornosti in 340 milijonov parametrov. Oba naša uporabljena modela temeljita na $\text{BERT}_{\text{BASE}}$ arhitekturi.



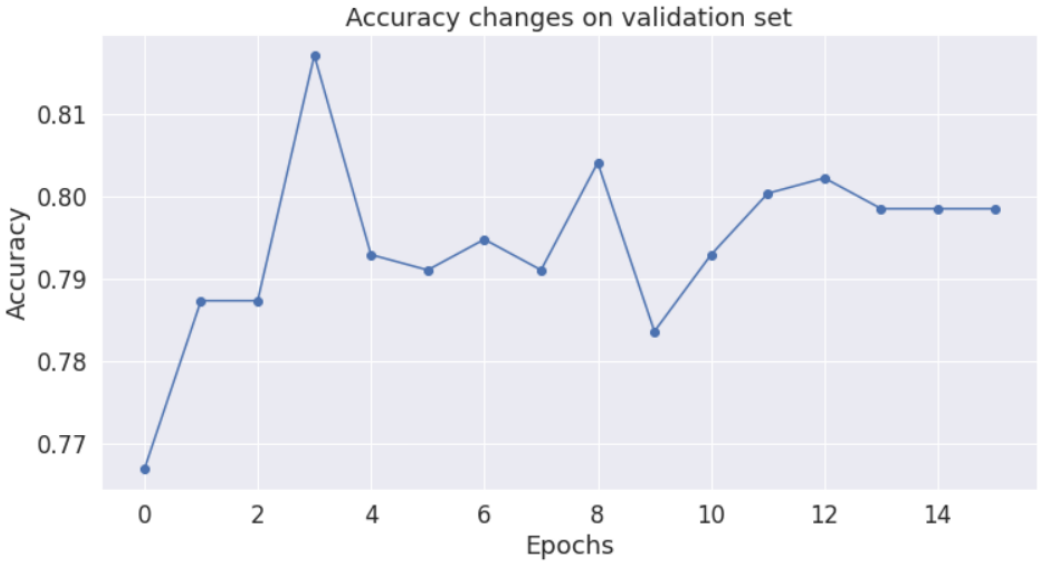
Slika 2.1: Arhitektura modela BERT

Model BERT učimo v dveh korakih. Prvi korak je predhodno učenje (angl. pre-training), drugi korak pa dodatno učenje (angl. fine-tuning). V fazi predhodnega učenja se model uči splošnih značilnosti jezika na neoznačenih podatkih, kot nalogo pa uporablja maskirani jezikovni model. V fazi dodatnega učenja je model inicializiran s parametri iz predhodnega učenja in se dodatno nauči na označenih podatkih.

Za potrebe dodatnega učenja in medjezikovnega prenosa napovednih modelov smo uporabili storitev Google Colaboratory, kjer so na voljo GPU za učenje velikih napovednih modelov. Učenje je potekalo v več iteracijah, po vsaki iteraciji smo izračunali izgubo, ki je nastala med učenjem na učni množici in klasifikacijsko točnost na validacijski množici. Obe izračunani vrednosti smo si shranili, da smo lahko izrisali grafa (slika 2.2 in slika 2.3), kjer so vidne spremembe točnosti oz. izgube skozi iteracije.



Slika 2.2: Spreminjanje izgube na učni množici med učenjem modela.

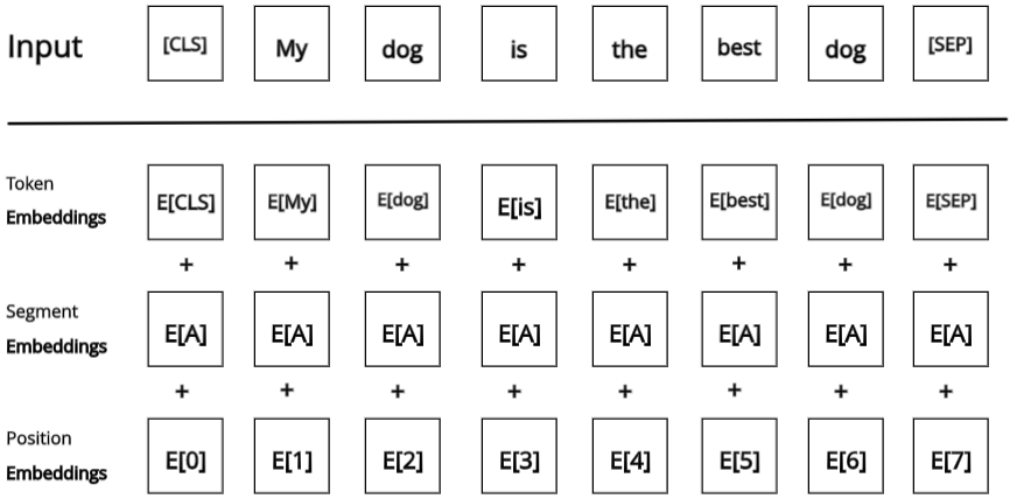


Slika 2.3: Spreminjanje klasifikacijske točnosti na validacijski množici med učenjem modela.

Kot ustavitveni pogoj učenja smo uporabili trikratni zaporedni padec oz. enakost klasifikacijske točnosti na validacijski množici. Model, ki je imel

najvišjo klasifikacijsko točnost na validacijski množici, smo shranili za poskus medjezikovnega prenosa napovednega modela.

Za uspešno učenje modela morajo biti vhodni in izhodni podatki v vektorski obliki. Za preslikavo v vektorsko obliko poskrbi tokenizator (angl. tokenizer), ki zaporedja besed razdeli na posamezne besede, ki se nahajajo v slovarju. Če se ne nahajajo v slovarju, jih tokenizator razdeli na krajše delčke [22], ki jih je s pomočjo slovarja moč preslikati v vektorsko obliko. Na sliki 2.4 je opisan postopek viden v vrstici, ki je označena s *Token Embeddings*. Vse besede, ki pripadajo istemu stavku, dobijo isto oznako, da lahko tokenizator loči med različnimi stavki, kar je vidno v vrstici, označeni s *Segment Embeddings*. V zadnji vrstici, označeni s *Position Embeddings*, določimo pozicijo vsake besede. Vidimo, da vsaki besedi, tudi če sta isti, v stavku pripada drugačna vektorska preslikava.



Slika 2.4: Preslikava besedila v vektorsko obliko.

2.1.1 Večjezikovni BERT

V knjižnici transformers [20] sta na voljo dva vnaprej naučena večjezikovna modela BERT. Uporabili smo model *bert-base-multilingual-cased*, ki je novejši

in natančnejši od drugega modela *bert-base-multilingual-uncased*. Napovedni model je bil predhodno naučen na 104 jezikih, ki imajo največ besedil na Wikipediji, med drugim tudi na slovenščini.

2.1.2 Trojezikovni BERT

Trojezikovni model je bil vnaprej naučen na UL FRI [17] na angleškem, hrvaškem in slovenskem jeziku. Za učenje so potrebovali 44 iteracij in skupno 3,96 milijona korakov, kar je trajalo približno 3 tedne. Model je namenjen medjezikovnim prenosom iz angleščine v slovenščino in hrvaščino.

Poglavje 3

Podatkovne množice

V tem poglavju opišemo uporabljene podatkovne množice za sovražni govor in njihovo pripravo za učenje.

3.1 Opis podatkov

Podatkovne množice v tej nalogi smo pridobili iz raziskave dr. Ljubešića in sodelavcev [8], v kateri so zbrali komentarje iz objav na družbenem omrežju Facebook npr. 24ur.com, BBC in The Guardian. Ključna prednost teh podatkov pred ostalimi podatki s področja sovražnega govora je, da so množice v različnih jezikih označene na enoten način.

Vsaka podatkovna množica je bila prvotno v JSON obliki in je vsebovala objave novic zbranih iz različnih virov. Njihovo število lahko vidimo v tabeli 3.1. Za vsako novico smo izluščili več komentarjev, njihovo število je v tabeli 3.3.

	ANG	HRV	SLO
LGBT	14	22	93
migranti	16	57	30

Tabela 3.1: Število objav novic za vsako posamezno temo.

Struktura novice in komentarja sta vidni na sliki 3.1 ter na sliki 3.2.

Združili smo komentarje vseh novic in iz njih ustvarili novo podatkovno množico. Namesto celotnih komentarjev smo uporabili samo besedilo in razred vsakega komentarja. Na sliki 3.2 lahko vidimo, da sta ti dve polji "text"in "type_mode".


```
{
  "title": "Tretjina Slovencev bi sirskim beguncem odprla vrata svojega doma",
  "post": "Ali bi sprejeli begunca pod svojo streho? #begunci",
  "url": "https://www.facebook.com/325510355503_10156064017960504",
  "source": "SiOL.net.Novice",
  "timestamp": "2015-09-15T09:18:22+0000",
  "comments": [
    {
      "text": "Lorem ipsum hate speech",
      "user_name": "Domen Novak",
      "user_id": "760129184149826",
      ...
    },
    ...
  ],
}
```


Slika 3.1: Struktura novice v JSON obliki


```
{
  "text": "Lorem ipsum hate speech",
    "user_name": "Domen Novak",
    "user_id": "760129184149826",
    "url": "https://www.facebook.com/325510355503_10156064017960504_10156064022950504",
    "timestamp": "2015-09-15T09:21:29+0000",
    "annotations": [
      {
          "type": "Background offensive",
          "target": "Migrants",
          "annotator": "26"
      },
      ...
    ],
    "type_mode": "Background offensive",
    "type_mode_prob": 0.9444444444444444,
    "target_mode": "Migrants",
    "target_mode_prob": 0.9444444444444444,
    "type_professional1": "Background offensive",
    "target_professional1": "Migrants",
    "type_professional2": "Background offensive",
    "target_professional2": "Migrants"
}
```


Slika 3.2: Struktura komentarja v JSON obliki

Če je bil komentar označen v podatkih kot sovražen, neprimeren ali žaljiv,

je v novi množici pripadel razredu 0, če pa je bila vsebina komentarja primerna, je pripadel razredu 1. Tako smo ločili komentarje na dva razreda.

Besedilo komentarjev je bilo potrebno pretvoriti v vektorsko obliko, ki jo napovedni model sprejme kot vhod [20]. Za model BERT je beseda vsebovana v slovarju ali pa se razbije na krajše delčke, ki obstajajo v slovarju in jih je možno preslikati v vhodni vektor. Na začetek vsakega komentarja je bilo pred preslikavo potrebno dodati poseben znak '[CLS]', na konec komentarja pa znak '[SEP]', da lahko napovedni model loči med različnimi komentarji. Zadnji korak priprave podatkov je zahteval enako dolžino vseh komentarjev, zato je bilo potrebno vsa zaporedja besed podaljšati na enako število besed s posebnimi znaki '[PAD]'. Tako so pri migrantski temi vsi komentarji vsebovali 505 besed, pri LGBT temi pa 499 besed, kar vidimo v tabeli 3.2. V našem primeru smo ignorirali komentarje, ki so vsebovali več kot 512 besed [6] in takšne, ki niso pripadali nobenemu razredu.

	ANG	HRV	SLO
LGBT	446	481	499
migranti	505	499	461

Tabela 3.2: Največje število besed v komentarju.

Podatki so ločeni glede na temo (LGBT in migranti) in glede na jezik (angleščina, hrvaščina in slovenščina), tako da smo skupno dobili 6 ločenih podatkovnih množic. Število komentarjev, ki jih vsebuje vsaka množica, je prikazano v spodnji tabeli 3.3.

	ANG	HRV	SLO
LGBT	5895	5681	4454
migranti	5825	5359	6460

Tabela 3.3: Število komentarjev v podatkovnih množicah.

Za uspešno učenje je koristno, da so množice uravnotežene, kar pomeni, da je število komentarjev, ki so označeni za sovražne, približno enako številu

komentarjev, ki niso označeni kot sovražni. Uravnoteženost je vidna v tabeli 3.4.

	ANG	HRV	SLO
LGBT	25,3	63,3	52,9
migranti	46,0	51,4	55,5

Tabela 3.4: Delež sovražnih komentarjev v %

3.2 Predpriprava podatkov

Za učenje smo uporabili python knjižnico PyTorch [1], zato morajo biti vhodni in izhodni podatki za uporabo modela BERT v posebnem formatu, imenovanem tensor. Tensor se razlikuje od navadne tabele v tem, da se lahko uporablja za računanje na CPU ali na GPU.

Vhodni podatki za model BERT morajo biti v vektorski obliki ter enake dolžine. Za preslikavo iz besedila v vektorsko obliko smo uporabili BertTokenizer [16], ki nam hkrati stavke še tokenizira. Za podaljševanje vektorjev na enako dolžino smo si pomagali s knjižnico Keras [4].

Po končani preslikavi besedila v vektorsko obliko smo vse komentarje razdelili na učno, validacijsko in testno množico. Učna množica je predstavljala 80%, testna in validacijska pa sta predstavljali vsaka po 10% celotne množice komentarjev. Zadnji korak priprave podatkov je pretvorba v format tensor, za kar poskrbi knjižnica PyTorch.

Poglavje 4

Evalvacija in rezultati

V tem poglavju bomo opisali opravljene poskuse in njihove rezultate. V podpoglavju 4.1 bomo navedli imena vseh knjižnic, s katerimi smo si pomagali pri poskusih in razlago enačbe za računanje klasifikacijske točnosti. V naslednjem podpoglavju 4.2 bomo razložili, kako deluje privzeti klasifikator, ki nam bo služil kot osnova za primerjavo z nevronsko mrežo tipa BERT. Podpoglavje 4.3 vsebuje opis poskusov, kjer smo nevronsko mrežo BERT naučili na določenem jeziku, izračunali klasifikacijsko točnost na tem istem jeziku ter naredili primerjavo s privzetim klasifikatorjem. V podpoglavju 4.4 smo že naučene napovedne modele BERT iz prejšnjega podpoglavja uporabili za napovedovanje sovražnega govora na jeziku, na katerem model ni bil naučen in naredili primerjavo izračunanih klasifikacijskih točnosti. V podpoglavju 4.5 smo že naučen napovedni model BERT iz podpoglavja 4.3 dodatno naučili z različnimi količinami podatkov na jeziku, na katerem smo potem izračunali in primerjali klasifikacijske točnosti. Zadnje podpoglavje 4.6 vsebuje povzetek in dodatno obrazložitev vseh izračunanih klasifikacijskih točnosti.

4.1 Implementacija poskusov

Vsa koda je bila napisana v programskem jeziku Python. Za inicializacijo napovednega modela smo uporabili knjižnico transformers [20] in pytorch-pretrained-bert [16], ki nam omogočata uporabo že delno naučenih mode-

lov in tokenizatorjev. Za učenje modela na GPU smo uporabili knjižnico PyTorch [1]. Pri predprocesiranju podatkov smo si pomagali s knjižnicami NumPy [11], json, csv, Keras [4] in Pandas [19]. Pri delitvi podatkov na učno, validacijsko in testno množico smo uporabili sklearn [12], za risanje grafov pa matplotlib [7].

Zaradi nedostopnosti GPU smo modele BERT učili s pomočjo storitve Google Colaboratory, ki omogoča učenje velikih napovednih modelov. Dodatno učenje modela na GPU Tesla P100-PCIE-16GB je v povprečju potrebovalo 15 iteracij, vsaka iteracija učenja pa je trajala približno 5 minut, skupno 75 minut.

Za primerjavo napovednih modelov smo uporabili klasifikacijsko točnost, ki predstavlja delež vseh primerov, ki so bili pravilno klasificirani. Izračunamo jo tako, da število pravilno klasificiranih primerov delimo s številom vseh primerov. Izračunana vrednost bo vedno med 0 in 1, zato jo lahko enostavno izrazimo z odstotki.

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

Razlaga vrednosti v enačbi:

- TP - število pravilno pozitivno klasificiranih primerov
- TN - število pravilno negativno klasificiranih primerov
- FP - število napačno pozitivno klasificiranih primerov
- FN - število napačno negativno klasificiranih primerov

Naredili smo 4 različne vrste poskusov klasifikacije sovražnega govora, ki jih bomo bolj podrobno razložili.

4.2 Učenje in klasifikacija na istem jeziku s privzetim klasifikatorjem

Privzeti klasifikator (angl. dummy classifier [12]) je klasifikator, ki za napovedovanje uporablja najverjetnejšo oznako. Uporaben je kot osnova za primerjavo z drugimi, bolj kompleksnimi klasifikatorji.

Možnih je več osnovnih klasifikatorjev. Uporabili smo:

- stratificiranega (angl. stratified) - napovedovanje glede na porazdelitev razredov učne množice.
- enakomernega (angl. uniform) - za vsak primer naključno napove en razred, tako da je v vseh razredih približno enako število primerov.

	ANG	HRV	SLO
LGBT	62,6	54,1	49,0
migranti	52,8	52,2	48,6

Tabela 4.1: Klasifikacijska točnost stratificiranega privzetega klasifikatorja v %.

	ANG	HRV	SLO
LGBT	51,6	51,7	49,7
migranti	51,5	51,5	51,2

Tabela 4.2: Klasifikacijska točnost enakomernega privzetega klasifikatorja v %.

4.3 Učenje in klasifikacija na istem jeziku z nevronsko mrežo BERT

Naredili smo pet poskusov, kjer smo vnaprej naučen model BERT inicializirali s pomočjo knjižnice transformers [20]. Podatke smo razdelili na učno,

validacijsko in testno množico kot je opisano v podpoglavju 3.2 in začeli z dodatnim učenjem na celotni učni množici nekega jezika. Ko se je dodatno učenje zaključilo, smo izvedli še napovedovanje na testni množici tega istega jezika.

Izračunali smo klasifikacijsko točnost in si vrednost shranili, da smo lahko potem izračunali povprečje klasifikacijskih točnosti petih poskusov in na tak način dobili zanesljivejši rezultat. Naučen napovedni model smo shranili za poskuse medjezikovnega prenosa. Enak postopek smo za oba modela BERT, za obe temi in za vsak jezik ponovili petkrat, skupno smo naredili šestdeset poskusov.

	ANG	HRV	SLO
LGBT	83,0	77,7	70,4
migranti	74,9	71,8	73,9

Tabela 4.3: Klasifikacijska točnost večjezikovnega modela, naučenega in testiranega na istem jeziku, v %.

	ANG	HRV	SLO
LGBT	83,4	82,5	75,7
migranti	73,7	76,5	76,7

Tabela 4.4: Klasifikacijska točnost trojezikovnega modela, naučenega in testiranega na istem jeziku, v %.

Ob primerjavi klasifikacijskih točnosti modela BERT in privzetega klasifikatorja vidimo, da se točnost zelo poveča. Če vzamemo primer teme LGBT v angleškem jeziku, lahko opazimo, da je točnost stratificiranega privzetega klasifikatorja 62,6%, točnost večjezikovnega modela BERT pa je 83,0%. Točnost napovedovanja se poveča za 20,4%. Podobna izboljšava točnosti je opazna tudi pri ostalih jezikih.

Iz tabel 4.3 in 4.4 lahko razberemo, da so klasifikacijske točnosti za angleški jezik približno enake pri obeh uporabljenih modelih. Klasifikacijske

točnosti za slovenski in hrvaški jezik so višje za približno 5% pri trojezikovnem modelu BERT.

4.4 Medjezikovni prenos napovednega modela BERT brez dodatnega učenja

Dokončno naučen napovedni model BERT smo inicializirali s pomočjo knjižnice transformers [20]. Model, ki je bil naučen na nekem jeziku (npr. na angleščini), smo uporabili za napovedovanje na nekem drugem jeziku (npr. na hrvaškem ali slovenskem jeziku), zato smo podatke tega drugega jezika razdelili na učno, validacijsko in testno množico, kot je opisano v podpoglavju 3.2 in brez dodatnega učenja izvedli napovedovanje na testni množici. Po končanem testiranju smo izračunali klasifikacijsko točnost. Enak postopek smo ponovili za oba modela BERT in za vsak jezik.

	večjezikovni model naučen na			trojezikovni model naučen na		
	SLO	ANG	HRV	SLO	ANG	HRV
LGBT	70,4	49,3	66,7	75,7	46,1	47,0
migranti	73,9	42,6	65,1	76,7	42,4	42,8

Tabela 4.5: Klasifikacijska točnost napovednega modela testiranega na slovenščini v %.

	večjezikovni model naučen na			trijezikovni model naučen na		
	HRV	ANG	SLO	HRV	ANG	SLO
LGBT	77,7	39,6	69,1	82,5	35,6	44,8
migranti	71,8	49,6	54,7	76,5	47,8	47,8

Tabela 4.6: Klasifikacijska točnost napovednega modela testiranega na hrvaščini v %.

	večjezikovni model naučen na			trijezikovni model naučen na		
	ANG	SLO	HRV	ANG	SLO	HRV
LGBT	83,0	72,2	76,0	83,4	65,7	72,4
migranti	74,9	54,2	58,8	73,7	53,4	54,2

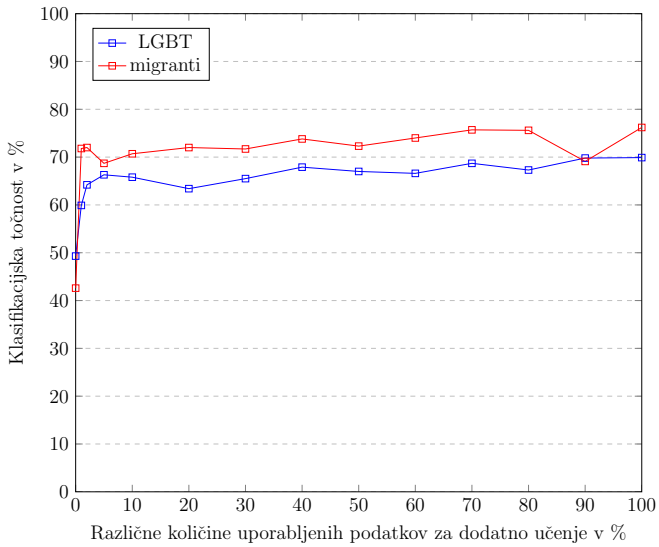
Tabela 4.7: Klasifikacijska točnost napovednega modela testiranega na angleščini v %.

Vsebina tabel 4.5, 4.6 in 4.7 nam razkrije, da ima vsak naučen večjezikovni model večjo klasifikacijsko točnost kot trojezikovni model, ko ga prenesemo na drug jezik za napovedovanje. Na primer večjezikovni model za migrantsko temo naučen na hrvaškem jeziku napove pravilno v 65,1% primerov na slovenskem jeziku, medtem ko trojezikovni model za isto temo naučen na hrvaškem jeziku napove pravilno v samo 42,8% primerov na slovenskem jeziku.

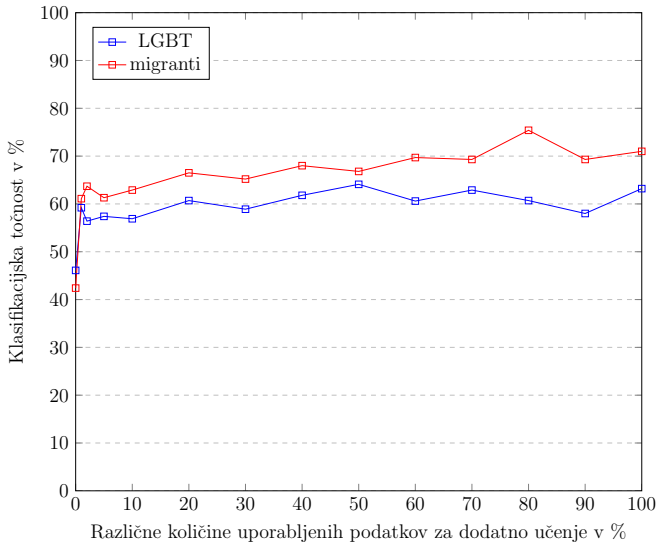
4.5 Medjezikovni prenos napovednega modela BERT z dodatnim učenjem

Dokončno naučen napovedni model BERT smo inicializirali s pomočjo knjižnice transformers [20]. Model, ki je bil naučen na nekem jeziku (npr. na angleškem jeziku), smo uporabili za napovedovanje na nekem drugem jeziku (npr. na slovenskem jeziku), zato smo podatke tega drugega jezika razdelili na učno, validacijsko in testno množico, kot je opisano v podpoglavju 3.2.

Za dodatno učenje napovednega modela smo učno množico razdelili na različno velike množice, po končanem dodatnem učenju pa smo izvedli še napovedovanje na testni množici in izračunali klasifikacijsko točnost.



Slika 4.1: Klasifikacijska točnost večjezikovnega modela, naučenega na angleščini z različnimi količinami dodatnega učenja in testiranjem na slovenščini v %.



Slika 4.2: Klasifikacijska točnost trojezikovnega modela, naučenega na angleščini z različnimi količinami dodatnega učenja in testiranjem na slovenščini v %.

Iz slik 4.1 in 4.2 lahko razberemo, da je možno boljši medjezikovni prenos napovednega modela opraviti z večjezikovnim modelom. Klasifikacijska točnost večjezikovnega in trojezikovnega modela BERT, po dodatnem učenju na 1% velikosti učne množice, je zelo podobna, se razlikuje samo za 0,7%. Po dodatnem učenju na 50% velikosti učne množice, se točnost razlikuje za 2,9%, po dodatnem učenju na 100% velikosti učne množice pa se točnost razlikuje za 5,7%.

4.6 Povzetek vseh rezultatov

Na slikah 4.3, 4.4 in 4.5 so tabele, kjer so združeni rezultati vseh poskusov.

	Klasifikacijska točnost na angleščini	
Privzeti klasifikator naučen na	**LGBT (%)**	**migranti (%)**
angleščini (stratificiran)	62,6	52,8
angleščini (enakomerni)	51,6	51,5
Večjezikovni model naučen na		
angleščini	**83,0**	**74,9**
slovenščini	**72,2**	**54,2**
hrvaščini	**76,0**	**58,8**

	Klasifikacijska točnost na angleščini	
Privzeti klasifikator naučen na	**LGBT (%)**	**migranti (%)**
angleščini (stratificiran)	62,6	52,8
angleščini (enakomerni)	51,6	51,5
Trojezikovni model naučen na		
angleščini	**83,4**	**73,7**
slovenščini	**65,7**	**53,4**
hrvaščini	**72,4**	**54,2**

Slika 4.3: Povzetek vseh rezultatov poskusov, kjer je bila uporabljena angleška testna množica

	Klasifikacijska točnost na hrvaščini	
Privzeti klasifikator naučen na	**LGBT (%)**	**migranti (%)**
hrvaščini (stratificiran)	54,1	52,2
hrvaščini (enakomerni)	51,7	51,5
Večjezikovni model naučen na		
hrvaščini	**77,7**	**71,8**
angleščini	**39,6**	**49,6**
slovenščini	**69,1**	**54,7**

	Klasifikacijska točnost na hrvaščini	
Privzeti klasifikator naučen na	**LGBT (%)**	**migranti (%)**
hrvaščini (stratificiran)	54,1	52,2
hrvaščini (enakomerni)	51,7	51,5
Trojezikovni model naučen na		
hrvaščini	**82,5**	**76,5**
angleščini	**35,6**	**47,8**
slovenščini	**44,8**	**47,8**

Slika 4.4: Povzetek vseh rezultatov poskusov, kjer je bila uporabljena hrvaška testna množica

	Klasifikacijska točnost na slovenščini	
Privzeti klasifikator naučen na	**LGBT (%)**	**migranti (%)**
slovenščini (stratificiran)	49,0	48,6
slovenščini (enakomerni)	49,7	51,2
Večjezikovni model naučen na		
slovenščini	**70,4**	**73,9**
hrvaščini	**66,7**	**65,1**
angleščini	**49,3**	**42,6**
angleščini z 1% slovenščine	59,9	71,8
angleščini z 2% slovenščine	64,2	72,0
angleščini z 5% slovenščine	66,3	68,7
angleščini z 10% slovenščine	65,8	70,7
angleščini z 20% slovenščine	63,4	72,0
angleščini z 30% slovenščine	65,5	71,7
angleščini z 40% slovenščine	67,9	73,8
angleščini z 50% slovenščine	67,0	72,3
angleščini z 60% slovenščine	66,6	74,0
angleščini z 70% slovenščine	68,7	75,7
angleščini z 80% slovenščine	67,3	75,6
angleščini z 90% slovenščine	69,8	69,1
angleščini z 100% slovenščine	69,9	76,2

	Klasifikacijska točnost na slovenščini	
Privzeti klasifikator naučen na	**LGBT (%)**	**migranti (%)**
slovenščini (stratificiran)	49,0	48,6
slovenščini (enakomerni)	49,7	51,2
Trojezikovni model naučen na		
slovenščini	**75,7**	**76,7**
hrvaščini	**47,0**	**42,8**
angleščini	**46,1**	**42,4**
angleščini z 1% slovenščine	59,2	61,1
angleščini z 2% slovenščine	56,4	63,7
angleščini z 5% slovenščine	57,4	61,3
angleščini z 10% slovenščine	56,9	62,9
angleščini z 20% slovenščine	60,7	66,5
angleščini z 30% slovenščine	58,9	65,2
angleščini z 40% slovenščine	61,8	68,0
angleščini z 50% slovenščine	64,1	66,8
angleščini z 60% slovenščine	60,6	69,7
angleščini z 70% slovenščine	62,9	69,3
angleščini z 80% slovenščine	60,7	75,4
angleščini z 90% slovenščine	58,0	69,3
angleščini z 100% slovenščine	63,2	71,0

Slika 4.5: Povzetek vseh rezultatov poskusov, kjer je bila uporabljena slovenska testna množica.

Ob primerjavi klasifikacijskih točnosti večjezikovnega in trojezikovnega modela, ki so v tabelah na slikah 4.3, 4.4 in 4.5, lahko vidimo, da trojezikovni napovedni model za približno 5% bolje napoveduje na jeziku, na katerem je bil prvotno naučen. Ta trditev ne velja popolnoma za napovedni model, ki je bil naučen na angleškem jeziku, saj imata v tem primeru večjezikovni in trojezikovni model skoraj enako klasifikacijsko točnost. Pri poskusu medjezikovnih prenosov napovednih modelov brez dodatnega učenja opazimo, da veliko bolje napoveduje večjezikovni napovedni model, v nekaterih primerih je razlika med klasifikacijskima točnostima skoraj 25%, v večini primerov pa je razlika manjša kot 10%. Medjezikovni prenos napovednih modelov z različnimi količinami podatkov za dodatno učenje nam razkrije, da v takšnih primerih večjezikovni napovedni model bolje napoveduje, vendar razlika ni tako velika kot pri medjezikovnem prenosu brez dodatnega učenja in znaša približno 5%.

Poglavje 5

Zaključek

V diplomski nalogi smo analizirali napovedni model na področju procesiranja naravnega jezika BERT, kot učno množico smo uporabili dve vrsti sovražnega govora iz treh jezikov. Naredili smo primerjavo med večjezikovnim in trojezikovnim modelom BERT brez ter z dodatnim učenjem. Na podlagi izračunanih klasifikacijskih točnosti smo ugotovili, da trojezikovni napovedni model brez dodatnega učenja bolje kot večjezikovni napovedni model brez dodatnega učenja napoveduje na jeziku, na katerem je bil prvotno naučen. Trojezikovni model v takšnem primeru napoveduje za približno 5% natančneje kot večjezikovni napovedni model. Večjezikovni napovedni model brez dodatnega učenja bolje kot trojezikovni napovedni model brez dodatnega učenja napoveduje na jeziku, na katerem ni bil prvotno naučen. Z njim tako naredimo boljši medjezikovni prenos napovednega modela. To se izkaže za resnično tudi v primeru medjezikovnega prenosa z dodatnim učenjem, kjer za dodatno učenje uporabimo različne količine podatkov. Večjezikovni model v tem primeru napoveduje za približno 5% natančneje kot trojezikovni model, v primeru brez dodatnega učenja pa je ta številka višja za med 3% in 25%.

Možnih je več izboljšav opravljenega dela. V eksperimente bi lahko vključili še kakšen jezik več. Uporabili smo samo angleščino, hrvaščino in slovenščino, ker podatkovne množice za druge jezike še niso bile pripravljene

ter dokončno označene. Lahko bi naredili tudi več poskusov medjezikovnega prenosa z dodatnim učenjem z različnimi količinami podatkov. Takšni poskusi so sicer časovno zelo zahtevni.

Razvita koda je dostopna v repozitoriju `https://github.com/zanpecovnik/diplomaPoskusi`.

Literatura

[1] Soumith Chintala Adam Paszke, Sam Gross and Gregory Chanan. Pytorch. `https://github.com/pytorch/pytorch`. Dostopano: 1. 5. 2020.

[2] Mikel Artetxe and Holger Schwenk. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610, 2019.

[3] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, page 759–760, 2017.

[4] François Chollet et al. Keras. `https://github.com/fchollet/keras`, 2015. Dostopano: 30. 4. 2020.

[5] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(76):2493–2537, 2011.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *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), pages 4171–4186, June 2019.

[7] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.

[8] Nikola Ljubesic, Darja Fiser, and Tomaz Erjavec. The FRENK datasets of socially unacceptable discourse in Slovene and English. *Proceedings of Text, Speech, and Dialogue*, TSD 2019.

[9] Rok Marinšek. Cross-lingual embeddings for hate speech detection in comments. Master's thesis, Univerza v Ljubljani, Fakulteta za računalništvo in informatiko, 2019.

[10] Kristian Miok, Dong Nguyen-Doan, Blaž Škrlj, Daniela Zaharie, and Marko Robnik-Šikonja. Prediction uncertainty estimation for hate speech classification. In *Statistical Language and Speech Processing*, pages 286–298, 2019.

[11] Travis E Oliphant. *A guide to NumPy*, volume 1. Trelgol Publishing USA, 2006.

[12] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[13] Georgios K. Pitsilis, Heri Ramampiaro, and Helge Langseth. Detecting offensive language in tweets using deep learning. *Applied Intelligence*, 48, 2018.

[14] Xin Rong. word2vec Parameter Learning Explained. *ArXiv*, abs/1411.2738, 2014.

[15] Anna Schmidt and Michael Wiegand. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, 2017.

[16] Tim Rault Thomas Wolf, Victor Sanh. Pytorch pretrained BERT: The big & extending repository of pretrained transformers. `https://pypi.org/project/pytorch-pretrained-bert/`. Dostopano: 30. 4. 2020.

[17] Matej Ulčar and Marko Robnik Šikonja. FinEst BERT and CroSloEngual BERT: less is more in multilingual models. Technical report, University of Ljubljana, 2020.

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008. 2017.

[19] Wes McKinney. Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference*, pages 56 – 61, 2010.

[20] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. Huggingface's transformers: State-of-the-art natural language processing. `https://huggingface.co/transformers/model_doc/bert.html#bertforsequenceclassification`, 2019. Dostopano: 30. 4. 2020.

[21] Shijie Wu and Mark Dredze. Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT. `https://arxiv.org/pdf/1904.09077.pdf`, 2019.

[22] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao,

Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. https://arxiv.org/pdf/1609.08144.pdf%20(7.pdf, 2016.

[23] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. Predicting the type and target of offensive posts in social media. https://arxiv.org/pdf/1902.09666.pdf, 2019.

[24] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). https://arxiv.org/pdf/1903.08983.pdf, 2019.