# Logično sklepanje v naravnem jeziku za slovenščino

*Tim KMECL*
Fakulteta za računalništvo in informatiko, Univerza v Ljubljani

*Marko ROBNIK-ŠIKONJA*
Fakulteta za računalništvo in informatiko, Univerza v Ljubljani

Na področju strojnega razumevanja naravnega jezika so v zadnjih letih najuspešnejši veliki jezikovni modeli. Pomemben problem s tega področja je logično sklepanje v naravnem jeziku, za reševanje katerega morajo modeli vsebovati dokaj široko splošno znanje, strojno generiranje razlag sklepov pa nam omogoča dodaten vpogled v njihovo delovanje.
Preizkusili smo različne pristope za logično sklepanje v naravnem jeziku za slovenščino. Uporabili smo dva slovenska velika jezikovna modela, SloBERTa in SloT5, in mnogo večji angleški jezikovni model GPT-3.5-turbo. Za učenje modelov smo uporabili slovensko podatkovno množico SI-NLI, strojno pa smo prevedli še 50.000 primerov iz angleške množice ESNLI.
Model SloBERTa, prilagojen na SI-NLI, doseže na testni množici SI-NLI klasifikacijsko točnost 73,2 %. Z vnaprejšnjim učenjem na prevodih ESNLI smo točnost izboljšali na 75,3 %. Ugotovili smo, da modeli delajo drugačne vrste napak kot ljudje in da slabo posplošujejo med različnimi domenami primerov. SloT5 smo na množici ESNLI prilagodili za generiranje razlag pri logičnem sklepanju. Ustreznih je manj kot tretjina razlag, pri čemer se model dobro nauči pogostih stavčnih oblik v razlagah, večinoma pa so pomensko nesmiselne. Predvidevamo, da so slovenski veliki jezikovni modeli z nekaj sto milijoni parametrov zmožni iskanja in uporabe jezikovnih vzorcev, njihovo poznavanje jezika pa ni povezano s poznavanjem resničnosti.

---

*Kmecl, T. et al.: Logično sklepanje v naravnem jeziku za slovenščino. Slovenščina 2.0, 12(1): 1–53.*

*1.01 Izvirni znanstveni članek / Original Scientific Article*

DOI: https://doi.org/10.4312/slo2.0.2024.1.1-53

Za uvrščanje primerov in generiranje razlag smo uporabili tudi večji model GPT-3.5-turbo. Pri učenju brez dodatnih primerov doseže na testni množici SI-NLI točnost 56,5 %, pri pravilno uvrščenih primerih pa je ustreznih 81 % razlag. V primerjavi z manjšimi slovenskimi modeli kaže ta model dokaj dobro razumevanje resničnosti, pri čemer pa ga omejuje slabše poznavanje slovenščine.

Ključne besede: logično sklepanje v naravnem jeziku, veliki jezikovni modeli, arhitektura transformer, SloBERTa, SloT5, GPT-3.5-turbo, ChatGPT, razlage, slovenščina, prilagajanje modelov

## 1 Uvod

Procesiranje naravnega jezika je vse od začetkov računalništva v sredini prejšnjega stoletja pomembno raziskovalno in aplikativno področje. Zajema številne probleme, ki so povezani z naravnim jezikom. Problemi so lahko klasifikacijske ali generativne narave. Pod klasifikacijske naloge spada na primer označevanje besednih vrst, razpoznava čustev v besedilu in zaznavanje neželene pošte. Pri generativnih problemih je izhod besedilo, kot na primer pri povzemanju vhodnega besedila in odgovarjanju na vprašanja. V zadnjih letih so na tem področju najuspešnejši veliki jezikovni modeli (angl. *large language models*, kratica *LLM*).

Veliki jezikovni modeli so posebna vrsta globokih nevronskih mrež z od nekaj deset milijonov do več sto milijard parametri, naučenih za modeliranje jezika, konkretno za napovedovanje naslednje ali manjkajoče besede v nizu. Učijo se na ogromnih korpusih besedil, najpogosteje s svetovnega spleta. Po splošnem učenju, ki jim zagotovi poznavanje jezika, jih lahko prilagodimo za različne naloge. V središče pozornosti širše javnosti so veliki jezikovni modeli vstopili novembra 2022 s predstavitvijo spletnega klepetalnega robota ChatGPT (OpenAI, 2022), ki temelji na velikem jezikovnem modelu GPT-3.5. Ker zna ChatGPT odgovarjati na vprašanja, slediti navodilom in ima znanje s praktično vseh področij človeškega delovanja, daje marsikomu vtis inteligence, podobne človeški.

Ljudje pri svojem razmišljanju in reševanju (jezikovnih) problemov med drugim uporabljamo zdravorazumsko sklepanje in poznavanje

sveta, pridobljeno skozi izkušnje in neposredno interakcijo s svetom. Po drugi strani pa se veliki jezikovni modeli učijo le na korpusih besedil in neposrednega dostopa do resničnega sveta nimajo. To poraja vprašanje, ali se modeli naučijo zgolj različnih jezikovnih vzorcev in hevristik, ki zadoščajo za reševanje različnih nalog, ali pa učenje le iz besedila omogoča globlje razumevanje resničnosti. Problem, ki nam lahko da vpogled v to, je logično sklepanje v naravnem jeziku (angl. *natural language inference*, kratica *NLI*).

Pri tipični formulaciji problema logičnega sklepanja v naravnem jeziku sta vhod dve povedi. Prvo imenujemo premisa, drugo hipoteza. Cilj je ugotoviti, v kakšnem logičnem razmerju sta podani povedi. Če je hipoteza logična posledica premise, oziroma če lahko ob predpostavki, da je premisa resnična, utemeljeno sklepamo na resničnost hipoteze, imenujemo to razmerje *implikacija* (angl. *entailment*). Če so informacije v hipotezi v nasprotju s tistimi v premisi, oziroma lahko iz resničnosti premise utemeljeno sklepamo na neresničost hipoteze, to imenujemo *kontradikcija* (angl. *contradiction*). Če pa iz premise ne moremo sklepati niti na resničnost niti na neresničnost hipoteze, torej če informacije v premisi hipoteze niti ne potrjujejo niti ne zavračajo, je tak primer *nevtralen* (angl. *neutral*). Tako formuliran problem je v osnovi klasifikacijski, lahko pa ga razširimo tako, da zahtevamo poleg klasifikacije (uvrščanja) še razlago zanjo. Če zmore jezikovni model, ki problem rešuje, odgovor utemeljiti, omogoča to dodaten vpogled v njegov pristop k reševanju problema.

Na primer, če je dana premisa *Mož in žena sedita v dnevni sobi* in hipoteza *V dnevni sobi sedita zakonca*, gre za implikacijo, ker sta mož in žena zakonca. Če bi bila hipoteza namesto tega *Dnevna soba je prazna*, bi bil to primer kontradikcije, ker sta glede na premiso v sobi dva človeka in zato ne more biti prazna. Če pa bi bila hipoteza *Ura je šest popoldne*, bi šlo za nevtralen primer, ker v premisi podatek o času ni podan, ljudje pa v dnevni sobi lahko sedijo kadarkoli.

V angleščini obstajajo številne podatkovne množice s primeri logičnega sklepanja v naravnem jeziku (Bowman idr., 2015; Camburu idr., 2018; McCoy idr., 2019). Prav tako so se z uporabo velikih jezikovnih modelov za reševanje tega problema ukvarjali mnogi raziskovalci (H. Liu idr., 2023; McCoy idr., 2019; Poth idr., 2021; Wang idr., 2021;

Zhong idr., 2023), tudi s strojnim generiranjem razlag zanje (Camburu idr., 2018; Kumar in Talukdar, 2020).

Logično sklepanje v naravnem jeziku za slovenščino je precej manj raziskano področje. Obstaja le ena izvirno slovenska podatkovna množica primerov logičnega sklepanja, imenovana SI-NLI (Klemen idr., 2022). Na spletni platformi SloBench (CJVT UL, 2023) sta objavljena dva rezultata vrednotenja na testni množici SI-NLI, oba pristopa uporabljata model SloBERTa (Ulčar in Robnik Šikonja, 2021). S strojnim generiranjem razlag pri logičnem sklepanju se v slovenščini ni ukvarjal še nihče.

Naš cilj je preizkusiti več pristopov za reševanje tega problema v slovenščini, pri tem pa testirati več velikih jezikovnih modelov, tako na slovenski podatkovni množici kot na strojnem prevodu angleške. Zanima nas, kako uspešni so različni modeli pri reševanju klasifikacijskega problema, kako dobro zmorejo generirati razlage in ali so sposobni posploševanja med različnimi domenami primerov istega problema. Na osnovi rezultatov skušamo poleg vrednotenja pristopov ugotoviti tudi, ali veliki jezikovni modeli, uporabni za slovenščino, premorejo dejansko razumevanje sveta ali pa se naučijo le jezikovnih vzorcev. Poskuse razdelimo v štiri sklope.

V prvem sklopu uporabimo slovensko množico SI-NLI za učenje klasifikacijskega modela SloBERTa. Preizkusiti želimo, kako zmogljiv je ta model, ali je število učnih primerov zadostno, kako na učenje vplivajo primeri, ki jih narobe uvrstijo ljudje, in ali so napake, ki jih naredi model, podobne človeškim.

V drugem sklopu poskusov se ukvarjamo z uporabo strojnega prevajanja iz angleščine, konkretno prevoda množice ESNLI (Camburu idr., 2018), in prenosom znanja. Model SloBERTa učimo na prevodih te množice in primerjamo rezultate s tistimi iz prejšnjega sklopa. Zanima nas, kako dobro jezikovni modeli posplošujejo med različnimi učnimi množicami, ali je strojno prevajanje lahko primeren nadomestek slovenskim podatkovnim množicam in ali lahko strojne prevode uporabimo za izboljšanje napovedovanja na SI-NLI.

Cilj tretjega sklopa je prilagajanje generativnega slovenskega modela SloT5 (Ulčar in Robnik-Šikonja, 2023) za generiranje razlag, ki jih kvalitativno ocenimo in s tem skušamo razložiti, kako jezikovni modeli rešujejo problem logičnega sklepanja.

Zadnji sklop poskusov temelji na uporabi angleškega modela GPT-3.5-turbo, ki poganja tudi ChatGPT. Uporabimo ga tako za klasifikacijo kot za generiranje razlag. Ugotoviti želimo, ali je ta model uporaben za slovenščino in ali mu nekaj redov velikosti več parametrov in učnih podatkov omogoča boljše razumevanje in logično sklepanje, tudi če ni bil naučen specifično za to nalogo.

Članek je razdeljen na sedem razdelkov. V razdelku 2 najprej predstavimo delovanje arhitekture transformer, ki je osnova za velike jezikovne modele. Zatem predstavimo tri jezikovne modele, ki smo jih uporabili za reševanje problema – slovenska modela SloBERTa in SloT5 ter angleški GPT-3.5-turbo. V razdelku 3 predstavimo in analiziramo slovensko podatkovno množico SI-NLI in angleško množico ESNLI ter opišemo njeno strojno prevajanje v slovenščino. V razdelku 4 opišemo postopek učenja modelov in izbiro parametrov zanje po štirih, prej opisanih sklopih poskusov. Predstavimo še evalvacijske metrike in način vrednotenja rezultatov. Rezultati vrednotenja so po sklopih podani v razdelku 5, ki vsebuje kvantitativne in kvalitativne ocene pristopov ter interpretacijo rezultatov. V razdelku 6 združimo najpomembnejše rezultate in jih postavimo v širši kontekst. Članek zaključimo v 7. razdelku s povzetkom narejenega in predlogi za nadaljnje delo ter izboljšave.

## 2 Veliki jezikovni modeli in predhodno delo

V tem razdelku predstavimo velike jezikovne modele, ki smo jih uporabili za logično sklepanje v naravnem jeziku. Začnemo s predstavitvijo modelov SloBERTa in SloT5, slovenskih različic modelov BERT in T5. Nato predstavimo največji model, ki smo ga uporabili, angleški GPT-3.5-turbo. Na koncu predstavimo še predhodno delo s področja uporabe velikih jezikovnih modelov za logično sklepanje v angleščini.

### 2.1 Modela BERT in SloBERTa

BERT (*Bidirectional Encoder Representations from Transformers*) (Devlin idr., 2019) je jezikovni model, ki temelji na uporabi kodirnika arhitekture nevronskih mrež transformer (Vaswani idr., 2017). Pri osnovni različici je kodirnik sestavljen iz 12 plasti in uporablja vektorje dimenzije 768. Skupno ima model 110 milijonov parametrov.

Za njegovo učenje je uporabljeno t. i. samonadzorovano učenje. Naučen je za dve nalogi, napovedovanje zaporednosti dveh stavkov in napovedovanje maskirane besede. Pri prvi je vhod v model sestavljen iz dveh stavkov, model pa mora ugotoviti, ali sta stavka zaporedna ali ne. Pri napovedovanju maskirane besede se 15 % členov na vhodu zamenja s posebnim členom [mask], model pa napoveduje člen, ki je bil tam pred zamenjavo. Z uporabo le druge naloge in večje količine podatkov kot pri originalnem modelu BERT je bila naučena izboljšana različica, imenovana RoBERTa (Y. Liu idr., 2019).

Učenje tako naučenih velikih jezikovnih modelov se kasneje nadaljuje na drugih nalogah, pri čemer je potrebna znatno manjša količina učnih primerov kot sicer, saj model že vsebuje neko razumevanje jezika, ki ga je treba zgolj prilagoditi konkretni nalogi. To imenujemo prilagoditev modela (angl. *fine-tuning*), za model, ki ga imamo za osnovo, pa rečemo, da je vnaprej naučen (angl. *pre-trained*).

Na enaki metodi učenja in enaki arhitekturi kot RoBERTa temelji slovenski model SloBERTa (Ulčar in Robnik-Šikonja, 2021). Ta model je bil naučen na korpusih slovenskih besedil, in sicer Gigafida 2.0 (Krek idr., 2020) (besedila iz knjig, revij, časopisov in interneta), Janes (Fišer idr., 2016) (besedila z družabnih omrežij), KAS (Erjavec idr., 2021) (akademska besedila), siParl (Pančur in Erjavec, 2020) (parlamentarni transkripti) in slWaC (Ljubešić in Erjavec, 2011) (slovenske spletne strani). Učna množica skupno vsebuje približno 3,4 milijarde besed oziroma 4,7 milijard členov. Učenje modela je trajalo 98 epoh (angl. *epochs*). Ta model smo uporabili kot vnaprej naučeni model in ga prilagodili za reševanje klasifikacijskega problema logičnega sklepanja v naravnem jeziku.

## 2.2 Modela T5 in SloT5

T5 (*Text-to-Text Transfer Transformer*) (Raffel idr., 2020) je družina modelov več velikosti, ki po zgradbi sledijo arhitekturi transformer in vsebujejo tako kodirnik kot dekodirnik. Ideja pristopa T5 je, da se vse naloge, tudi klasifikacijske, obravnava kot transformacijo enega besedila v drugo, npr. polnega besedila v povzetek ali besedila iz enega

jezika v drugega. Model so vnaprej učili na množici besedil s spleta, pri čemer je vhod modela besedilo, v katerem so nekatere besede zamenjane s posebno oznako, izhod pa mora biti besedilo, ki vsebuje manjkajoče besede v pravem zaporedju. Tako vnaprej naučene modele se nato prilagaja za različne naloge, kot so povzemanje besedil, prevajanje, razpoznava čustev in odgovarjanje na vprašanja.

Slovenska različica modela T5 se imenuje SloT5 (Ulčar in Robnik-Šikonja, 2023). Pravzaprav gre za dva modela, ki sta po zgradbi enaka dvema modeloma iz družine originalnih modelov T5. Manjši T5-sl-small ima 8 plasti v kodirniku in 8 v dekodirniku, skupaj 60 milijonov parametrov, večji model T5-sl-large pa ima 24 plasti v kodirniku in 24 v dekodirniku, skupaj 750 milijonov parametrov. Za učenje so bili uporabljeni isti korpusi kot za učenje modela SloBERTa. Manjši model so učili 5 epoh, večjega pa eno, pri čemer je učenje manjšega na 4 grafičnih karticah A100 s 40 GB pomnilnika trajalo 12 dni, večjega pa tri tedne.

Oba slovenska modela T5 smo uporabili za generiranje razlag za primere logičnega sklepanja.

## 2.3 Model GPT-3.5-turbo

Model GPT-3 (*Generative Pre-trained Transformer 3*) (Brown idr., 2020) temelji na uporabi dekodirnika arhitekture transformer, pri čemer je vhod v model že na začetku vstavljen v dekodirnik. Model je naučen na korpusu besedil s spleta CommonCrawl, ki vsebuje približno 400 milijard členov, z nalogo napovedovanja naslednje besede v besedilu. Ta model je bistveno večji od modelov, predstavljenih v prejšnjih razdelkih, saj vsebuje kar 175 milijard parametrov. Model je v lasti podjetja OpenAI in ni prosto dostopen, uporablja pa se ga lahko prek programskega vmesnika API, ki ga podjetje ponuja.

GPT-3 se lahko brez dodatnega učenja uporablja za številne naloge z dvema tehnikama. Pri prvi, imenovani učenje brez dodatnih primerov (angl. *zero-shot learning*), se modelu kot vhod posreduje navodilo oziroma opis naloge v naravnem jeziku in morebitni kontekst. To je lahko na primer besedilo nekega članka in navodilo, naj model članek povzame. Zaradi velikosti modela in velike količine

učnih podatkov, kar mu omogoča dobro posploševanje, zna model mnogim navodilom pravilno slediti in dobimo pravilen izhod. Za razliko od prilagoditve, ki se uporablja pri modelih arhitekture BERT in T5, pri učenju brez dodatnih primerov ne potrebujemo učnih primerov za želeno nalogo. Druga tehnika uporabe GPT modelov brez dodatnega učenja je uporaba nekaj dodatnih primerov (angl. *few-shot learning*), ki je podobna prejšnji, le da tu poleg navodil in konteksta na vhod dodamo še nekaj že rešenih primerov, ki lahko modelu olajšajo razumevanje naloge.

InstructGPT (Ouyang idr., 2022) je družina modelov, katerih največji je po velikosti enak GPT-3. Osnova je vnaprej naučen model GPT-3, s prilagoditvijo pa so izboljšali sposobnost modela za odgovarjanje na vprašanja in s tem tudi rezultate, ki jih lahko dosežemo z učenjem brez ali z nekaj dodatnimi primeri. Izboljšanje je bilo doseženo z uporabo spodbujevanega učenja s človeško povratno informacijo (angl. *reinforcement learning from human feedback*, RLHF). RLHF poteka v več korakih. Na začetku pripravimo podatkovno množico različnih navodil, za katera bi želeli, da jih model zna upoštevati pri uporabi, na primer ukaz, da napiše povzetek nekega besedila. Nato človeški demonstratorji spišejo odgovore za vsako od navodil, ki se uporabijo za nadzorovano učenje vnaprej naučenega modela (prilagoditev). V naslednjem koraku za vsako od navodil s tem modelom generiramo odgovor, človeški ocenjevalci odgovore ocenijo, ta informacija pa se nato uporabi za dodatno učenje z algoritmom spodbujevanega učenja, ki ne zahteva zvezne funkcije izgube.

GPT-3.5-turbo je eden najzmogljivejših modelov, ki jih ponuja OpenAI (OpenAI, 2023b). Je variacija zmogljivega modela InstructGPT, točni podatki o zgradbi in učenju pa niso objavljeni. To je model, ki poganja znani spletni vmesnik ChatGPT (OpenAI, 2022).

Čeprav model GPT-3 za razliko od predstavljenih v prejšnjih razdelkih nima slovenske različice, je med učnimi podatki tudi nekaj slovenščine. Modela ne prilagajamo za specifično nalogo, saj do njega nimamo neposrednega dostopa. Model GPT-3 nam služi za preizkus, ali več redov velikosti večje število parametrov in učnih podatkov pri vnaprejšnjem učenju lahko odtehta te pomanjkljivosti.

## 2.4 Uporaba jezikovnih modelov za logično sklepanje

Področje logičnega sklepanja v naravnem jeziku je v angleščini dobro raziskano. Zadnja leta najboljše rezultate dosegajo veliki jezikovni modeli, predstavljeni v prejšnjih razdelkih.

Poth idr. (2021) so vnaprej naučena modela BERT in RoBERTa prilagajali za reševanje različnih nalog, med drugim so testirali tudi več podatkovnih množic za logično sklepanje v naravnem jeziku. Z modelom RoBERTa, s katerim so dosegli boljše rezultate, so dosegli klasifikacijsko točnost 41,5 % na množici ANLI, 87,5 % na množici MNLI in 91,1 % na SNLI. Točnost na SNLI je bila do takrat najvišja dosežena.

Njihov rezultat so z lastnim velikim jezikovnim modelom nadgradili Wang idr. (2021). Uporabljeni model je po zgradbi podoben modelu RoBERTa, uporabili pa so drugačen način učenja. Z namenom povečanja učne množice so primere iz podatkovnih množic za druga področja razumevanja naravnega jezika, nepovezana z logičnim sklepanjem, reformulirali kot probleme logičnega sklepanja, te pa so nato uporabili za vnaprejšnje učenje modela. S to prilagoditvijo so na množici SNLI dosegli najboljši objavljen rezultat, 93,1 %.

Zhong idr. (2023) so primerjali zmogljivost modelov RoBERTa in GPT-3.5-turbo z učenjem brez ali z nekaj dodatnimi primeri. Z učenjem brez dodatnih primerov so presegli rezultate modela RoBERTa, njihova točnost je znašala 89,3 %, uporaba enega ali pet dodatnih primerov pa rezultata ni izboljšala. Liu idr. (2023) so primerjali modele RoBERTa, GPT-3.5-turbo in GPT-4 na podatkovnih množicah LogiQA in ReClor. Tudi oni so ugotovili, da točnost GPT-3.5-turbo za nekaj odstotkov preseže model RoBERTa, še večjo pa doseže GPT-4.

Z generiranjem razlag za logične sklepe so se ukvarjali Camburu idr. (2018). Pokazali so, da so pri tej nalogi veliki jezikovni modeli arhitekture transformer boljši od prejšnjih pristopov. Z učenjem lastnega modela na množici ESNLI so na tej množici dosegli klasifikacijsko točnost 81,7 %, pri čemer je pri pravilno klasificiranih primerih ustreznih 64 % razlag. Njihov pristop sta izboljšala Kumar in Talukdar (2020) z uporabo vnaprej naučenega modela GPT-2, predhodnika GPT-3.5-turbo. Preizkusila sta različne pristope, med drugim tudi z učenjem treh ločenih modelov za generiranje razlag za primere posameznih

razredov in dodatnega modela, ki na koncu izbere najboljšo. S tem sta izboljšala tako klasifikacijsko točnost kot tudi delež ustreznih razlag.

Problem NLI za slovenščino je manj raziskan. Na spletni platformi SloBench (CJVT UL, 2023) sta objavljena dva rezultata vrednotenja na testni množici SI-NLI, oba pristopa uporabljata prilagoditev modela SloBERTa. S strojnim generiranjem razlag pri logičnem sklepanju se v slovenščini ni ukvarjal še nihče.

## 3 Uporabljene podatkovne množice

V tem razdelku predstavimo podatkovni množici s področja logičnega sklepanja v naravnem jeziku, ki smo ju uporabili za učenje jezikovnih modelov. Najprej, v razdelku 3.1, opišemo in analiziramo sestavo slovenske množice SI-NLI, v razdelku 3.2 pa predstavimo še angleško množico ESNLI, ki vsebuje tudi razlage. Opišemo tudi postopek prevajanja te množice v slovenščino.

### 3.1 Slovenska množica SI-NLI

SI-NLI (Klemen idr., 2022) je podatkovna množica s skupno 5937 pari povedi v slovenščini. Vsak par vsebuje premiso in hipotezo ter je označen z oznako *entailment* (implikacija), *neutral* (nevtralno) ali *contradiction* (kontradikcija), ki označuje razmerje med povedmi. Pri konstrukciji množice so sodelovali človeški označevalci (angl. *annotators*). Množica je bila ustvarjena na osnovi povedi, ki se pojavijo v korpusu ccKres (Logar idr., 2013), ki vsebuje različne tipe besedil, kot so članki v časopisih in revijah, literarna in neliterarna besedila ter besedila z interneta. Označevalci so spreminjali hipoteze tako, da so ustrezale vsaki od možnih treh kategorij. Po en primer za vsako oznako je prikazan v Tabeli 1.

Delitev SI-NLI na učno, validacijsko in testno množico je predstavljena v Tabeli 2. Avtorji zagotavljajo, da so težji in lažji primeri enakomerno razporejeni med vsemi tremi množicami. Vse tri množice vsebujejo premiso in hipotezo za vsakega od primerov. Učna in testna množica vsebujeta poleg tega še oznako oz. klasifikacijo primera in po tri stolpce s klasifikacijami posameznih označevalcev, njihovimi identifikatorji ter morebitnimi komentarji. Testna množica tega ne vsebuje.

**Tabela 1:** *Trije primeri iz podatkovne množice SI-NLI, po eden za vsako oznako*

| | |
|---|---|
| Premisa | Med večletnimi širokolistnimi pleveli prevladujejo slak, osat, gabez in ščavje, najpogostejše večletne trave v koruzi pa so pirnica in divji sirek. |
| Hipoteza | Slak, gabez in osat spadajo med širokolistni plevel, enako tudi ščavje, na drugi strani pa med večletne trave v koruzi prištevamo pirnico in divji sirek. |
| Oznaka | *implikacija* |
| Premisa | "Res je," je zavzdihnila in se zravnala na sedežu. |
| Hipoteza | Z globokim poraženim izdihom je morala priznati, da dejstev ne gre zanikati. |
| Oznaka | *nevtralno* |
| Premisa | Večina delničarjev ga je potrdila za predsednika, Janez Pestotnik pa je postal novi predsednik nadzornega sveta Banke Karantanija. |
| Hipoteza | Ker se delničarji s svojimi glasovi niso uspeli uskladiti, je Banka Karantanija še vedno brez predsednika nadzornega sveta. |
| Oznaka | *kontradikcija* |

**Tabela 2:** *Število primerov v učni, validacisjki in testni množici SI-NLI, ESNLI in ESNLIsi*

| Podatkovna množica | *Učna* | *Validacijska* | *Testna* |
|---|---|---|---|
| SI-NLI | 4392 | 547 | 998 |
| ESNLI | 550.000 | 10.000 | 10.000 |
| ESNLIsi | 49.922 | 3000 | 3000 |

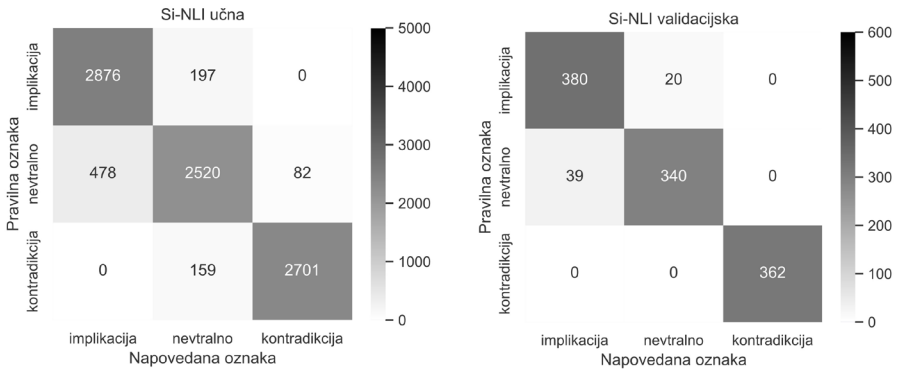
## Analiza

Analizirali smo sestavo učne in validacijske množice ter morebitne razlike med oznakami primerov in klasifikacijami posameznih označevalcev, torej primere, v katerih so označevalci glede na dodeljeno oznako storili napako.

V učni množici s 4392 primeri je 34,6 % vseh implikacij, 32,5 % nevtralnih in 33,0 % kontradikcij. Množica je torej skoraj uravnotežena. Podobno je pri validacijski množici, kjer 35,3 % primerov predstavlja implikacije, 31,6 % je nevtralnih in 33,1 % opisuje kontradikcije.

Vsakega od primerov v učni množici sta označila dva ali trije označevalci. Od oznake primera se vedno razlikuje največ ena oznaka označevalcev. 89,8 % oznak označevalcev se ujema z oznako primera (torej jih lahko privzamemo za pravilne oziroma ta odstotek predstavlja točnost človeških označevalcev), 79,1 % primerov pa je takih, da se nobena od oznak označevalcev ne razlikuje od oznake primera

(soglasni primeri). Na levi strani Slike 1 je predstavljena matrika zamenjav (angl. *confusion matrix*). Vidimo lahko, da ni napak, kjer bi označevalec zamenjal razreda implikacija in kontradikcija. Največ napak, skoraj tri četrtine, je posledica zamenjave med razredi implikacija in nevtralno, najpogostejša napaka pa je označitev nevtralnega primera kot primera implikacije.




**Slika 1:** Matriki zamenjav za oznake označevalcev na učni in validacijski množici SI-NLI.

V validacijski množici je soglasnih primerov 89,2 %, pravilnih oznak označevalcev pa 94,8 %. Struktura napak, prikazana na desni strani Slike 1, je podobna.

Edine napake so posledice zamenjav med razredoma implikacija in nevtralno, ki so tudi v učni množici najpogostejše, največ napak pa je tudi tu predstavljala označitev nevtralnega primera za implikacijo. V nadaljevanju skušamo ugotoviti, ali veliki jezikovni modeli delajo podobne tipe napak kot človeški označevalci in kako vpliva odstranitev primerov, kjer označevalci niso soglasni.

Metrika, ki je za oceno strinjanja med različnimi ocenjevalci bolj robustna od odstotka soglasnih primerov, je Cohenova kappa, ki ima nabor vrednosti med –1 in 1, kjer 0 pomeni količino ujemanj, ki jih lahko pripišemo naključju, 1 pa popolno ujemanje (McHugh, 2012). Podrobno analizo strinjanja med označevalci so opravili Klemen idr. (2024), ki so za vse pare izračunali Cohenovo kappo. Njihovo povprečje znaša 0,74, kar kaže na visoko konsistentnost med označevalci.

Ker oznake primerov testne množice niso javno objavljene, smo za naše potrebe validacijsko množico uporabili kot testno, učno množico pa smo razdelili na učno in validacijsko. V učni množici SI--NLI se večina premis ponovi trikrat, nobena premisa pa se ne pojavi hkrati v učni in validacijski množici, kar smo zagotovili tudi pri naši delitvi. Naključno smo izbrali 200 premis iz učne množice ter pripadajoče pare (590 primerov) shranili kot validacijsko množico, preostalih 3802 pa kot učno. V nadaljevanju se učna, validacijska in testna množica SI-NLI nanašajo na našo delitev, razen kjer je posebej navedeno drugače. Dodatno smo shranili tudi podmnožico naše učne množice, pri kateri so bili označevalci soglasni in vsebuje le 3015 primerov.

## 3.2 Množica z razlagami ESNLI

ESNLI (Camburu idr., 2018) je angleška podatkovna množica s 570 tisoč primeri (njena delitev je prikazana v Tabeli 2), od katerih vsak vsebuje premiso, hipotezo, eno od treh možnih oznak in razlago. Osnova te množice je angleška množica SNLI (Bowman idr., 2015), v kateri so premise opisi slik, človeški označevalci pa so jim dopisali po eno hipotezo za vsako od treh kategorij. ESNLI dodatno vsebuje še razlage, zakaj dani par premise in hipoteze pripada dodeljeni kategoriji. Razlage so pisali ljudje, želeli pa so, da so samozadostne, torej da za njihovo razumevanje ni potrebno predhodno prebrati premise in hipoteze. Primer take razlage je *Kdorkoli lahko plete, ne le ženske*, primer neustrezne pa *Ne moremo sklepati, da so to ženske* (Camburu idr., 2018). Pisci razlag so morali tudi označiti, katere besede v premisi in hipotezi so ključne za izbor dane oznake. Primeri v učni množici vsebujejo vsak po eno razlago, tisti v validacijski in testni pa po tri. Po en primer za vsako oznako (samo z eno razlago) iz množice je podan v Tabeli 3.

**Tabela 3:** *Trije primeri iz podatkovne množice ESNLI*

| | |
|---|---|
| Premisa | Mlad fant poljublja starca na čelo. |
| Hipoteza | Tam je fant, ki izkazuje naklonjenost staremu moškemu. |
| Razlaga | Poljubljanje je način izkazovanja naklonjenosti. |
| Oznaka | *implikacija* |
| Premisa | Črno-beli pes skače čez rdeče-belo palico. |
| Hipoteza | Pes spi. |
| Razlaga | Psi ne skačejo, ko spijo. |
| Oznaka | *kontradikcija* |
| Premisa | Mlada ženska, oblečena v belo jopico in kratke hlače, sedi na robu ploščadi, ki je dvignjena nad vodno površino. |
| Hipoteza | Ženska sedi na pomolu in opazuje sončni zahod. |
| Razlaga | Ni nujno, da ženska, ki sedi na ploščadi, opazuje sončni zahod. |
| Oznaka | *nevtralno* |

*Opomba.* Primeri so bili prevedeni v slovenščino, po eden za vsako oznako.

### *3.2.1 Prevajanje*

Ker se osredotočamo na logično sklepanje v slovenščini, smo del množice ESNLI strojno prevedli. Najprej smo se morali odločiti, kateri strojni prevajalnik uporabiti. Na voljo sta bili dve storitvi v oblaku, Google Prevajalnik (Google Prevajalnik, 2023) in prevajalnik DeepL (DeepL Translate API, 2023), ki ponujata programski vmesnik API. Dodatno smo imeli na voljo še prostodostopni strojni prevajalnik NeMo iz projekta RSDO (Lebar Bajec idr., 2022), naučen za prevajanje iz angleščine v slovenščino.

Za odločitev, katerega od prevajalnikov uporabiti, smo najprej prevedli manjše število primerov z vsakim od treh prevajalnikov, prevode uporabili za učenje klasifikacijskih modelov in na osnovi uspešnosti posameznega modela izbrali najboljšega. Učenje modelov na treh različnih prevodih je opisano v razdelku 4.2.1. Na podlagi njihovih rezultatov, predstavljenih v istem razdelku, smo za prevajanje večje množice uporabili Google Prevajalnik.

Naključno je bilo izbranih 50 tisoč primerov iz učne množice, tri tisoč iz validacijske in tri tisoč iz testne. Njihovi prevodi so predstavljali učno, validacijsko in testno množico za učenje modelov. V nekaterih izbranih

primerih učne množice so manjkale razlage ali hipoteze, poleg tega pa so bile nekatere zelo kratke razlage zelo slabe zaradi nesmiselnosti ali nerazumljivosti (npr. *his new it*, *man and guy*, *Runs in runs* ...), podobno pa so bile nesmiselne tudi nekatere zelo kratke hipoteze, ki so bile ali napačne oziroma zmotno skrajšane ali pa je šlo za primer le ene besede brez konteksta (npr. *Two wom*, *Fetch*, *f*, *A baby is* ...). Prevajanje takšnih primerov bi bilo nesmiselno, ročno preverjanje vsakega od njih pa preveč zamudno. Učno množico smo zato filtrirali tako, da smo odstranili vse primere, kjer je bila razlaga krajša od 14 znakov ali hipoteza krajša od 10 znakov. Naša ocena je namreč, da je večina tako odstranjenih primerov nesmiselna, če pa katero od mej zvišamo, bi odstranili dosti primerov, ki vsebujejo kratke, a ustrezne razlage oziroma hipoteze. Po filtriranju vsebuje učna množica 49.922 primerov.

Vse tri množice smo nato prevedli z Google Prevajalnikom, pri čemer smo za primere iz testne in validacijske množice prevedli le prvo od treh razlag. Množice, ki vsebujejo identifikator primera, izvirno angleško hipotezo, premiso in razlago ter njihove prevode v slovenščino, smo v formatu TSV objavili na spletu.[1] V nadaljevanju bo ta podatkovna množica imenovana ESNLIsi.

**Napake v prevodih** Po prevajanju smo naključno izbran vzorec prevodov še pregledali, da bi ocenili njihovo kvaliteto. Čeprav je lahko naša ocena nekoliko subjektivna, menimo, da je večina primerov prevedenih ustrezno. Pri nekaterih je prevedena formulacija rahlo nerodna, vendar kljub temu slovnično in pomensko pravilna. V nekaj odstotkih prevodov se pojavljajo napake. Primere napačnih prevodov prikazuje Tabela 4.

Prva vrsta napake je napačen prevod ene od besed, kar lahko vidimo v razlagi prvega primera. Tam je prevajalnik angleško besedo *batter* prevedel kot *udarec* namesto *udarjalec*. Razlaga je zato nerazumljiva, če pa do podobne napake pride v premisi ali hipotezi, je lahko nerazumljiv cel primer.

V drugem primeru je prikazano nekonsistentno prevajanje iste besede. Besedna zveza *gave up* je enkrat prevedena z glagolom *opustil*, drugič pa *obupal*. Razlaga je zato neustrezna, saj bi morala biti uporabljena ali enaka formulacija (kot to velja v angleškem izvirniku) ali pa bi morala

1 https://github.com/timkmecl/nli-slovene?tab=readme-ov-file#podatkovne-mno%C5%BEice

razlaga vsebovati še dodatno obrazložitev, da kdor obupa, nekaj opusti.

V nekaterih primerih prevajanje celo spremeni kategorijo, v katero bi bilo primer pravilno uvrstiti, kar je prikazano v tretjem primeru tabele. Tam je angleška besede "*man*" v premisi prevedena kot "*človek*" namesto "*moški*". Izvirnik je označen kot kontradikcija, saj je kraljica ženska in ne moški (razlaga je ustrezno prevedena). Ker pa je kraljica človek, v slovenskem prevodu pa je namesto angleške besede "*his*", uporabljene v izvirni premisi, spolno nevtralna oblika pridevnika "*svoj*", bi bila ustrezna klasifikacija slovenskega prevoda za nevtralni primer zaradi dejstva, da je kraljica tudi človek.

**Tabela 4:** *Primeri napak pri strojnem prevajanju podatkovne množice ESNLI*

| | |
|---|---|
| Oznaka | *implikacija* |
| Premisa | A man wearing a red shirt with the number 54 hits a baseball, while a catcher prepares to catch the ball. |
| Hipoteza | A red shirted batter hits the pitch. |
| Razlaga | A man in a red shirt who hits a baseball is known as a **batter**. |
| Premisa | Moški, oblečen v rdečo majico s številko 54, udari žogico za baseball, medtem ko se lovilec pripravlja, da ujame žogo. |
| Hipoteza | Udarjalec v rdeči majici pride na igrišče. |
| Razlaga | Moški v rdeči majici, ki udari bejzbolsko žogico, je znan kot **udarec**. |
| Oznaka | *kontradikcija* |
| Premisa | An elderly male is blowing air into an object. |
| Hipoteza | The elderly man **gave up** blowing air into the object |
| Razlaga | If he **gave up** how can he be blowing air. |
| Premisa | Starejši moški piha zrak v predmet. |
| Hipoteza | Starejši moški je **opustil** vpihovanje zraka v predmet |
| Razlaga | Če je **obupal**, kako lahko piha zrak. |
| Oznaka | *kontradikcija* |
| Premisa | **A man** sits on his throne behind the drums. |
| Hipoteza | The Queen of England sits behind some drums. |
| Razlaga | A Queen is a woman not **a man**. |
| Premisa | **Človek** sedi na svojem prestolu za bobni. |
| Hipoteza | Angleška kraljica sedi za bobni. |
| Razlaga | Kraljica je ženska in ne **moški**. |

*Opomba.* Za vsakega od treh primerov je najprej podana oznaka primera, potem premisa, hipoteza in razlaga v izvirniku, nato pa še njihovi slovenski prevodi. Napake in nekonsistentnosti so označene odebeljeno.

# 4 Učenje in vrednotenje modelov

V tem razdelku v štirih sklopih predstavimo načine, kako smo učili in uporabili jezikovne modele. Pristopi prvih treh podrazdelkov temeljijo na prilagoditvi (angl. *fine-tuning*) vnaprej naučenih slovenskih velikih jezikovnih modelov. V podrazdelkih 4.1 in 4.2 je opisano učenje jezikovnega modela SloBERTa, najprej več pristopov učenja na množici SI-NLI, nato pa še na prevedeni množici ESNLIsi. V podrazdelku 4.3 je opisan poskus generiranja razlag z generativnimi modeli družine SloT5, ki smo jih učili na razlagah množice ESNLIsi. V podrazdelku 4.4 je opisana uporaba velikega angleškega modela GPT-3.5-turbo. Na koncu, v razdelku 4.5, pa je predstavljen še način vrednotenja pristopov.

Cilj vrednotenja je bil preveriti, kako uspešni so različni pristopi pri klasifikaciji primerov iz testne množice SI-NLI, ki služi kot merilo za vse pristope. V razdelkih 4.3 in 4.4 je predstavljena ocena uspešnosti generiranja razlag, rezultati so predstavljeni v razdelku 5.

Vsa programska koda je bila napisana v programskem jeziku Python v okolju Jupyter Notebook. Za učenje modelov smo uporabili storitev Kaggle Notebooks, ki omogoča zaganjanje datotek Jupyter Notebook v oblaku. Vsi modeli so bili učeni na virtualnem stroju na grafični kartici Nvidia Tesla P100 s 16 GB pomnilnika. Za delo s podatki je bila uporabljena knjižnica Pandas, za učenje modelov pa HuggingFace Transformers (Wolf idr., 2020). Koda je javno dostopna na spletu.[2]

## 4.1 Učenje klasifikatorja SloBERTa na SI-NLI

V tem sklopu smo najprej prilagodili model SloBERTa[3] na učni množici SI-NLI. Ta model služi kot izhodišče za primerjavo ostalih pristopov. Učenje modela smo trikrat ponovili ob različnih delitvah na učno in validacijsko množico, da bi ocenili občutljivost na izbiro učnih podatkov. Z učenjem modela na podmnožici učne množice, ki vsebuje le primere, pri katerih nobeden od označevalcev ni storil napake, skušamo odgovoriti na vprašanje, kako ti primeri vplivajo na uspešnost učenja

2 https://github.com/timkmecl/nli-slovene
3 https://huggingface.co/EMBEDDIA/sloberta

in ali gre pri napakah le za človeško površnost oziroma ali so ti primeri inherentno težji od ostalih ali dvoumni. Za primerjavo je naučen še en model na naključno izbrani podmnožici iste velikosti.

### *4.1.1 Izbira parametrov in učenje modelov*

Parametre učenja smo poskušali nastaviti tako, da bi z učenjem dosegli čim večjo klasifikacijsko točnost na validacijski množici, pri čemer pa smo želeli, da se model neha izboljševati prej kot v 20 epohah (angl. *epoch*), saj bi sicer učenje trajalo predolgo. Za učenje vseh modelov tega razdelka smo uporabili enake parametre, zato jih navajamo le enkrat. Število epoh učenja smo tako nastavili na 20, stopnja učenja (angl. *learning rate*), ki se je izkazala za najbolj uspešno, je $10^{-5}$, delež ogrevanja (angl. *warmup ratio*) pa 0,05. Uporabljena velikost paketa (angl. *batch size*) za učenje je 32. Nastavili smo jo tako, da je čim večja, saj vzporedna obravnava večjega števila primerov pohitri učenje, hkrati pa je dovolj majhna, da zahteve po pomnilniku grafične kartice ne presežejo 16 GB. Uporabljen je privzeti algoritem učenja AdamW s privzetimi parametri $\beta_1 = 0.9$, $\beta_2 = 0.99$, $\epsilon = 10^{-8}$. Po vsaki epohi učenja model generira napovedi za validacijsko množico, kot končni model pa je shranjen model tiste epohe, po kateri je bila klasifikacijska točnost na validacijski množici največja. Po enakem postopku smo model SloBERTa učili tudi na podmnožici učne množice s soglasnimi označevalci.

Zgolj neposredna primerjava rezultatov tega modela z osnovnim ne bi bila ustrezna, saj lahko na kvaliteto modela vpliva tudi količina učnih podatkov. Zato smo iz učne množice naključno izbrali toliko primerov, kot jih je v podmnožici s soglasnimi ocenjevalci (3015). Še na teh primerih smo učili model SloBERTa, rezultati pa dajejo oceno, ali bi s povečanjem števila primerov v podatkovni množici SI-NLI lahko bistveno izboljšali rezultate na njej učenih modelov ali pa je že pri trenutnem številu primerov omejujoč dejavnik zmogljivost modela in ne velikost učne množice.

## 4.2 Učenje s prenosom iz ESNLIsi

V tem sklopu se ukvarjamo z učenjem s prenosom, za kar uporabljamo podatkovno množico ESNLI in lasten strojni prevod njenega dela v slovenščino, ESNLIsi. Zanima nas predvsem, koliko je pri logičnem sklepanju v slovenščini uporabno prevajanje angleških podatkovnih množic kot potencialna rešitev za majhno količino ustreznih učnih primerov, ki so na voljo v slovenščini. Vira stavkov pri SI-NLI in ESNLI sta povsem različna (članki, dokumenti, knjige itd. pri SI-NLI in opisi slik pri ESNLI (Bowman idr., 2015; Klemen idr., 2022; Logar idr., 2013)). To dejstvo smo uporabili za ugotavljanje, kako dobro modeli, naučeni za logično sklepanje na specifičnem tipu stavkov, posplošujejo na drugačen tip. Metoda predobdelave podatkov, izbire parametrov, učenja modelov in izbire končnega modela je enaka kot v prejšnjem razdelku, zato so v tem razdelku navedene le morebitne razlike v metodi oziroma izbranih parametrih.

### *4.2.1 Izbira prevajalnika*

Izvedba zastavljenih ciljev je zahtevala slovenski prevod angleške množice ESNLI, za kar je bila najprej potrebna izbira enega od treh strojnih prevajalnikov (DeepL, Google in NeMo).

Naključno smo izbrali 200 primerov vsakega od treh razredov iz testne množice in po 200 iz validacijske ter tako dobili testno in validacijsko množico s po 600 primeri. Naključno smo izbrali tudi po 500 primerov iz vsakega od treh razredov iz učne množice in dobili učno množico s 1500 primeri. Nato smo vse tri množice prevedli z vsakim od prevajalnikov. Za prevajanje z Google Prevajalnikom in DeepL smo uporabili njihove vmesnike API za Python, za prevajalnik NeMo pa smo najprej prenesli objavljeni model[4] in ga zagnali na lastnem računalniku z uporabo modula nemo_toolkit za Python.

Na vsakem od prevodov smo prilagodili model SloBERTa pri stopnji učenja $10^{-5}$ in deležu ogrevanja 0,02. Točnosti modelov na pripadajočih testnih množicah so predstavljene v Tabeli 5. Glede na majhno količino učnih primerov je razlika med njimi relativno majhna, dodatno pa bi težava z uporabo teh točnosti za izbiro najboljšega prevoda lahko

4 https://www.clarin.si/repository/xmlui/handle/11356/1736

bila uporaba prevodov istega prevajalnika tudi za vrednotenje. Tako bi lahko prevajalnik, katerega prevodi so slabši, a bolj konsistentni, morda dal boljši rezultat kot prevajalnik, ki je boljši, a morda prevaja na manj konsistenten način.

**Tabela 5:** *Točnost napovedi v % za tri strojne prevajalnike*

| | *Deepl* | *Google* | *NeMo* |
|---|---|---|---|
| Prevedena množica ESNLIsi | **74,3** | 73,0 | 71,2 |
| Slovenska množica SI-NLI | 50,2 | **52,7** | 48,0 |

Temu se lahko izognemo z vrednotenjem modelov na primerih, ki so izvorno v slovenščini. V ta namen smo generirali napovedi za validacijsko množico SI-NLI, kot to prikazuje Tabela 5. Na osnovi teh podatkov smo za prevajanje izbrali storitev Google Prevajalnik. Dodatna prednost tega prevajalnika je, da je bilo, za razliko od storitve DeepL, prevajanje dane količine podatkov brezplačno, v primerjavi s prevajalnikom NeMo, kjer prevajanje poteka na lastnem računalniku, pa precej hitrejše. Postopek izbire podatkov in prevajanje, s katerim smo dobili množico ESNLIsi, je opisan v razdelku 3.2.1.

### *4.2.2 Učenje modelov*

Na slovenskih prevodih učne množice ESNLIsi smo 10 epoh pri stopnji učenja $10^{-5}$ in deležu ogrevanja 0,02 učili model SloBERTa. Kriterij za selekcijo končnega modela je bila klasifikacijska točnost na validacijski množici ESNLIsi.

Podobno kot v prejšnjem sklopu smo tudi tu želeli ugotoviti, koliko bi učenje modela na večjem številu primerov izboljšalo rezultat, s čimer bi določili tudi smiselnost nadaljnjega prevajanja primerov iz ESNLI. Model SloBERTa smo zato učili na podmnožici 40 tisoč naključno izbranih primerov (80 %) iz učne množice ESNLIsi z istimi parametri kot prvič in z nespremenjeno validacijsko množico.

Poleg tega smo želeli za primerjavo s SI-NLI določiti še zmogljivost modela, ki bi bil naučen na podobnem številu primerov. V ta namen smo učili še en model SloBERTa na naključno izbranih 4000 primerih učne množice ESNLIsi (velikost učne množice SI-NLI je 3802 primerov) pri stopnji učenja $10^{-5}$ in deležu ogrevanja 0,02, za validacijsko

množico pa smo iz validacijske množice ESNLIsi naključno izbrali 600 primerov (velikost validacijske množice SI-NLI je 590).

Z vsemi tremi modeli smo generirali napovedi za testno množico ESNLIsi. Za vrednotenje učenja s prenosom smo z njimi generirali še napovedi za testno množico SI-NLI. Z osnovnim modelom, naučenim na SI-NLI, pa smo generirali napovedi za testno množico ESNLIsi.

### *4.2.3 Prilagajanje na SI-NLI*

Ugotoviti smo želeli, če oziroma koliko lahko z uporabo ESNLIsi, ki ima več kot 10-krat več primerov kot SI-NLI, izboljšamo model glede na osnovnega, naučenega le na SI-NLI. To smo poskusili storiti s prilagajanjem modela, ki smo ga najprej učili na učni množici ESNLIsi (gl. 4.2.2). Učenje modela se je nato nadaljevalo na učni množici SI-NLI, za validacijsko množico pa je bila uporabljena validacijska množica SI-NLI, saj je bil cilj dobiti model, ki bi čim bolje napovedoval na testni množici SI-NLI. Model smo učili 6 epoh, tokrat pri manjši stopnji učenja $4 \cdot 10^{-6}$. S tako dobljenim modelom smo generirali napovedi za testno množico SI-NLI.

## 4.3 Generiranje razlag s SloT5

V prejšnjih dveh sklopih je bil cilj čim bolj uspešna klasifikacija primerov. Problem logičnega sklepanja pa lahko razširimo tako, da ne zahtevamo le končne oznake primera, temveč želimo imeti še razlago, ki nam pojasni, zakaj je ta oznaka primerna. Naloga je torej generativnega in ne klasifikacijskega tipa, zato v tem delu uporabljamo vnaprej naučena slovenska generativna modela družine SloT5 (manjši model t5-sl-small[5] in večji t5-sl-large[6]). Ker podatkovna množica SI-NLI ne vsebuje razlag primerov, je učenje potekalo le z uporabo množice ESNLIsi. Končni namen je uporaba tako naučenih modelov za generiranje razlag za primere iz testne množice SI-NLI.

Postopek predobdelave podatkov, učenja modelov in izbire parametrov zanje je podoben kot pri učenju modelov SloBERTa, opisanih v prejšnjih razdelkih, le da so tu uporabljeni ekvivalentni razredi

5 https://huggingface.co/cjvt/t5-sl-small
6 https://huggingface.co/cjvt/t5-sl-large

knjižnice HuggingFace, namenjeni učenju generativnih modelov. Razlika je v tem, da generativni model pri učenju kot ciljni izhod zahteva besedilo, podano kot zaporedje členov. Ciljno besedilo je v našem primeru razlaga primera brez dodatne predobdelave.

Naučili smo tri modele. Za učenje dveh smo začeli z modelom t5-sl-small. Učenje prvega je potekalo na naključno izbrani podmnožici učne množice ESNLIsi s 4000 primeri iz prejšnjega razdelka. Stopnja učenja je bila $8 \cdot 10^{-5}$, padanje uteži (angl. *weight decay*) 0,01, velikost paketa 32, ogrevanje pa ni bilo uporabljeno.

Model se je učil 10 epoh. Drugi model smo učili na celotni učni množici ESNLIsi pri stopnji učenja $4 \cdot 10^{-5}$ in ostalih parametrih, enakih kot v prejšnjem primeru.

Na podmnožici s 4.000 primeri smo učili še model t5-sl-large s stopnjo učenja $10^{-4}$, padanjem uteži 0,01 in velikostjo paketa 4, kar je bila največja možna velikost za pomnilnik uporabljene grafične kartice. Model se je učil tri epohe.

Z vsemi tremi modeli smo generirali razlage za primere testne množice ESNLIsi. Za vrednotenje razlag smo uporabili prvih 50 primerov te množice (19 primerov implikacije, 19 kontradikcije in 12 nevtralnih primerov). Zaradi slabih rezultatov že na tej množici modelov nismo vrednotili še na primerih iz množice SI-NLI. Pri tem bi namreč pričakovali še slabše rezultate, saj bi šlo za učenje s prenosom med dvema različnima podatkovnima množicama — to se je pokazalo že pri vrednotenju klasifikacijskih modelov prejšnjega sklopa.

Zaradi rezultatov opisanih treh modelov, predstavljenih v razdelku 5.3, smo se odločili, da večjega SloT5 modela ne bomo učili na celotni množici ESNLIsi. To učenje bi zahtevalo velik časovni vložek, saj bi trajalo več ur, hkrati pa boljšega rezultata od že predstavljenih ni pričakovati.

## 4.4 Uporaba GPT-3.5-turbo

Vsi do sedaj opisani pristopi so temeljili na prilagoditvi vnaprej naučenih jezikovnih modelov SloBERTa in SloT5 na določeni podatkovni množici. Uporabljeni modeli so bili vnaprej naučeni na slovenščini, vsebujejo pa nekaj sto milijonov parametrov. V tem razdelku je predstavljen alternativni pristop. Za napovedovanje oznak v testni množici

SI-NLI uporabimo učenje brez dodatnih primerov (angl. *zero-shot learning*) in z nekaj dodatnimi primeri (angl. *few-shot learning*) z modelom GPT-3.5-turbo, ki ni vnaprej naučen specifično na slovenščini, temveč primarno na angleških besedilih, vsebuje pa nekaj redov velikosti več parametrov od že uporabljenih slovenskih modelov. Enak pristop je bil uporabljen tudi za generiranje razlag.

Pri tem pravzaprav ne gre za učenje nevronskih mrež v običajnem pomenu besede. Pri učenju brez dodatnih primerov modelu zgolj postavimo vprašanje oziroma mu kot vhod damo navodilo, izhod pa je že generirano besedilo, odgovor oziroma napoved. Učna množica, na kateri bi se model učil, tako ni potrebna. Pri učenju z nekaj dodatnimi primeri vhod modela poleg navodila vsebuje še nekaj primerov problema s podanim odgovorom. Tu učno množico potrebujemo le kot vir primerov, zato je lahko zelo majhna (zgolj nekaj primerov). Skrajna različica učenja z nekaj dodatnimi primeri je učenje v kontekstu (angl. *in-context learning*), kjer model kot del vhoda dobi tudi več sto rešenih primerov. Pri obeh tipih učenja je pomembna izbira navodila, saj razumljivost navodila vpliva na kvaliteto odgovorov.

Testiranja smo se lotili v več korakih. V prvem koraku smo uporabili spletni vmesnik OpenAI Playground, v katerem smo na nekaj primerih iz učne množice SI-NLI ročno preizkušali različna navodila. Cilj tega je bil predvsem izločitev navodil, ki bi bila modelu očitno nerazumljiva. Želeli smo tudi doseči, da bi bil odgovor modela v želenem formatu, npr. ena sama beseda pri klasifikaciji in jasno ločena razlaga in končna napoved oznake pri generiranju razlag.

V drugem koraku smo izbrali naključno množico 100 primerov iz učne množice SI-NLI (naključni izbor je vseboval 41 primerov implikacije, 35 kontradikcije in 24 nevtralnih primerov). Napisali smo program, ki za vsak primer generira vhodni tekst za model na podlagi določenega navodila, nato z uporabo programskega vmesnika OpenAI API za Python (razred openai.ChatCompletion) avtomatsko pošilja vhodna besedila modelu, od njega prejme odgovore in shrani napovedane oznake (oziroma generirane razlage). Parameter temperatura (angl. *temperature*) smo nastavili na 0, kar zagotavlja deterministične odgovore, saj pri generiranju teksta model za naslednjo besedo vedno izbere tisto, ki je označena kot najbolj verjetna.

Zadnji korak je bila uporaba enakega pristopa za napovedovanje oznak za celotno testno množico SI-NLI, kjer smo uporabili le navodila, katerih rezultati prejšnjega koraka so bili najboljši. S tem, da so bili primeri v prejšnjem koraku vzeti iz učne množice in ne iz testne, smo preprečili, da bi prek izbire navodil prišlo do prekomernega prileganja podatkom v testni množici.

Žal je bil v času naše uporabe programski vmesnik OpenAI zelo nestabilen, verjetno zaradi preobremenjenosti strežnikov. Strežnik namreč programu pogosto ni vrnil odgovora, zato je bilo treba izvajanje prekiniti. Poleg tega je vmesnik namesto odgovora večkrat javljal napako, da je strežnik preobremenjen. Po eni takšni napaki je bilo za nadaljevanje uporabe modela potrebno čakanje, pogosto nekajminutno, sicer je ob vsaki naslednji poslani zahtevi strežnik vrnil isto napako. Zaradi tega ni bilo možno implementirati avtomatskega ponovnega zaganjanja v primeru napake. Pogostost prekinitev je bila večja v primeru daljših vhodnih besedil in daljših generiranih izhodov, kjer je do napake prišlo na vsakih nekaj primerov. Celovito testiranje pristopa v teh okoliščinah bi bilo bistveno preveč zamudno, zato smo omejili število različnih pristopov, na celotni učni množici pa smo kasneje vrednotili le navodilo, ki se je pri vrednotenju na 100 primerih izkazalo za najboljše.

### *4.4.1 Učenje brez dodatnih primerov*

Najprej smo poskusili učenje brez dodatnih primerov. Na množici 100 primerov smo testirali štiri različna navodila:

– *navodilo-1-en*, ki so ga uporabili Zhong idr. (2023):

```
Given the sentence „{premise}“, determine if the following statement is entailed or contradicted or neutral: „{hypothesis}“
The answer (entailed or contradicted or neutral) is:
```

– *navodilo-1-si*, slovenski prevod prejšnjega:

```
Glede na stavek „{premise}“ ugotovite, ali je naslednja izjava posledica, kontradikcija ali nevtralna: „{hypothesis}“
Odgovor (posledica ali kontradikcija ali nevtralna) je:
```

– *navodilo-2-en*, prirejeno po H. Liu idr. (2023):

```
Instructions: You will be presented with a premise and a
    hypothesis about that premise in Slovene. You need to
    decide whether the hypothesis is entailed by the premise
    by choosing one of the following answers: 'entailment':
    The hypothesis follows logically from the information
    contained in the premise. 'contradiction': The hypothe-
    sis is logically false from the information contained in
    the premise. 'neutral': It is not possible to determine
    whether the hypothesis is true or false without further
    information. Read the passage of information thorough-
    hly and select the correct answer from the three answer
    labels. Read the premise thoroughly to ensure you know
    what the premise entails.
Premise: {premise}
Hypothesis: {hypothesis}
Answer (just one word, either entailment/neutral/
    contradiction):
```

– *navodilo-2-si*, slovenski prevod prejšnjega:

```
Navodila: Predstavljena vam bo premisa in hipoteza o tej
    premisi. Odločiti se morate, ali je hipoteza posledica
    premise, tako da izberete enega od naslednjih odgovo-
    rov: 'posledica': Hipoteza logično sledi iz informacij,
    ki jih vsebuje premisa. 'kontradikcija': Hipoteza je
    logično napačna glede na informacije, ki jih vsebuje
    premisa. 'nevtralno': Brez dodatnih informacij ni mo-
    goče ugotoviti, ali je hipoteza resnična ali napačna.
    Natančno preberite odlomek informacij in izberite pra-
    vilen odgovor med tremi oznakami odgovorov. Temeljito
    preberite predpostavko, da boste vedeli, kaj sledi iz
    predpostavke.
Premisa: {premise}
Hipoteza: {hypothesis}
Odgovor (samo ena beseda, bodisi posledica/nevtralno/
    kontradikcija):
```

V vseh prikazanih navodilih je za vhod v model niz {premise} zamenjan s premiso, {hypothesis} pa s hipotezo primera (tako pri angleških kot pri slovenskih navodilih sta premisa in hipoteza v slovenščini). S testiranjem v OpenAI Playground smo predhodno ugotovili, da je izhod modela kdaj res le ena beseda kot zahtevano (npr. *kontradikcija*), kdaj pa odgovori v celem stavku (npr. *Izjava je kontradikcija*.). Končna klasifikacija je zato določena na podlagi pojavitve podniza v izhodu oziroma odgovoru. Če odgovor vsebuje podniz *posledica* ali *entail*, štejemo, da je napovedana oznaka implikacija, če vsebuje *kontradikcija* ali *contradict* je kontradikcija, če *nevtral* ali *neutral* pa je nevtralna. Za procesiranje odgovorov smo napisali funkcijo v jeziku Python. Ta način klasifikacije odgovorov je bil zadosten, saj je v vseh primerih odgovor vseboval enega od teh podnizov.

Kot najboljše se je izkazalo *navodilo-1-en*, posamezni rezultati so predstavljeni v naslednjem razdelku. Z uporabo tega navodila so bile na enak način kot prej napovedane oznake za celotno testno množico SI-NLI.

Za razliko od vnaprej naučenih modelov iz prejšnjih razdelkov, ki so bili učeni na slovenščini, je GPT-3.5-turbo v največji meri učen na angleških besedilih. Zato smo želeli preveriti, če in koliko slabši je rezultat zaradi morebitnega slabšega razumevanja v učnih podatkih bistveno manj zastopane slovenščine v primerjavi z angleščino. Napovedi za izbrano množico 100 primerov smo tako generirali še z uporabo njihovih angleških prevodov (in navodila *navodilo-1-en*).

### *4.4.2 Učenje z nekaj dodatnimi primeri*

Najboljše navodilo (*navodilo-1-en*) smo uporabili za osnovo testiranja učenja z nekaj dodatnimi primeri. Sledili smo načinu, ki so ga uporabili Zhong idr. (2023). Iz učne množice smo naključno izbrali po en primer vsake od treh oznak ter njihove premise in hipoteze vstavili v navodilo. Na koncu navodil smo dodali odgovor entailed, če je bil primer v razredu implikacija, če kontradikcija *contradicted* in *neutral* za nevtralne primere. Isti nabor treh dopolnjenih navodil smo dodali na začetek vhoda za vse primere pri napovedovanju.

Na ta način smo generirali napovedi za prej opisano množico 100 primerov. Poskus smo skupaj ponovili trikrat, vsakič z drugimi tremi naključnimi izbranimi primeri. Rezultati so predstavljeni v razdelku 5.4. Ker ta pristop v primerjavi z učenjem brez dodatnih primerov skoraj ni izboljšal rezultatov, zaradi tehničnih težav strežnikov OpenAI, opisanih na začetku razdelka, tega pristopa nismo dodatno vrednotili na testni množici SI-NLI. Zaradi počasnosti strežnika tudi nismo poskušali uporabiti več kot treh podanih primerov naenkrat.

### *4.4.3 Generiranje razlag*

Podobno metodo smo uporabili še za hkratno generiranje razlag in klasifikacijo. S pomočjo testiranja znotraj vmesnika OpenAI Playground smo *navodilo-1-en* modificirali tako, da je model kot izhod najprej podal razlago v slovenščini, v naslednji vrstici pa klasifikacijo. Navodilo, ki smo ga na koncu uporabili, je bilo oblike:

```
Given the sentence „{premisa}“, determine if the following
    statement is entailed or contradicted or neutral: „{hi-
    poteza}“. First give a short, one sentence reasoning or
    explanation for the decision in slovene, and then the
    final answer (one english word - „entailed“ or „contra-
    dicted“ or „neutral“) in a new line after that.

Razlaga v slovenščini:
```

Niza {premisa} in {hipoteza} sta kot prej vsakič zamenjani s premiso in hipotezo danega primera. Izhod modela nato razdelimo na dva dela glede na znak za novo vrstico. Prvi del vzamemo za razlago, drugi del pa obravnavamo enako kot odgovor pri prej opisanem pristopu in mu dodelimo oznako glede na vsebovan podniz. Pri ročnem testiranju smo ugotovili, da kljub zahtevi v navodilu, naj bo končni odgovor beseda v angleščini, kdaj odgovori v slovenščini, pri čemer namesto besede *entailed* uporabi besedo *potrjeno*. Zato tudi primere, ki vsebujejo podniz *potrjen*, pri napovedovanju uvrstimo kot implikacijo.

Z opisano metodo smo generirali napovedi in razlage za množico 100 primerov. Na koncu smo izbrali podmnožico 50 primerov, ki jo bomo uporabili za kvalitativno vrednotenje generiranih razlag. Želeli smo, da pogostost posameznih razredov v njej vsaj približno odraža pogostost razredov v celotni podatkovni množici, zato smo iz množice stotih naključno izbrali 18 primerov implikacije, 16 kontradikcije in 16 nevtralnih primerov.

## 4.5 Način vrednotenja

V tem razdelku opišemo, kako smo vrednotili različne pristope vrednotili. Najprej predstavimo načine za vrednotenje klasifikatorjev. Definiramo nekaj metrik in utemeljimo primernost njihove uporabe za naš problem. Opišemo tudi način vrednotenja generiranih razlag. Rezultati opisanega vrednotenja so predstavljeni v naslednjem razdelku.

### *4.5.1 Evalvacijske metrike za klasifikacijo*

Za kvantitativno vrednotenje klasifikatorjev poznamo več različnih metrik. Temeljna je klasifikacijska točnost (angl. *classification accuracy*) ali samo točnost, ki nam pove delež primerov, ki jih je klasifikator uvrstil v pravilen razred.

Zgolj ta podatek nam pogosto ne da zadostne informacije o delovanju klasifikatorja, predvsem v primeru neuravnoteženih podatkovnih množic. Poznamo metrike, s katerimi lahko bolje ovrednotimo tudi tovrstne primere. Natančnost (angl. *precision*) pove, kolikšen delež primerov, uvrščenih v izbrani ciljni razred, temu razredu dejansko pripada. Priklic (angl. *recall*) pove, kolikšen delež primerov ciljnega razreda je vsebovan med pozitivnimi napovedmi tega razreda. Ocena $F_1$ (angl $F_1$-*score*) je harmonična sredina natančnosti in priklica in služi kot nekakšen povzetek sposobnosti klasifikatorja za napovedovanje izbranega razreda, kar je lahko dobra alternativa klasifikacijski točnosti (Müller in Guido, 2016).

Če imamo pri klasifikacijskem problemu le en ciljni razred, lahko za vrednotenje napovedi uporabimo le metrike za ta razred. Pri problemu klasifikacije v več razredov pa jih najprej izračunamo za vsak razred posebej, nato pa izračunamo njihovo povprečje. Uporaba mikro

povprečja je priporočena, kadar je enako pomemben vsak posamezen primer, če pa je enako pomemben vsak posamezen razred, se uporabi makro povprečje (Müller in Guido, 2016).

### *4.5.2 Vrednotenje klasifikatorjev*

Obe uporabljeni testni množici sta skoraj uravnoteženi, zato smo pri vrednotenju klasifikatorjev najbolj upoštevali klasifikacijsko točnost in o njej poročamo pri vseh pristopih. Dodatno pri vsakem pristopu navajamo še povprečje ocene $F_1$, povprečje natančnosti in priklicev.

Pri problemu logičnega sklepanja v naravnem jeziku, s katerim se ukvarjamo, nobeden od treh razredov ni privilegiran glede na ostale. Ker so razredi za nas enakovredni, je smiselna uporaba makro povprečja metrik. V nadaljevanju besedila zato ocena $F_1$ danega pristopa pomeni makro povprečje ocen $F_1$ za vse tri razrede, enako velja tudi za natančnost in priklic. Glavni merili za primerjavo pristopov sta tako klasifikacijska točnost in ocena $F_1$, in sicer na testni množici SI-NLI, ki je izvirno slovenska, človeško ustvarjena množica.

Ker nas zanima tudi, kakšne tipe napak delajo modeli in ne le njihova sposobnost napovedovanja razredov, pri nekaterih rezultatih dodamo še matriko zamenjav (angl. *confusion matrix*), ki vsebuje po en stolpec in eno vrstico za vsakega od možnih razredov. Element matrike v stolpcu A in vrstici B pove število primerov razreda B, za katere je klasifikator napovedal razred A. Vsota diagonale te matrike je število pravilno uvrščenih primerov.

V poskusih iz drugega sklopa, kjer se ukvarjamo z učenjem s prenosom, navajamo še klasifikacijske točnosti modelov na testni množici ESNLIsi.

V poskusih uvrščanja z GPT-3.5-turbo zaradi opisanih težav za vse pristope izračunamo točnost, oceno $F_1$ in matrike zamenjav za izbrano množico 100 primerov, ki smo jo uporabili za vrednotenje. Ker v tej množici primeri niso tako enakomerno razporejeni med razredi kot v celotnih testnih množicah, je tu relativno bolj pomembna ocena $F_1$. Prej navedene metrike za celo testno množico SI-NLI izračunamo samo za najuspešnejši pristop na manjši množici, kot je opisano v razdelku 4.4.

### *4.5.3 Vrednotenje generiranja razlag*

Za generirana besedila obstaja več metrik, ki jih lahko izračunamo na podlagi primerjave generiranega besedila s ciljnimi odgovori, ki jih imamo v testni množici. Pristopi temeljijo na podobnosti dveh besedil. Ker lahko za en primer obstajata dve ali več pravilnih razlag, ki se med seboj razlikujejo, poleg tega pa lahko le ena beseda povsem spremeni pravilnost razlage, takšen povsem kvantitativen pristop tu ne bi bil ustrezen. Namesto tega smo 50 razlag za vsak pristop ročno pregledali in jih ocenili kot ustrezne ali ne.

Da je razlaga ocenjena kot ustrezna, mora biti pravilna, torej mora najprej dati argument za pravilno klasifikacijo primera. Zgolj pravilnost pa ne zadošča. Razlage, kot je na primer *Druga trditev pove isto kot prva* (kot utemeljitev, zakaj je nek primer implikacija) ali *Navedbe v drugi trditvi nasprotujejo prvi* (za kontradikcijo), so lahko sicer pravilne, a ne podajajo nobene dodatne informacije. Zahtevali smo, da je iz razlage razvidno neko razumevanje, torej da razlaga izpostavi, kaj v premisi in hipotezi privede do tega, da se primer uvršča v nek razred.

Pri tem nismo zahtevali, da je razlaga popolna. Na primer, če je razlogov za kontradikcijo več, zadošča navedba enega. Prav tako za pozitivno oceno nismo zahtevali slovnične pravilnosti, želeli smo le razumljivost, na oceno pa ni vplival niti slog.

Modele SloT5, učene na ESNLIsi, iz tretjega sklopa smo vrednotili na prvih 50 primerih testne množice ESNLIsi, ki vsebujejo 19 primerov implikacije, 19 kontradikcije in 12 nevtralnih primerov. Razlage, generirane z GPT-3.5-turbo, smo vrednotili na množici 50 primerov iz SI-NLI, in sicer 18 naključno izbranih primerov implikacije, 16 kontradikcije in 16 nevtralnih primerov.

## 5 Rezultati

V tem razdelku po sklopih predstavimo rezultate, pridobljene z metodami, opisanimi v prejšnjem razdelku. Pristope ocenimo tako kvantitativno kot tudi kvalitativno. Rezultate interpretiramo in na njihovi podlagi skušamo odgovoriti na izhodiščna vprašanja.

## 5.1 Učenje klasifikatorja SloBERTa na SI-NLI

Na testni množici SI-NLI smo vrednotili vse tri v tem sklopu naučene modele. Model SloBERTa, prilagojen na celotni učni množici SI-NLI, imenujemo model *SI-NLI-celotna*. Model *SI-NLI-soglasni* je model SloBERTa, prilagojen le na primerih učne množice, pri katerih so bili vsi označevalci med seboj soglasni in so hkrati izbrali pravilno oznako primera. Model *SI-NLI-manjša* pa je model SloBERTa, ki smo ga prilagodili na naključno izbrani podmnožici učne množice iste velikosti (približno 80 % cele množice). Rezultati vrednotenja so prikazani v Tabeli 6. Model *SI-NLI-celotna*, prilagojen na celotni učni množici SI--NLI, doseže najvišje ocene pri vseh metrikah.

Pri poskusu še trikratne naključne delitve izvirne učne množice SI-NLI na učno in validacijsko smo ugotovili, da znaša povprečje klasifikacijskih točnosti štirih različnih delitev 73,2 %, standardni odklon pa 0,8 %. Točnost je torej nekoliko odvisna od izbire učnih primerov, kar moramo upoštevati pri primerjavi modelov.

**Tabela 6:** *Metrike v %, izračunane na napovedih za testno množico SI-NLI*

| Model | Točnost | $F_1$ | Natančnost | Priklic |
|---|---|---|---|---|
| *SI-NLI-celotna* | **73,2** | **73,2** | **73,3** | **73,2** |
| *SI-NLI-manjša* | 72,2 | 72,2 | 72,4 | 72,4 |
| *SI-NLI-soglasni* | 72,9 | 73,0 | 73,3 | 72,8 |

*Opomba.* Za tri modele SloBERTa, učene na celotni učni množici SI-NLI in dveh podmnožicah.

Model *SI-NLI-manjša* ima za 1 % manjšo klasifikacijsko točnost od modela *SINLI-celotna*, podobno tudi ostale metrike. To je vpliv zmanjšanja učne množice na 80 % prvotne. Količina učnih primerov, vsebovana v podatkovni množici SI-NLI (nekaj tisoč), je trenutno torej tako majhna, da bi z razširjanjem množice z dodatnimi primeri rezultate modelov, učenih na njej, še lahko izboljševali.

Oglejmo si vpliv primerov, na katerih se označevalci tudi motijo, na učenje modela. Vidimo lahko, da izločitev takšnih primerov iz učne množice zmanjša točnost za manj kot odstotek. To zmanjšanje je le malo manjše kot takrat, ko so izločeni primeri naključno izbrani, razlika je manjša od standardnega odklona. Sklepamo lahko, da gre

predvsem za posledico manjše učne množice. Delež takih primerov v učni množici torej ne vpliva na uspešnost učenja.

V Tabeli 7 podajamo še točnosti vseh treh modelov na podmnožici učne množice s soglasnimi označevalci, podmnožici ostalih (torej tistih primerih, kjer je kateri od označevalcev naredil napako), in razlike teh dveh točnosti. Izkaže se, da so razlike med modeli le pri podmnožici soglasnih odločitev, medtem ko dajejo enake rezultate na ostalih primerih.

**Tabela 7:** *Primerjava točnosti napovedi v % za soglasne in nesoglasne primere testne množice SI-NLI*

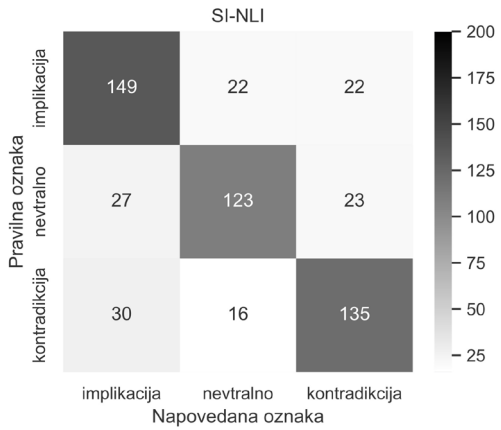
| Model | Točnost na soglasnih | Točnost na ostalih | Razlika |
|---|---|---|---|
| *SI-NLI-celotna* | **76,0** | 61,0 | **15,0** |
| *SI-NLI-manjša* | 73,6 | 61,0 | 12,5 |
| *SI-NLI-soglasni* | 74,4 | 61,0 | 13,4 |

Pri vseh treh modelih je točnost na soglasnih primerih skoraj 15 % večja kot na ostalih. Vidimo, da primere, pri katerih so imeli težave ljudje, slabše klasificirajo tudi naučeni modeli. To, ali so bili primeri takšne vrste vsebovani v učni množici ali ne, na točnost pri napovedovanju na njih ne vpliva. Najverjetneje gre vsaj delno za dvoumne primere (takšne, ki jih tudi ljudje razumemo na različne načine), kar hkrati razloži tako napake nekaterih človeških označevalcev kot slabše rezultate modelov.

Drugačne oznake nekaterih označevalcev tako vsaj do neke mere niso napake, temveč različne interpretacije istega primera. Posledično bi to lahko upoštevali pri učenju modelov. Možen pristop bi lahko namesto le enega ciljnega razreda pri učenju takim primerom podal verjetnostno porazdelitev, kjer bi bila verjetnost vsakega od razredov delež človeških oznak tega razreda. S tem se nismo ukvarjali in je predlog za nadaljnje delo.

Iz matrik zamenjav modela *SI-NLI-celotna* na Sliki 2 (napake drugih modelov so podobne) vidimo, da so vrste napak, ki jih modeli naredijo, relativno enakomerno razporejene med šestimi možnostmi (ne glede na točno izbiro tipa ali količine učnih primerov). Nobena vrsta napake posebej ne izstopa, prav tako modeli nimajo večjih težav s

primeri enega razreda kot z drugimi. Po tem se očitno razlikujejo od človeških označevalcev (Slika 1), ki razredov implikacija in kontradikcija sploh ne zamenjujejo, največkrat pa zamenjajo primere implikacije in nevtralne primere. Jasno je, da se strojni način reševanja problema razlikuje od človeškega. To opažanje skušamo dodatno razložiti prek generiranja razlag v razdelku 5.3.




**Slika 2:** Matrika zamenjav za napovedi modela, učenega na celotni učni množici SI-NLI.

## 5.2 Učenje s prenosom iz ESNLIsi

V tem poskusu smo evalvirali tri modele SloBERTa, prilagojene na množici ESNLIsi: model *ESNLIsi-celotna*, učen na celotni učni množici ESNLIsi; model *ESNLIsi-40k*, učen na 80 % primerov iste množice; model *ESNLIsi-4k*, učen na 4 tisoč primerih te množice, kar je podobno številu primerov v množici SI-NLI; in model *ESNLIsi-SI-NLI*, najprej učen na celi učni množici ESNLIsi, zatem pa prilagojen še na učni množici *SI-NLI*.

### *5.2.1 Klasifikator za ESNLIsi*

V Tabeli 8 so prikazane klasifikacijske točnosti na testni množici ESNLIsi za prve tri modele in model *SI-NLI-celotna* iz prejšnjega razdelka. Največjo točnost ima model *ESNLIsi-celotna*, le malo manjšo *ESNLIsi-40k*. Daleč najmanjša je točnost modela *SI-NLI-celotna*, kjer gre v tem primeru za prenos znanja iz ene množice na drugo.

**Tabela 8:** *Točnost napovedi v % za testno množico ESNLIsi treh modelov, učenih na ESNLIsi, in modela, učenega na SI-NLI*

| Model | *ESNLIsi-celotna* | *ESNLIsi-40k* | *ESNLIsi-4k* | *SI-NLI-celotna* | *Poth idr.* | *Wang idr.* |
|---|---|---|---|---|---|---|
| Točnost | 85,7 | 85,4 | 80,0 | 49,3 | 91,1 | **93,1** |

*Opomba.* Dodane so še točnosti, ki so jih na množici ESNLI dosegli Wang idr. (2021) z lastno arhitekturo in Poth idr. (2021) z modelom RoBERTa.

Naša največja dosežena točnost je manjša od največje točnosti na množici SNLI, ki so jo z lastno arhitekturo jezikovnega modela dosegli Wang idr. (2021). Manjša je tudi od točnosti, ki so jo s prilagoditvijo modela RoBERTa dosegli Poth idr. (2021). Njihov model je po zgradbi in velikosti enak modelu SloBERTa, ki smo ga uporabili mi. Razlika je delno posledica tega, da smo uporabili 10-krat manjšo učno množico (prevedli smo le del množice ESNLI), delno posledica napak pri prevajanju (opisanih v razdelku 3.2.1), delno pa zaradi vnaprejšnjega učenja modela na drugih korpusih v različnih jezikih. Razlika znaša približno 5 %, iz česar lahko sklepamo, da je to približek za delež primerov v podatkovni množici, pri katerih je prevod napačen do te mere, da ga ni več mogoče uvrstiti v ustrezni razred.

Točnost modela *SI-NLI-celotna* na množici SI-NLI v prejšnjem sklopu je več kot 5 % manjša od točnosti modela *ESNLIsi-4k* na ESNLIsi. Oba modela sta bila učena na podobni količini primerov, zato razlika ni posledica razlike v velikosti učnih množic. Prav tako ni posledica napak, ki bi jih povzročilo strojno prevajanje množice ESNLI iz angleščine, saj bi to kvečjemu zmanjšalo točnost na ESNLIsi. Množica SI-NLI vsebuje povedi, pridobljene iz različnih vrst besedil, medtem ko povedi v ESNLI temeljijo na opisih slik. SI-NLI je bolj raznolika, hkrati pa so primeri daljši in pogosto bolj abstraktni. Iz primerjave primerov vidimo, da so primeri logičnega sklepanja v množici ESNLIsi dejansko lažji kot v SI-NLI. Napovedovanje na bolj raznolikih in abstraktnih primerih je za model SloBERTa težje.

Povečanje števila primerov s 4 na 40 tisoč poveča klasifikacijsko točnost za nekaj več kot 5 %, dodatno povečanje na 50 tisoč pa za manj kot pol odstotka. Količina 50 tisoč primerov je za obravnavan par modela in podatkovne množice zadostna. S prevajanjem dodatnih primerov iz angleščine se sposobnost napovedovanja z uporabo

prilagoditve modela SloBERTe ne bi bistveno povečala, zato menimo, da nadaljnje prevajanje iste podatkovne množice ni smiselno.

### 5.2.2 Učenje s prenosom za napovedovanje na SI-NLI

V Tabeli 9 podajamo klasifikacijske točnosti in tri ostale metrike, izračunane na testni množici SI-NLI za štiri modele tega sklopa. Pri prvih treh modelih gre v tem primeru za prenos znanja. Točnost in ocena $F_1$ se večata z večanjem števila učnih primerov. Vse metrike so daleč največje za model *ESNLIsi-SI-NLI*, ki je edini še prilagojen na množici SI-NLI.

Najpogostejša napaka modelov je napačna napoved nevtralnega razreda namesto drugih dveh razredov. Model, naučen na večji učni množici, dela takšnih napak manj. Struktura napak modela *ESNLIsi-SI-NLI* je podobna strukturi napak modelov prejšnjega sklopa, učenih le na SI-NLI.

Vidimo, da ima tu večanje števila učnih primerov večji vpliv kot pri vrednotenju, ko učna in testna množica pripadata isti podatkovni množici. Iz razlike med točnostjo in oceno $F_1$ modelov *ESNLIsi-celotna* in *ESNLIsi-40k* lahko sklepamo, da bi tu s prevajanjem dodatnih primerov množice ESNLI lahko rezultate še nekoliko izboljšali.

**Tabela 9:** *Metrike v %, izračunane na napovedih za testno množico SI-NLI*

| Model | Točnost | $F_1$ | Natančnost | Priklic |
|---|---|---|---|---|
| *ESNLIsi-celotna* | 65,4 | 65,2 | 67,1 | 65,2 |
| *ESNLIsi-40k* | 64,0 | 63,8 | 67,6 | 64,1 |
| *ESNLIsi-4k* | 55,9 | 55,5 | 61,6 | 56,6 |
| *ESNLIsi-SI-NLI* | **75,3** | **75,3** | **75,3** | **75,4** |

*Opomba.* Za tri modele, učene na ESNLIsi, in model, ki je bil na koncu prilagojen še na SI-NLI.

Kljub temu, da se je model *ESNLIsi-celotna* učil na več kot 10-krat večji množici kot *SI-NLI-celotna* iz prejšnjega sklopa, sta zaradi učenja s prenosom tako točnost kot ocena $F_1$ na množici SI-NLI manjši za skoraj 10 %.

Še slabša sta rezultata modelov *SI-NLI-celotna* in *ESNLIsi-4k*, ki sta bila učena na nekaj tisoč primerih, na tuji množici. Čeprav je množica SI-NLI bolj raznolika in težja za napovedovanje, je prenos znanja modela, učenega na njej, celo nekoliko slabši. To je morda posledica

prevajalskih napak v testni množici ESNLIsi, ki so lahko vzrok za napačne klasifikacije in posledično slabše rezultate modela *SI-NLI-celotna*, vrednotenega na njej.

Iz teh ugotovitev lahko zaključimo, da je prenos znanja med različnimi množicami primerov logičnega sklepanja v naravnem jeziku relativno slab. Izboljšuje se z večanjem števila učnih primerov. Tudi z za red velikosti večjim številom učnih primerov pa ne dosegamo rezultatov modelov, učenih na istem tipu primerov, kot jih vsebuje testna množica. Model SloBERTa torej relativno slabo posplošuje z ene podatkovne množice na drugo, oziroma s problema logičnega sklepanja na povedih enega izvora na povedi drugega izvora.

Z uporabo množice ESNLIsi smo kljub temu uspeli izboljšati rezultat modela SloBERTa. Če smo to množico uporabili za vnaprejšnje učenje, nato pa model prilagodili še na SI-NLI, sta točnost in ocena $F_1$ za približno odstotek večja kot brez njene uporabe. Tudi ta model pa ne dosega najboljšega rezultata, objavljenega na SloBench (točnost 77,2 %) (CJVT UL, 2023), ki je sicer učen na nekoliko večji množici, ki vsebuje tudi našo testno množico. Za izboljšanja napovednih modelov za množico SI-NLI oziroma za logično sklepanje v slovenščini na splošno z uporabo učenja s prenosom bi bilo tako smiselno nadaljnje prevajanje množic, ki so v primerjavi z ESNLI po izvoru povedi bolj raznolike.

Kljub vsemu je rezultat učenja na 50 tisoč prevodih klasifikator, ki na novi, povsem drugače sestavljeni množici, dve tretjini primerov pravilno uvrsti. Ta pristop je enostavno posplošiti na druge jezike z malo viri, v katerih podatkovne množice primerov logičnega sklepanja ne obstajajo.

### 5.3 Generiranje razlag s SloT5

Vrednotili smo razlage, generirane za prvih 50 primerov testne množice ESNLIsi, generirane s tremi modeli: modelom *t5-large-4k*, dobljenim s prilagoditvijo večjega modela SloT5 (t5-sl-large) na podmnožici učne množice ESNLIsi s 4 tisoč primeri; modelom *t5-small-4k*, dobljenim s prilagajanjem manjšega modela SloT5 (t5-sl-small) na isti množici; in modelom *t5-small-50k*, dobljenim s prilagajanjem manjšega modela SloT5 na vseh 50 tisoč primerih učne množice ESNLIsi.

**Tabela 10:** *Rezultati vrednotenja razlag treh modelov SloT5 na množici 50 primerov*

| Model | *t5-large-4k* | *t5-small-4k* | *t5-small-50k* |
|---|---|---|---|
| Ustrezne razlage izmed 50 | 8 (16%) | 8 (16%) | **14** (28%) |

V Tabeli 10 so navedena števila ustreznih razlag. Kriterij za ustreznost razložimo in utemeljimo v razdelku 4.5.3. Vidimo, da je število ustreznih razlag zelo majhno. Nekoliko večje je le pri učenju na 50 tisoč primerih, a je delež ustreznih razlag še vedno manjši od tretjine. Večji model ni bolj uspešen kot manjši. Vzrok za to je verjetno podučenost večjega modela. Ulčar in Robnik-Šikonja (2023) namreč ugotavljata, da je manjši od modelov SloT5 pri večini testiranih nalog boljši zaradi podučenosti modela t5-sl-large. Zaradi te ugotovitve in naših lastnih rezultatov večjega modela nismo učili na celotni učni množici ESNLIsi, saj bi to zahtevalo prevelik časovni vložek glede na pričakovane rezultate.

Trije primeri iz testne množice ESNLIsi s premiso, hipotezo, podano razlago in razlagami, generiranimi s tremi modeli, so prikazani v Tabeli 11.

**Tabela 11:** *Trije primeri iz testne množice ESNLIsi (prvi kontradikcija, drugi nevtralen in tretji implikacija) in zanje generirane razlage treh modelov SloT5*

| | | |
|---|---|---|
| Premisa | Dve osebi hodita po ulicah. | |
| Hipoteza | Dve osebi sedita notri. | |
| Razlaga | Ljudje ali hodijo ali sedijo. Lokacija je zunaj na ulicah ali znotraj. | |
| *t5-large-4k* | Ljudje so bodisi na ulicah ali v zaprtih prostorih. | ✓ |
| *t5-small-4k* | Ljudje ne morejo hoditi in hoditi hkrati. | × |
| *t5-small-50k* | Ljudje ne morejo hoditi in sedeti hkrati. | ✓ |
| Premisa | V ospredju stoji semafor z dvema simboloma kolesa, zeleno puščico, ki zavija levo, in rdečim signalom za naprej, medtem ko moški v belem klobuku in modri srajci kolesari po rdeči poti. | |
| Hipoteza | Ta moški je športnik. | |
| Razlaga | Ne morete sklepati, da je športnik. | |
| *t5-large-4k* | Ni vsak moški športnik. | ✓ |
| *t5-small-4k* | Človek ne more biti športnik. | × |
| *t5-small-50k* | Človek ne more biti hkrati športnik in športnik. | × |
| Premisa | Kuhar srednjih let je v belih vratih mobilne kuhinje. | |
| Hipoteza | V bližini kuhinje je kuhar. | |

| Razlaga | Kuhar na vratih kuhinje je blizu kuhinje. | |
|---|---|---|
| *t5-large-4k* | Kuhar je v belih vratih mobilne kuhinje. | × |
| *t5-small-4k* | Če je kuhar, potem je v bližini kuhinje. | × |
| *t5-small-50k* | Kuhar srednjih let je kuhar. | × |

*Opomba.* Ustrezne razlage so označene s ✓, napačne pa z ×.

### *5.3.1 Kvalitativna ocena razlag*

Ob pregledu generiranih razlag nismo opazili očitnih značilnosti, ki bi razlage modelov medsebojno razlikovale. Kvalitativno so si zelo podobne, zato podajamo splošna opažanja.

Razlage so slovnično pravilne in modeli SloT5 se dobro naučijo forme razlag. V učni množici se pogosto pojavljajo razlage določenih oblik, npr. *X ne more Y in Z hkrati.* ali *X ne more biti Y* za kontradikcijo, *Ni vsak X Y* za nevtralne ali *Če je X, potem je Y* za implikacijo. Modeli si te oblike ali predloge zapomnijo in jih uporabljajo, a jih izbirajo na videz naključno. Pogosto je glede na oznako primera napačna že sama izbira oblike. To lahko vidimo pri drugi in tretji razlagi za drugi primer (Tabela 11).

Modeli se naučijo tudi tega, da morajo v razlagah uporabiti besede, ki se pojavijo v premisi in hipotezi. To pogosto privede do povsem nesmiselnih povedi, kot je druga razlaga za prvi primer in tretja razlaga za drugi primer v Tabeli 11. V nekaterih primerih model tako ustvari resnično poved, a ni ustrezna kot razlaga, takšna je tretja razlaga za tretji primer (Tabela 11).

Sklepamo, da modeli SloT5 nimajo zadostnega dejanskega poznavanja sveta, da bi ga lahko uporabili za generiranje razlag pri logičnem sklepanju. Naučijo se zgolj forme, niso pa zmožni pisati pomensko ustreznih razlag. To kaže na to, da so zmožni iskanja in uporabe jezikovnih vzorcev, poznavanje jezika pa ni povezano s poznavanjem resničnosti.

Poizkus generiranja razlag z modeli SloT5 ocenjujemo kot neuspešen.

McCoy idr. (McCoy idr., 2019) domnevajo, da jezikovni modeli za reševanja problema logičnega sklepanja uporabljajo različne jezikovne hevristike, ki temeljijo na vsebovanosti delov hipoteze v premisi. Uporaba takšnih hevristik lahko razloži nezmožnost generiranja ustreznih razlag naših modelov, saj gre pri njihovi uporabi le za jezikovno analizo in ne za dejansko razumevanje.

Uporaba podobnih hevristik namesto dejanskega razumevanja sveta je zato verjetno razlaga za slab prenos znanja in slabo posploševanje ter drugačno strukturo napak klasifikatorjev glede na človeške, ki smo jih omenjali v prejšnjih dveh sklopih. Njihov pristop k reševanju zastavljenega problema po tej domnevi temelji na procesiranju jezika namesto na poznavanju zakonitosti resničnega sveta in zdravorazumskega sklepanja, kot to počnemo ljudje. Naučene hevristike na eni podatkovni množici morda ne delujejo na drugi, če se domena povedi med njima preveč razlikuje.

## 5.4 Uporaba GPT-3.5-turbo

V tem razdelku podajamo rezultate vrednotenja različnih načinov uporabe modela GPT-3.5-turbo na množici 100 primerov podatkovne množice SI-NLI. Na celotni testni množici SI-NLI je ovrednoten najboljši od pristopov.

### *5.4.1 Učenje brez dodatnih primerov*

Kot lahko vidimo iz prvih štirih vrstic Tabele 12, je za uspešnost pri učenju brez dodatnih primerov pomembna izbira navodila, saj to lahko spremeni točnost napovedi za skoraj 10 %. Navodilo nekoliko vpliva tudi na vrsto napak. Ne glede na izbiro navodila model GPT največ primerov napačno uvrsti kot implikacijo. Najmanjkrat model pravilno klasificira nevtralni razred.

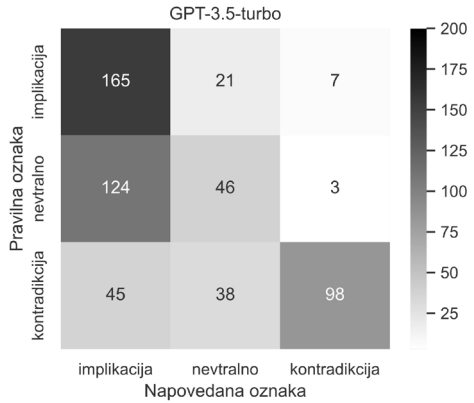
**Tabela 12:** *Rezultati klasifikacije v % pri uporabi modela GPT-3.5-turbo z učenjem brez dodatnih primerov za različna navodila*

| Navodilo | Točnost | $F_1$ | Natančnost | Priklic |
|---|---|---|---|---|
| *navodilo-1-en* | **59** | **54,6** | 57,3 | **54,6** |
| *navodilo-1-si* | 51 | 48,7 | 53,4 | 47,4 |
| *navodilo-2-en* | 51 | 47,3 | **58,6** | 50,5 |
| *navodilo-2-si* | 54 | 48,4 | 54,6 | 48,4 |
| *navodilo-razlaga* | 49 | 47,2 | 49,4 | 47,7 |
| *navodilo-1-en* angleški | *66* | *55,3* | *59,1* | *58,6* |

*Opomba.* Metoda navodilo-razlaga najprej generira razlago primera, nato pa primer še klasificira (glej razdelek 5.4.3). Pri pristopu v zadnji vrstici tabele so bile napovedi generirane na angleških prevodih primerov, zato pristop ni neposredno primerljiv s prejšnjimi.

**Tabela 13:** *Rezultati v % pri uporabi modela GPT-3.5-turbo z učenjem brez dodatnih primerov in uporabo navodila-1-en za napovedovanje oznak v celotni testni množici SI-NLI*

| Točnost | $F_1$ | Natančnost | Priklic |
|---|---|---|---|
| 56,5 | 54,5 | 61,3 | 55,4 |




**Slika 3:** Matrike zamenjav za napovedi GPT-3.5-turbo z učenjem brez dodatnih primerov in uporabo navodila-1-en za celotno testno množico SI-NLI.

Ker dosega od štirih testiranih navodil *navodilo-1-en* na množici 100 primerov najvišjo točnost in oceno $F_1$, smo to vrednotili na celotni testni množici SI-NLI. Metrike so podane v Tabeli 13, matrika zamenjav pa na Sliki 3. Najpogostejša napaka je tako kot pri človeških označevalcih (Slika 1) označitev nevtralnega primera za implikacijo. Model najslabše uvršča nevtralne primere.

Rezultati so očitno slabši kot rezultati modela SloBERTa, prilagojenega na SI-NLI. Vseeno so rezultati minimalno boljši kot pri uporabi učenja s prenosom z nekaj tisoč učnimi primeri (model *ESNLIsi-4k* v Tabeli 9), vendar slabši kot pri učenju s prenosom z nekaj deset tisoč učnimi primeri (model *ESNLIsi-40k*). GPT-3.5-turbo je torej sposoben reševanja problemov s področja logičnega sklepanja v naravnem jeziku, čeprav ni bil učen ali prilagojen za to nalogo. Učenje zelo velikih modelov na veliki količini podatkov s spleta da modelu zadostno razumevanje pomena jezika in razumevanje sveta, da je zmožen relativno uspešno reševati nalogo sklepanja v naravnem jeziku.

Glede na to, da je pri navodilih *navodilo-2-** kot del navodila podana tudi razlaga oznak, napake pri uvrščanju niso posledica nerazumevanja pomena treh NLI oznak. Če primerjamo metrike v Tabeli 12, vidimo, da uporaba angleških prevodov izboljša točnost za 7 %, izboljša pa tudi preostale metrike. Sklepamo lahko, da ima GPT-3.5-turbo nekaj težav z razumevanjem slovenščine. Očitno slovenščino pozna, saj tudi pri uporabi slovenskih izvirnikov doseže primerljive rezultate kot učenje s prenosom slovenskega modela SloBERTa, je pa zaradi majhne zastopanosti slovenščine njeno poznavanje slabše.

### *5.4.2 Učenje z nekaj dodatnimi primeri*

V Tabeli 14 so prikazane metrike za tri različne naključne izbire dodatnih primerov, za primerjavo pa še za isto navodilo brez dodatnih primerov. Vidimo, da ta pristop rahlo poveča točnost, v dveh primerih od treh pa zmanjša oceno $F_1$. Različne izbire dodatnih primerov metrike spremenijo za nekaj odstotkov. V primerjavi z učenjem brez dodatnih primerov je še večja pristranskost k označevanju primerov kot implikacije, manj primerov pa označi kot nevtralne.

Iz primerjave rezultatov ocenjujemo, da učenje s tremi dodatnimi primeri glede na učenje brez dodatnih primerov bistveno ne izboljša napovedovanja. Brown idr. (2020) so pri testiranju modela GPT-3 na različnih nalogah ugotovili, da učenje z enim dodatnim primerom v povprečju izboljša dosežke modela, še bolj pa ga izboljša učenje s 50 dodatnimi primeri. V prihodnje bi bilo zato smiselno preveriti, ali bi z uporabo večjega števila dodatnih primerov lahko izboljšali rezultate.

**Tabela 14:** *Rezultati v % pri uporabi modela GPT-3.5-turbo z učenjem z nekaj dodatnimi primeri*

| Pristop | Točnost | Ocena F1 | Natančnost | Priklic |
|---|---|---|---|---|
| *nekaj-primerov 1* | 60 | 53,0 | 59,3 | 53,9 |
| *nekaj-primerov 2* | **63** | **55,2** | **62,7** | **56,9** |
| *nekaj-primerov 3* | 61 | 53,3 | 60,7 | 54,5 |
| *brez in navodilo-1-en* | 59 | 54,6 | 57,3 | 54,6 |

*Opomba.* V prvih treh vrsticah so rezultati za tri različne naključne izbire dodatnih primerov, v zadnji pa so za primerjavo rezultati z istim navodilom brez dodatnih primerov.

### *5.4.3 Generiranje razlag*

Za preverjanje uspešnosti generiranja razlag je model GPT-3.5-turbo z uporabo učenja brez dodatnih primerov z navodilom *navodilo-razlaga*, prirejenem po *navodilo-1-en*, za 100 primerov najprej generiral razlage, nato pa je primere še klasificiral. Ker je razlaga generirana najprej, je zaradi mehanizma pozornosti v arhitekturi transformer vplivala tudi na klasifikacijo.

Klasifikacijske metrike za ta pristop so podane v predzadnji vrstici Tabele 12. Predhodno generiranje razlage napovedovanja ne izboljša. Vidimo, da je točnost manjša kot pri navodilih brez razlag, predvsem zaradi pogostega napačnega uvrščanja primerov drugih dveh razredov kot kontradikcije. Pogostost posameznih vrst napak je povsem spremenjena, kar še dodatno kaže na to, da je uspešnost reševanja problema z učenjem brez dodatnih primerov zelo občutljiva na točno formulacijo navodila in zahtevan pristop reševanja. S preizkušanjem več različnih navodil za ta pristop bi lahko rezultate verjetno izboljšali.

Na podmnožici 50 primerov smo razlage ročno ovrednotili. Za pravilno klasificirane primere je bilo ustreznih 81 % razlag (86 % za pravilno klasificirane primere implikacije, 85 % za primere kontradikcije in 71 % za nevtralne primere; opozarjamo, da je uporabljen vzorec premajhen za zanesljivo primerjavo po razredih).

Sposobnost logičnega sklepanja tako za razliko od manjših modelov, uporabljenih v prejšnjih treh sklopih, pri tem modelu ni pogojena le z uporabo relativno preprostih jezikovnih hevristik, pač pa dejansko daje rezultate, ki kažejo na razumevanje.

**Kvalitativno vrednotenje** Trije primeri z ustreznimi razlagami so prikazani v Tabeli 15. Razlage so jasne in jezikovno pravilne, nekatere pa so slogovno nerodne (npr. tretji primer v Tabeli 15). Več lahko izvemo iz pregleda napačnih razlag, saj nam to razloži vzroke napak.

Trije primeri s tipičnimi neustreznimi razlagami so prikazani v Tabeli 16. Prvi primer ima razlago, ki je sicer tehnično skoraj pravilna, a neustrezna. Hipoteza je res le parafrazirana premisa in je zato implicirana, vendar bi razlaga morala to bolj jasno ponazoriti (npr. poudariti, da *veliki čezmerni odmerki* pomenijo *prekomerno zaužitje*). Drugi

primer vsebuje napačno razlago, kljub temu pa je klasifikacija pravilna. Model očitno ne ve, da je stoletnica specifična vrsta obletnice, in verjame, da je to vzrok za kontradikcijo. V tretjem primeru sta napačni tako razlaga kot klasifikacija. Model je v hipotezi pri besedni zvezi *samo policisti* obravnaval le drugo besedo in zanemaril prvo, ki pa je ključna za pravilno klasifikacijo.

Napake, podobne tisti v prvem primeru, bi verjetno lahko odpravili z bolj natančnimi navodili ali podajanjem nekaj primerov ustreznih razlag poleg navodil. Te napake so namreč posledica tega, da model nima informacije o tem, kakšna mora biti ustrezna razlaga.

Napaka v drugem primeru kaže na pomanjkljivo razumevanje odnosov med slovenskimi besedami. Pogosta vrsta napake je, da model besedo iz hipoteze in besedo iz premise, ki sta sopomenki ali je ena nadpomenka in ena podpomenka, obravnava kot različni in povedi zato kot kontradiktorni ter na tej osnovi uvrsti primer kot kontradikcijo. To razloži tudi prepogosto uvrščanje primerov v razred kontradikcija v matriki zamenjav.

Iz vrednotenja generiranja razlag lahko zaključimo, da model GPT-3.5-turbo kaže dokaj dobro razumevanje resničnosti, pri čemer pa ga omejuje slabše poznavanje slovenščine. Večino klasifikacij je zmožen ustrezno utemeljiti, po čemer se razlikuje od manjših slovenskih modelov. Pristop bi verjetno lahko še izboljšali s preizkušanjem različnih, morda bolj natančnih navodil, in z uporabo učenja z dodatnimi primeri.

Glede na neuspeh slovenskih modelov SloT5 pri isti nalogi menimo, da je za generiranje razlag pri logičnem sklepanju v slovenščini trenutno najboljši pristop uporaba zelo velikih jezikovnih modelov z več milijardami parametrov, četudi niso bili naučeni specifično za uporabo na slovenščini.

**Tabela 15:** *Primeri iz množice SI-NLI s pravilnimi razlagami in klasifikacijami, generiranimi z modelom GPT-3.5-turbo*

| | |
|---|---|
| Oznaka | *nevtralno* |
| Premisa | Vstala je in stopila k oknu, s katerega se je videlo na dvorišče. |
| Hipoteza | Pogledala je skozi okno, da bi videla, kaj je na dvorišču. |
| Odgovor modela | Iz podanega stavka ni mogoče sklepati, ali je pogledala skozi okno ali ne, zato je izjava nevtralna. – *nevtralno* |
| Oznaka | *kontradikcija* |
| Premisa | Pisni izpit ocenijo zunanji ocenjevalci po merilih, ki jih pripravi DPK SM za matematiko. |
| Hipoteza | Vsak učenec je svoje pisno preverjanje znanja ocenil sam s pomočjo sošolcev, pri ocenjevanju so postavili kar svoje lasnte kriterije. |
| Odgovor modela | Izjava je v nasprotju s podanim stavkom, saj pravi, da so učenci sami ocenili svoje pisno preverjanje znanja, medtem ko je v podanem stavku navedeno, da pisni izpit ocenijo zunanji ocenjevalci. – *kontradikcija* |
| Oznaka | *implikacija* |
| Premisa | Slovenija je še vedno najmanj zadolžena članica držav evrobmočja in se je tej ravni zadolženosti zaradi hitre gospodarske rasti doslej približevala postopoma. |
| Hipoteza | V Evropski uniji Slovenija še vedno velja za eno izmed držav z najmanj dolgovi, k čemur je pripomogel hiter, a postopen, vzpon gospodarstva. |
| Odgovor modela | Izjava je v skladu s podanim stavkom, saj oboje trdi, da je Slovenija ena izmed držav z najmanj dolgovi v Evropski uniji, pri čemer je k temu prispevala hitra gospodarska rast. – *implikacija* |

**Tabela 16:** *Primeri iz množice SI-NLI z napačnimi ali neustreznimi razlagami, generiranimi z modelom GPT-3.5-turbo*

| | |
|---|---|
| Oznaka | *implikacija* |
| Premisa | Veliki čezmerni odmerki lahko povzročijo močno zaspanost, utrujenost in občutek vročine. |
| Hipoteza | Prekomerno zaužitje lahko sproži intenzivno utrujenost, zaspanost, celo občutek povišane temperature. |
| Odgovor modela | Izjava je enaka kot izvirna, zato je izjava potrjena. – *implikacija* |
| Oznaka | *kontradikcija* |
| Premisa | To častitljivo obletnico so počastili s številnimi prireditvami in z izdajo obsežnega zbornika. |
| Hipoteza | Praznovali so stoletnico in ob tem pripravili dva koncerta ter izdali kratko brošuro. |
| Odgovor modela | Izjava je v nasprotju s podanim stavkom, saj omenja stoletnico, medtem ko je v podanem stavku omenjena obletnica. – *kontradikcija* |
| Oznaka | *kontradikcija* |
| Premisa | Kraj nesreče sta si ogledali tudi preiskovalna sodnica in državna tožilka. |
| Hipoteza | Okoliščine nesreče so si prišli ogledat samo policisti. |
| Odgovor modela | Iz podanega stavka ni mogoče sklepati ali so si okoliščine nesreče ogledali samo policisti ali ne, saj se o tem ne govori. – *nevtralno* |

*Opomba.* Prvi primer ima razlago, ki je sicer tehnično skoraj pravilna, a neustrezna. Drugi primer vsebuje napačno razlago, kljub temu pa je klasifikacija pravilna. V tretjem primeru sta napačni tako razlaga kot klasifikacija.

# 6 Diskusija

V tem razdelku združimo najpomembnejše rezultate iz različnih sklopov poskusov, ki so bolj podrobno predstavljeni v prejšnjem razdelku. Dodatno jih interpretiramo in jih postavimo v širši kontekst obstoječega dela.

## 6.1 Uvrščanje

Za uvrščanje primerov smo preizkusili več pristopov z uporabo modelov SloBERTa in GPT-3.5-turbo. Metrike so predstavljene v Tabeli 17.

**Tabela 17:** *Metrike v %, izračunane na napovedih za testno množico SI-NLI, za štiri različne pristope*

| Model | Točnost | Ocena F1 | Natančnost | Priklic |
|---|---|---|---|---|
| SloBERTa (*SI-NLI-celotna*) | 73,2 | 73,2 | 73,3 | 73,2 |
| SloBERTa (*ESNLIsi*) | 65,4 | 65,2 | 67,1 | 65,2 |
| SloBERTa (*ESNLIsi SI-NLI*) | **75,3** | **75,3** | **75,3** | **75,4** |
| GPT-3.5-turbo | 56,5 | 54,5 | 61,3 | 55,4 |

*Opomba.* V oklepajih so navedene učne množice, pri zadnjem pristopu dodatnega učenja ni bilo.

Model SloBERTa, učen na množici SI-NLI, je dosegel klasifikacijsko točnost 73,2 %, bi pa lahko dosegel boljše rezultate, če bi bila učna množica večja. Izločitev učnih primerov, na katerih se človeški označevalci motijo, zmanjša točnost napovedi, podobno oz. enako kot izločitev enakega števila naključno izbranih primerov. Na primerih, kjer se človeški označevalci motijo, se pogosteje kot pri ostalih motijo tudi jezikovni modeli. Domnevamo, da gre vsaj delno za dvoumne primere, torej takšne, ki jih tudi ljudje razumejo na različne načine; drugačne oznake nekaterih označevalcev torej niso napake, pač pa različne interpretacije istega primera. Posledično bi to informacijo lahko upoštevali pri učenju modelov in morda tako izboljšali rezultate.

Model SloBERTa, učen na prevodih ESNLI, je, kljub znatno večji učni množici, na testni množici SI-NLI dosegel skoraj 10 % manjšo točnost. Ugotavljamo, da je prenos znanja med različnimi množicami primerov logičnega sklepanja relativno slab, izboljšuje pa se z večanjem števila učnih primerov. Kljub temu pa model, učen na strojnih

prevodih, pravilno uvršča skoraj dve tretjini primerov na drugi množici. Ta pristop je enostavno posplošiti na druge jezike z malo viri, v katerih podatkovne množice primerov logičnega sklepanja morda ne obstajajo, zato je pristop, kljub slabšim rezultatom, uporaben.

Najboljši rezultat dosežemo z vnaprejšnjim učenjem modela SloBERTa na množici ESNLI in z nadaljnjo prilagoditvijo na množici SI--NLI; na testni množici SI-NLI je bila dosežena točnost 75,3 %. Angleške prevode podatkovnih množic torej lahko uporabimo za izboljšanje rezultatov na SI-NLI, vendar pa naš rezultat ne dosega najboljšega rezultata, objavljenega na SloBench (točnost 77,2 %) (CJVT UL, 2023).

Z uporabo GPT-3.5-turbo smo dosegli slabši rezultat, tako pri učenju brez dodatnih primerov kot tudi z nekaj dodatnimi primeri, čeprav je model mnogo večji od modela SloBERTa. Občutljiv je na izbiro ukaznega navodila in na izbor primerov pri učenju z nekaj dodatnimi primeri. Z bolj širokim testiranjem različnih navodil bi rezultat najverjetneje lahko izboljšali.

## 6.2 Generiranje razlag

Razlage smo poskusili generirati z dvema različno velikima modeloma, SloT5 in GPT-3.5-turbo. Poskus generiranja razlag z modelom SloT5 je bil neuspešen. Ustrezne razlage so modeli generirali za manj kot tretjino primerov (28 %). Ugotovili smo, da tudi večji model SloT5 ni boljši od manjšega. Modeli se dobro naučijo zgolj forme, niso pa zmožni ustvariti pomensko smiselnih razlag.

Sklepamo, da sta ta dva slovenska velika jezikovna modela z do nekaj sto milijonov parametrov zmožna iskanja in uporabe jezikovnih vzorcev, poznavanje jezika pa ni v celoti povezano s poznavanjem resničnosti. To je v skladu z ugotovitvami McCoy idr. (2019), da jezikovni modeli za reševanja problema logičnega sklepanja uporabljajo različne jezikovne hevristike.

Uporaba podobnih hevristik namesto dejanskega razumevanja sveta je zato verjetno razlaga za slab prenos znanja, slabo posploševanje, drugačne tipe napak glede na človeške in nezmožnost generiranja pravilnih razlag. Pristop velikih jezikovnih modelov k reševanju zastavljenega problema po tej domnevi temelji na procesiranju jezika

namesto na poznavanju zakonitosti resničnega sveta in zdravorazumskega sklepanja, kot to počnemo ljudje.

Boljše rezultate smo dosegli z modelom GPT-3.5-turbo in učenjem brez dodatnih primerov. Pri pravilno uvrščenih primerih, ki so predstavljali približno polovico, je bilo ustreznih 81 % razlag. Glede na ta rezultat in slab rezultat SloT5 menimo, da je za nadaljnje raziskovanje generiranja razlag pri logičnem sklepanju za slovenščino najbolj smiselna uporaba velikih jezikovnih modelov z več milijardami parametrov, tudi če ti niso bili učeni specifično za slovenščino. Koristil bi tudi veliki model, naučen na zadostnem številu slovenskih besedil.

GPT-3.5-turbo je torej sposoben reševanja problemov s področja logičnega sklepanja v naravnem jeziku, čeprav ni bil učen ali prilagojen za to nalogo. Kaže dokaj dobro razumevanje resničnosti, pri čemer pa ga za našo nalogo omejuje slabše poznavanje slovenščine. Večino pravilnih klasifikacij je zmožen ustrezno utemeljiti, v čemer se razlikuje od manjših slovenskih modelov. Učenje zelo velikih modelov na veliki količini podatkov s spleta da modelom zadostno razumevanje pomena jezika in razumevanje sveta za uspešno reševanje problema logičnega sklepanja in utemeljitev sklepov.

## 7 Zaključek

Raziskovali smo različne pristope k logičnemu sklepanju v naravnem jeziku za slovenščino. Preizkusili smo več velikih jezikovnih modelov za uvrščanje primerov in generiranje razlag in v slovenščino strojno prevedli ter objavili 50.000 primerov iz angleške podatkovne množice ESNLI.

Ugotovili smo, da je prenos znanja med različnimi množicami primerov logičnega sklepanja relativno slab, izboljšuje pa se z večanjem števila učnih primerov. Najboljši rezultat smo dosegli z vnaprejšnjim učenjem modela SloBERTa na množici ESNLI in nadaljnjo prilagoditvijo na množici SI-NLI. Angleške prevode podatkovnih množic torej lahko uporabimo za izboljšanje rezultatov na slovenski množici SI-NLI.

Poskus generiranja razlag z modelom SloT5 je bil neuspešen, bolj uspešen pa je bil z mnogo večjim GPT-3.5-turbo. Sklepamo, da so slovenski veliki jezikovni modeli z nekaj sto milijoni parametrov zmožni iskanja in uporabe jezikovnih vzorcev, poznavanje jezika pa ni v celoti

povezano s poznavanjem resničnosti. Za nadaljnje raziskovanje generiranja razlag pri logičnem sklepanju za slovenščino predlagamo uporabo velikih jezikovnih modelov z več milijardami parametrov, tudi če niso bili učeni specifično za slovenščino. Koristil bi tudi veliki model, naučen na zadostni množici slovenskih besedil.

Testirane pristope bi lahko še izboljšali. Če bi namesto ali poleg prevodov množice ESNLI za vnaprejšnje učenje modela SloBERTa uporabili prevode množice, ki je bolj raznolika od ESNLI ali pa so primeri v njej po izvoru bolj podobni tistim v SI-NLI, bi verjetno lahko na testni množici SI-NLI dosegli boljše rezultate. Prav tako bi rezultate morda lahko izboljšali z uporabo informacije o dvoumnih primerih oziroma o nestrinjanju označevalcev. Oboje zahteva nadaljnje testiranje.

Največ potenciala za izboljšanje je pri generiranju razlag in uporabi GPT-3.5-turbo. Z bolj širokim testiranjem različnih navodil, tako za klasifikacijo kot za generiranje razlag, bi najverjetneje lahko izboljšali rezultate pri obeh nalogah. Prav tako bi bilo smiselno preizkusiti učenje z več deset ali več sto dodatnimi primeri, ki ponavadi izboljša dosežke modela (Brown idr., 2020). To zlasti velja za generiranje razlag, saj smo tam preizkusili le eno navodilo in učenje brez dodatnih primerov.

V času našega testiranja je bil dostop do modela GPT-4 (OpenAI, 2023a), ki je izboljšana različica modela GPT-3.5-turbo, omejen. Ta model dosega boljše rezultate pri večini nalog. V prihodnje bi bilo smiselno preveriti, če in koliko boljše napovedi in razlage bi lahko dobili z njim.

Tako GPT-4 kot GPT-3.5-turbo sta v lasti podjetja OpenAI in sta dostopna le prek vmesnika tega podjetja, točni podatki o njuni zgradbi pa niso znani. Obstajajo javno dostopni odprti modeli, ki so po zmogljivosti primerljivi ali skoraj primerljivi z njima, kakršen je na primer LLaMa-2 (Touvron idr., 2023). Podobne preizkuse bi lahko izvedli tudi na katerem od teh odprtih modelov.

## Zahvala

Delo je podprla Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije (ARIS) iz državnega proračuna preko raziskovalnega programa št. P6-0411 (Jezikovni viri in tehnologije za slovenski jezik) in projekta PROP – Empirična podlaga za digitalno podprt razvoj pisne jezikovne zmožnosti (št. J7-3159).

## Literatura

Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, ..., & Askell, A. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877–1901.

Camburu, O.-M., Rocktäschel, T., Lukasiewicz, T., & Blunsom, P. (2018). e-SNLI: Natural Language Inference with Natural Language Explanations. *Advances in Neural Information Processing Systems*, *31*.

CJVT UL. (2023). *SloBench – Natural language inference (SI-NLI) leaderboard*. Pridobljeno s https://slobench.cjvt.si/leaderboard/view/9

*DeepL Translate API*. (2023). Pridobljeno s https://www.deepl.com/pro-api

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long and Short Papers)* (str. 4171–4186).

Erjavec, T., Fišer, D., & Ljubešić, N. (2021). The KAS corpus of Slovenian academic writing. *Lang. Resour. Eval.*, *55*(2), 551–583.

Fišer, D., Erjavec, T., & Ljubešić, N. (2016). JANES v0.4: Korpus slovenskih spletnih uporabniških vsebin. *Slovenščina 2.0: empirične, aplikativne in interdisciplinarne raziskave*, *4*(2), 67–99.

*Google Prevajalnik*. (2023). Pridobljeno s https://translate.google.com/?hl=sl

Klemen, M., Žagar, A., Čibej, J., & Robnik-Šikonja, M. (2022). *Slovene Natural Language Inference Dataset SI-NLI*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1707

Klemen, M., Žagar, A., Čibej, J., & Robnik-Šikonja, M. (2024). SI-NLI: A Slovene Natural Language Inference Dataset and its Evaluation. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Torino, Italia (str. 14859–14870). ELRA and ICCL. Pridobljeno s https://aclanthology.org/2024.lrec-main.1294.pdf

Krek, S., Arhar Holdt, Š., Erjavec, T., Čibej, J., Repar, A., Gantar, P., Ljubešić, N., Kosem, I., & Dobrovoljc, K. (2020). Gigafida 2.0: The Reference Corpus of Written Standard Slovene. *Proceedings of the Twelfth Language*

*Resources and Evaluation Conference*, Marseille, France (str. 3340–3345). European Language Resources Association. Pridobljeno s https://aclanthology.org/2020.lrec-1.409

Kumar, S., & Talukdar, P. (2020). NILE: Natural Language Inference with Faithful Natural Language Explanations. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (str. 8730–8742). Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.771

Lebar Bajec, I., Repar, A., Demšar, J., Bajec, Ž., Rizvič, M., Kumperščak, B., & Bajec, M.(2022). *Neural Machine Translation model for Slovene-English language pair RSDO-DS4-NMT 1.2.6*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1736

Liu, H., Ning, R., Teng, Z., Liu, J., Zhou, Q., & Zhang, Y. (2023). Evaluating the logical reasoning ability of ChatGPT and GPT-4. Pridobljeno s file:///C:/Users/student1/Downloads/Evaluating_the_Logical_Reasoning_Ability_of_ChatGP.pdf

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, ..., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. Pridobljeno s https://arxiv.org/pdf/1907.11692

Ljubešić, N., & Erjavec, T. (2011). hrWaC and slWac: compiling web corpora for Croatian and Slovene. *Proceedings of the 14th International Conference on Text, Speech and Dialogue* (str. 395–402). doi: 10.1007/978-3-642-23538-2_50

Logar, N., Erjavec, T., Krek, S., Grčar, M., & Holozan, P. (2013). *Written corpus ccKres 1.0*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1034

McCoy, T., Pavlick, E., & Linzen, T. (2019). Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy (str. 3428–3448). Association for Computational Linguistics. doi: 10.18653/v1/P19-1334

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica*, *22*(3), 276–282.

Müller, A., & Guido, S. (2016). *Introduction to Machine Learning with Python: A Guide for Data Scientists*. O'Reilly Media.

OpenAI. (2022). *Introducing ChatGPT*. Pridobljeno s https://openai.com/blog/chatgpt

OpenAI. (2023a). GPT-4 Technical Report. Pridobljeno s https://arxiv.org/pdf/2303.08774

OpenAI. (2023b). *Models – OpenAI API*. Pridobljeno s https://platform.openai.com/docs/models/gpt-3-5

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., ..., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, 27730–27744. Pridobljeno s https://arxiv.org/abs/2203.02155

Pančur, A., & Erjavec, T. (2020). The siParl corpus of Slovene parliamentary proceedings. *Proceedings of the Second ParlaCLARIN Workshop* (str. 28–34).

Poth, C., Pfeiffer, J., R"uckl'e, A., & Gurevych, I. (2021). What to Pre-Train on? Efficient Intermediate Task Selection. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (str. 10585–10605).

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, *21*(1), 5485–5551.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., ..., & Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. Pridobljeno s https://arxiv.org/abs/2307.09288

Ulčar, M., & Robnik-Šikonja, M. (2021). SloBERTa: Slovene monolingual large pretrained masked language model. *Proceedings of SI-KDD within the Information Society 2021* (str. 17–20).

Ulčar, M., & Robnik-Šikonja, M. (2023). Sequence-to-sequence pretraining for a less-resourced Slovenian language. *Frontiers in Artificial Intelligence*, *6*, 1–12. doi: 10.3389/frai.2023.932519

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Wang, S., Fang, H., Khabsa, M., Mao, H., & Ma, H. (2021). Entailment as few-shot learner. Pridobljeno s https://arxiv.org/pdf/2104.14690

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., ..., & Rush, A. (2020). Transformers: State-of-the-Art Natural Language Processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (str. 38–45).

Zhong, Q., Ding, L., Liu, J., Du, B., & Tao, D. (2023). Can ChatGPT understand too? A comparative study on ChatGPT and fine-tuned BERT. Pridobljeno s https://arxiv.org/pdf/2302.10198

## Natural language inference for Slovene

In recent years, large language models have been the most successful approach to natural language processing. An important problem in this field is natural language inference, which requires models to contain relatively broad general knowledge. Moreover, the requirement for models to explain their reasoning can offer additional insights into their functioning. We tested several approaches for natural language inference in Slovene. We used two Slovene large language models, SloBERTa and SloT5, as well as a much larger English model GPT-3.5-turbo. Training data consisted of the Slovene dataset SI-NLI and an additional 50,000 machine-translated samples from the English dataset ESNLI. The SloBERTa model was fine-tuned on both datasets. Fine-tuning it on the SI-NLI dataset achieved a classification accuracy of 73.2% on the SI-NLI test set. Pretraining it on the ESNLI dataset improved its accuracy to 75.3%. We observe that models make different types of errors compared to humans and that they generalize poorly across different datasets.

The SloT5 model was also fine-tuned on ESNLI to generate explanations for natural language inference samples. Less than a third of explanations were appropriate, with the model learning common sentence patterns from the domain and producing semantically meaningless explanations. We assume that the tested Slovene large language models with up to several hundred million parameters are capable of identifying and using language patterns, but their language understanding is not necessarily sufficient to understand reality. When the considerably larger GPT-3.5-turbo was used both for classification and explanation generation, it achieved an accuracy of 56.5% on the SI--NLI test set using zero-shot learning, but with 81% of the explanations being appropriate for the correctly classified samples. In comparison with smaller Slovene models, this model shows a reasonable understanding of reality but is limited by its lower Slovene proficiency.

Keywords: natural language inference, large language models, transformer architecture, SloBERTa, SloT5, GPT-3.5-turbo, ChatGPT, explanations, Slovene, fine-tuning