University *of Ljubljana*
Faculty *of Computer and Information Science*

Klasifikacija kratkih sporočil - IMapBook

Marko Katrašnik and Erica Zago

Povzetek

Klasifikacija besedil je dobro znan problem področja obdelave naravnega jezika. Posebno podpodročje predstavljajo spletni pogovori, saj so sporočila običajno zelo kratka in pogosto vsebujejo slovnične napake. V okviru te seminarske naloge sva se ukvarjala s sporočili, ki so nastala v okviru aplikacije IMapBook, v kateri se učenci pogovarjajo o oblikovanju skupnega odgovora na vprašanje o prebrani knjigi. Zbirka je razdeljena na dva dela. V prvem (CREW) je poleg sporočil podan še končni odgovor na katerega se nanašajo sporočila, v drugem (Discussion) tega podatka ni, so pa sporočila povprečno nekoliko daljša in jih je manj.
Podatke sva analizirala, očistila in preizkusila nekaj različnih postopkov klasifikacije besedil. Najprej sva oblikovala nabor ročnih značilk in uporabila MLP klasifikator. Ta klasifikator sva preizkusila tudi z uporabo tf-idf vektorjev. Uporabila sva še dve globoki metodi, in sicer BERT in DistilBERT. Napovedi sva poskusila izboljšati z uporabo podobnosti sporočil s končnim skupnim odgovorom in vsebino knjige.
Na delu zbirke CREW sva najboljše rezultate dosegla z uporabo modela BERT, vrednost mere makro F1 je bila $0,541$. Na sporočilih Discussion pa z naborom ročnih značilk, makro F1 je bil $0,749$. Uporaba podobnosti se ni izkazala za koristno pri klasifikaciji sporočil.

Ključne besede
Klasifikacija besedil, spletni pogovori, kratka besedila

Mentor: Slavko Žitnik

1. Uvod

Klasifikacija besedil je podpodročje obdelave naravnega jezika in je proces razvrščanja besedil v organizirane skupine. Namen te seminarske naloge je pripraviti algoritem, ki bo klasificiral sporočila, ki so nastala z uporabo aplikacije IMapBook. Sporočila so bila napisana v okviru knjižnih krožkov, kjer so sodelovale različne skupine. Vsaka skupina je imela nalogo, da skupaj sestavi odgovor na podano vprašanje, nanašajoče se na prebrano knjigo/poglavje. Preko sporočil so se dogovorili kako odgovoriti.

Ukvarjala sva se torej s klasifikacijo besedil (angl. text classification), bolj specifično s klasifikacijo govornih dejanj (angl. speech act classification) in klasifikacijo kratkih besedil (angl. short text classification) [1]. Za klasifikacijo kratkih besedil je značilno, da je klasifikacija nekoliko težja zaradi majhnega števila besed v posameznem sporočilu. Po drugi strani pa so pogosto na voljo dodatni podatki, kot na primer časovno sosledje sporočil in avtor posameznega sporočila. V zbirki IMapBook so poleg tega na voljo še besedila skupnih končnih odgovorov in vsebine knjig na katere se nanašajo sporočila, zato sva se odločila, da se osredotočiva na uporabo teh podatkov, ki so posebnost uporabljene zbirke.

V tem delu predstavljava štiri pristope h klasifikaciji kratkih sporočil spletnih pogovorov, in sicer nabor ročno pridobljenih značilk v kombinaciji z MLP (MultiLayer Perceptron), TF-IDF vektorje v kombinaciji z MLP, BERT in DistilBERT. Preizkusila sva tudi, če bi uporaba podobnosti sporočil z vsebino knjige in končnim skupnim odgovorom pomagala pri klasifikaciji. Preizkusila sva dve metodi, in sicer z uporabo TF-IDF vektorjev in BERT vektorskih vložitev. V obeh primerih se je za računanje podobnosti med vektorji uporabila kosinusna podobnost.

Preostanek poročila je sestavljen iz naslednjih poglavij. V poglavju 2 so predstavljene ugotovitve iz literature, v kateri so se ukvarjali s podobnimi problemi. V 3 so opisane metode, ki sva jih uporabila za klasifikacijo sporočil. V poglavju 4 je podrobneje predstavljena uporabljena zbirka podatkov in izvedba eksperimentov. V naslednjem poglavju 5 so predstavljeni rezultati, v 6 sledi še zaključek s kratko diskusijo in idejami za nadaljnje delo.

2. Sorodna dela

V tem poglavju je predstavljenih nekaj relevantnih del, v katerih so se ukvarjali s podobnimi problemi, kot je bil naš.

V članku [1] se avtorji ukvarjajo s problemom kratkih besedil, kjer izpostavljajo, da tradicionalne metode, zaradi omejenega števila besed, ne morejo predstavljati prostora značilk in razmerja med besedami in dokumenti. Kratka besedila imajo svoje značilnosti, kot so redkost (angl. sparseness), obsežnost (angl. large-scale), neposrednost (angl. immediacy) in nestandardnost (angl. non-standardability). Predlagana rešitev klacifikacje predvidi zmanjševanje dimenzij značilk in ekstrakcijo značilk z uporabo semantičnih razmerij, kot so npr. LSA, LDA klasifikacijski modeli. Ker je na voljo veliko več neoznačenega (angl. unlabeled text), kot označenega besedila, so priporočili uporabo polnadzorovanega (angl. semi-supervised) algoritma klasifikacije. Za izboljšanje točnosti (angl. accuracy) se lahko uporabi klacifikacijo z ansambli (angl. ensemble classification).

Med najbolj znanimi socialnimi omrežji, ki vsakodnevno prispeva na milijone kratkih sporočil je Twitter. Delo [2] se ukvarja s prepoznavanjem govornega dejanja (angl. speech act recognition) na Twitterju, tako da ga obravnava kot večrazredni problem klasifikacije. Sporočila so bila razdeljena in ročno označena v 5 razredov in sicer: trditev (angl. assertion), priporočilo (angl. recommendation expression), vprašanje (angl. question), zahteva (angl. request) in razno (angl. miscellaneous). Izbrane značilke za obravnavo sporočil so bile razdeljene na semantične in sintaktične. Semantične značilke zavzemajo mnenjske besede (angl. opinion words), vulgarne besede, čustvene simbole, glagole govornega dejanja (angl. speech act verbs) in N-grame. Sintaktične značilke pa obravnavajo ločila, značilne Twitterjeve znake, okrajšave in podobno. Za nadzorovano klasificiranje so bili uporabljeni Naivni Bayes, odločitvena drevesa, logistična regresija, SVM in osnovni maksimalni klasifikator (angl. baseline max classifier). Najboljše rezultate so dosegli z uporabo logistične regresije.

Najbolj ranljiva skupina, ko se govori o socialnih omrežjih in o kratkem in hitrem sporočanju so seveda otroci in najstniki. S to problematiko se sooča članek [3], kjer avtorji predstavijo samodejno analizo razpoloženja otrok glede na njihova kratka besedila. Prvotno je bilo opredeljenih 7 razredov nasilnih tem, ki so bili nato združeni v 4. Prvi zavzema agresijo in nasilje, drugi stisko, tesnobo in depresijo, tretji spolnost in četrti snovi/mamila. Za klasifikacijo so bili uporabljeni SVM, random forest (RF) in BERT. Za izvedbo RF je bilo uporabljenih 10 dreves z največjo globino 2. Za implementacijo SVM in RF so uporabili zbirko orodij Scikit-learn. Model BERT je bil vzpostavljen z uporabo implementacije TensorFlow, v večjezični različici. Ob upoštevanju mere F1 so najboljše rezultate dosegli z SVM in BERT.

V članku [4] so predstavili raziskavo o klasifikaciji sporočil klepetalnice. V delu so bili uporabljeni 3 nabori podatkov in sicer "chat lines", "chat topics" in "broadcast messages", pogovori so potekali med skupnim ogledom filma. Prvotnih 8 razredov so združili v tri, in sicer pogovor o filmu, pogovor o osebnih temah in pogovor o študiju. Analiza se je pričela s pretvorbo vsake vrstice klepeta (ali bloka teme) v vektor besed. Sledilo je korenjene besed (angl. steamming), odstranjevanje stop besed in frekvenčno izločanje (angl. frequency cutoff). Frekvenčno izločanje se ukvarja z odstranjevanjem besed iz nabora podatkov, ki se ne pojavljajo pogosto. Za klasifikacijo so uporabili dva algoritma, in sicer Naivni Bayes in Viterbijev algoritem za skrite Markove modele (HMM). Naivni Bayes se je izkazal za boljšega in je imel višjo točnost od HMMja.

Za klasifikacijo besedil se sicer v zadnjem času uporabljajo tudi naprednješi algoritmi, ki upoštevajo tudi vrstni red besed[1]. Za vektorizacijo se uporablja na primer GloVe embedding, za klasifikacijo pa nato konvolucijske (CNN) ali rekurenčne (RNN) nevronske mreže. Težava teh pristopov pa je, da za uspešno učenje nevronskih mrež potrebujemo velike zbirke podatkov, zato takih pristopov na podatkih IMapBook nisva uporabila.

3. Metode

V tem razdelku so na kratko predstavljene metode, ki sva jih uporabila v okviru te seminarske naloge.

3.1 Ročne značilke

Na podlagi značilnih lastnosti sporočil v posameznem razredu, sva oblikovala nabor ročnih značilk, za katerega sva predvidevala, da bo omogočal dobro razlikovanje med razredi. V nabor sva vključila naslednje značilke:

- število besed v sporočilu,
- povprečna dolžina besede v sporočilu,
- dolžina najdaljše besede v sporočilu,
- frekvenca besed, ki se začnejo na w ali h,
- frekvenca velikih tiskanih črk v sporočilu,
- prisotnost spletne povezave v sporočilu,
- prisotnost besede "ok",
- prisotnost besede "thank",
- prisotnost pozdrava ("hi", "hello", "good morning"),
- ali je v sporočilu le emoji,
- prisotnost emoji-ja v sporočilu,
- prisotnost vprašaja,
- prisotnost klicaja.

Poskusila sva dodati tudi značilko, ki zajame frekvenco besed, ki niso pravilno napisane, vendar se je izkazalo, da je slovar angleških besed preobsežen, posledično je iskanje vseh besed zahtevalo preveč časa, da bi bila taka značilka praktično uporabna.

3.2 TF-IDF

TF-IDF je vrsta meritve, ki ocenjuje, koliko je neka beseda pomembna za dokument v zbirki dokumentov. Izračuna se tako, da se pomnoži dve metriki: kolikokrat se beseda pojavi

[1] https://developers.google.com/machine-learning/guides/text-classification

v dokumentu (angl. Term Frequency - tf) in obratno pogostost dokumenta v nizu dokumentov (angl. Inverse Document Frequency - idf).

3.3 MLP

MLP (angl. Multilayer Perceptron) je klasični tip nevronske mreže. Sestavljen je iz ene ali več plasti nevronov. Podatki so podani na vhodno plast, nato pa gredo skozi eno ali več skritih plasti, ki zagotavljajo ravni abstrakcije. Na izhodni plasti se pa izračunajo napovedi oziroma verjetnosti posameznih razredov.

3.4 BERT

BERT (angl. Bidirectional Encoder Representations from Transformers) [5] je dvosmerni transformator, ki je bil predhodno treniran s kombinacijo modeliranja z zamaskiranim jezikom in napovedovanja naslednjega stavka na velikem korpusu, ki ga sestavljata Toronto Book Corpus in Wikipedia.

3.5 DistilBERT

DistilBERT [6] oziroma prečiščena različica modela BERT je bila zasnovana z namenom, da bi bila na voljo manjša in hitrejša različica tega kompleksnega modela, ki pa za razliko od večine poskusov poenostavljanja tega modela ni fokusirana na specifično nalogo. Tako je uporaba zelo podobna uporabi klasičnega modela BERT, le da je model približno pol manjši in omogoča hitrejše izvajanje.

3.6 Uporaba besedila knjig in skupnega končnega odgovora

Želela sva ugotoviti tudi, ali ima uporaba podatkov besedila knjig in skupnega končnega odgovora pozitiven vpliv na uspešnost napovedovanja razredov posameznih sporočil. Glede na to, da ena podmnožica sporočil (discussion) nima podanega podatka o končnem odgovoru, ima pa precej daljša besedila sporočil, sva lahko primerjala tudi koristnost teh dveh lastnosti. Pomembno je poudariti tudi, da bi bil algoritem, ki uporablja skupni končni odgovor uporaben le za kakšno analizo pogovora, ko je ta že zaključen, saj končni odgovor med pogovorom ni še znan.

Besedila knjig in skupne končne odgovore sva uporabila tako, da sva izračunala podobnost s sporočilom. Predvidevala sva namreč, da so si sporočila, v katerih so govorili o vsebini (npr. content discussion) bolj podobna besedilu knjige in končnemu odgovoru. Za vektorizacijo sporočil sva najprej uporabila TF-IDF, kasneje pa sva ga nadomestila z BERT vektorskimi vložitvami. Podobnost sva računala s kosinusno podobnostjo.

Podatka o podobnosti z besedilom knjige in končnim odgovorom, sva nato skupaj z napovedjo razreda sporočila (ta napoved je bila pridobljena z zgoraj opisanimi postopki, ki uporabljajo samo vsebno sporočila) poslala na vhod še enega algoritma strojnega učenja. Uporabila sva naključni gozd (angl. random forest).

Shema algoritma klasifikacije sporočil z uporabo besedil knjig in končnih skupnih odgovorov je prikazana na sliki 1 .

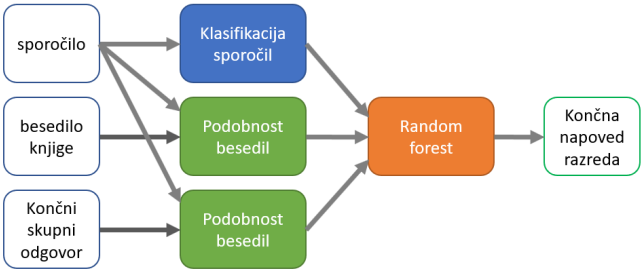

Figure 1. Shema algoritma z uporabo besedila knjig in končnih odgovorov.

Preizkusila sva dva načina združevanja podatkov. Pri prvem sva združila napoved razreda, ki ga vrne MLP izhodiščnega algoritma in kosinusno podobnost. Tako sva imela le dve ali tri vhodne značilke za končni klasifikator z naključnim gozdom. Pri drugem načinu pa sva združila vektor z verjetnostmi posameznih razredov in kosinusne podobnosti.

3.7 Metrike za evalvacijo

Uporabila sva metrike: točnost (angl. accuracy), makro preciznost (angl. precision), makro priklic (angl. recall) in makro mero F1. Pri analizi rezultatov nama je zelo prav prišla konfuzijska matrika (angl. confusion matrix). S pomočjo te sva lahko videla v katere razrede in v kolikšni meri so bila napačno klasificirana sporočila.

4. Eksperimenti

4.1 Zbirka podatkov

Za razvoj algoritmov, učenje modelov in testiranje sva uporabila zbirko podatkov, ki je nastala ob uporabi aplikacije IMapBook.

4.1.1 Izbira uporabljenih podatkov iz zbirke

Zbirka podatkov je podana v obliki Excel tabele, v kateri je v vsakem stolpcu podan en tip podatkov. Po pregledu zbirke sva identificirala podatke, ki sva jih uporabila za klasifikacijo.

- Najpomembnejši podatek je “Message”, vsebuje vsebino sporočila, kateremu želimo napovedati kategorijo.
- “CodePreliminary” je oznaka kategorije in se uporablja za učenje algoritmov in za evalvacijo.
- “Collab Response” in “Book ID” (povezano z besedilom knjige) sta uporabljena za računanje podobnosti s sporočilom.
- “isAnswer” in “Page” lahko zanemarimo, kot lahko tudi “Topic”, ker vprašanje ni dovolj specifično.

4.1.2 Obdelava zbirke podatkov

Zbirka podatkov IMapBook ni bila še prečiščena, zato sva podatke pregledala in naredila nekaj sprememb.

Nekaj sporočil je imelo oznake razredov, ki niso bile pričakovane glede na opise razredov, ki so podani v ločeni datoteki v zbirki podatkov. Glede na to, da je bilo v vsakem od takih razredov zelo malo sporočil in ker so bili po vsebini

podobni sporočilom v predvidenih razredih, sva jih pridružila k razredom, ki so bili glede na vsebino sporočil najbolj primerni. H katerim razredom sva pridružila sporočila iz posameznega nepričakovanega razreda je prikazano v tabeli 1.

Table 1. V tabeli je prikazano h katerim razredom so bile pridružene nepričakovane oznake razredov.

Izvroni razred	Dodano k razredu
assignment instructions question	assignment instruction
general discussion	general comment
observation	general comment
outside material	external material

Pri štirih sporočilih je manjkala oznaka id-ja, ki označuje na katero knjigo se nanaša diskusija. Glede na to, da so sporočila očitno pripadala k sporočilom pred in za takimi sporočili, sva dodala enak id knjige kot je bil določen v okolici.

V treh primerih skupina ni podala končnega odgovora. Oznake v teh primerih niso bile usklajene, zato sva pri vseh vnesla <Nothing>, kot je bilo zapisano pri enem od treh primerov brez končnega odgovora.

V zbirki podatkov je bilo nekaj razredov z zelo malo pripadajočimi sporočili. Ta težava je še posebej izrazita v delu zbirke discussion. S testiranjem delovanja in v želji, da bi v obeh delih zbirke podatkov uporabila enako pravilo za odstranjevanje razredov z malo primeri, sva določila, da so odstranjena spročila iz razredov z manj kot petimi primeri. Odstranjeni razredi so označeni v tabelah 2 in 3.

V dveh primerih je manjkal podatek o avtorju sporočila, oziroma njegov psevdonim (Pseudonym). Na ti dve mesti sva napisala besedo “blank”.

4.1.3 Analiza podatkov

Prečiščene podatke sva nato analizirala, da bi dobila boljši vpogled v zbirko in na tej podalgi lažje določila nadalnje korake.

Število sporočil:

- 705 sporočil v prvem zavihku (CREW data) - 711 pred odstranjevanjem razredov
- 122 sporočil v drugem zavihku (Discussion only) - 130 pred odstranjevanjem razredov

V tabeli 2 so prikazani razredi sporočil v zavihku CREW data, v tabeli 3 pa so prikazani razredi iz zavihka Discussion. Pri obeh je podano tudi število sporočil, ki pripadajo posameznemu razredu.

Opazna je zelo neenakomerna porezdelitev števila sporočil po kategorijah, kar je bilo potrebno upoštevati pri uporabi algoritmov, še bolj pa pri uporabi in interpretaciji metrik za evalvacijo.

Na slikah 2 in 3 sta prikazani distribuciji dolžin sporočil (število besed v sporočilu) v zavihkih CREW data in Discussion. Distribuciji sta si precej različni, v prvi je veliko več kratkih sporočil, druga je bolj enakomerna, prisotnih je tudi nekaj daljših sporočil. Povprečna dolžina sporočila v CREW data je 11 besed, v Discussion pa 44,5 besed.

Table 2. CREW data - Razredi in število sporočil v vsakem razredu.

Razred	št. sporočil
assignment instructions	35
assignment question	10
content discussion	299
content question	18
discussion wrap-up	12
emoticon/non-verbal	13
external material	13
feedback	31
general comment	66
general question (odstranjeno)	2
greeting	11
incomplete/typo	10
instruction question (odstranjeno)	4
logistics	80
opening statement	8
response	99

Table 3. Discussion only data - Razredi in število sporočil v vsakem razredu.

Razred	št. sporočil
assignment instructions (odstranjeno)	1
content discussion	94
content question	5
emoticon/non-verbal (odstranjeno)	3
feedback (odstranjeno)	3
general comment	6
incomplete/typo	10
instruction question (odstranjeno)	1
response	7

Analizirala sva tudi kateri učenci so bili najbolj aktivni. Aktivnost študenta je bila določena na podlagi števila sporočil, ki jih je poslal. Za CREW data so bili najbolj aktivni: edf-16 s 104 sporočili, edf-15 s 104, pim-11 s 54, pim-12 s 44 in pim-04 s 39 sporočili. Za Discussion only pa: pim-30 s 10 sporočili, dig-11 s 8, pim-05 s 5, pim-14 s 5 in pim-07 s 5 sporočili.

Poleg tega sva analizirala tudi, kdo je največ prispeval h končnemu odgovoru, oziroma sporočila katerega učenca so si povprečno najbolj podobna s končnim odgovorom. Rezultati so prikazani v tabeli 4. Podobnosti sva računala z uporabo kosinusne podobnosti med tf-idf vektorji sporočil in končnimi odgovori.

Po istem principu je možno ugotoviti kdo je pisal sporočila, ki so bila najbolj podobna vsebini knjige. Za knjigo ”The Lady or the Tiger” je oseba s preudonimom pim-02 dosegla podobnost $0,230$, s knjigo ”Just Have Less” so bila najbolj podobna sporočila učenca s preudonimom dig-04 s podobnostjo $0,260$

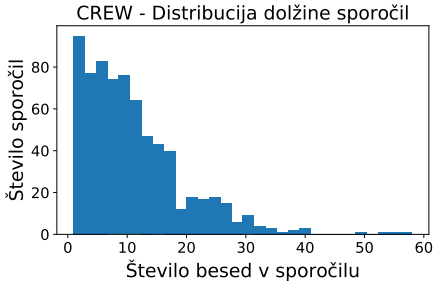

Figure 2. Distribucija dolžine sporočil v zavihku CREW data.

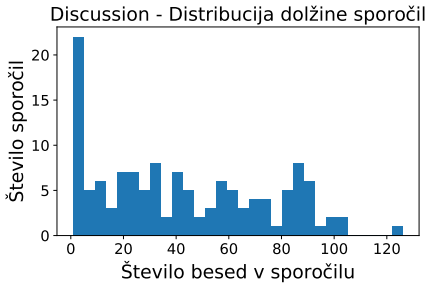

Figure 3. Distribucija dolžine sporočil v zavihku discussion.

in za knjigo ”Design for the Future When the Future Is Bleak” je oseba s preudonimom dig-14 dosegla podobnost $0,231$.

4.1.4 Podobnosti vsebine sporočil s končnimi odgovori in vsebinami knjig

Glede na to, da je eden od namenov te seminarske naloge, da ugotoviva, ali podatki o končnem odgovoru in vsebini knjige na katero se nanašajo sporočila pripomorejo k uspešnosti napovedovanja razredov sporočil, sva preverila kakšna je povprečna podobnost med sporočili posameznega razreda in končnimi odgovori, ter vsebinami knjig.

Tudi na tem mestu sva podobnost računala z uporabo kosinusne podobnosti med TF-IDF vektorji sporočil in končnimi odgovori, ter vsebinami knjig.

Ugotovila sva, da se povprečne podobnosti med razredi precej razlikujejo, iz česar sklepava, da obstaja dobra podlaga za izboljšanje uspešnosti razreda sporočil. Opazila sva tudi, da so si vrednosti kosinusne podobnosti s končnim odgovorom in vsebino knjige zelo podobne in so višje pri razredih povezanih z vsebino knjige. Povprečne podobnosti so prikazane v tabelah 5 in 6.

4.2 Delitev zbirke podatkov

Ker je zastopanost razredov v uporabljeni zbirki podatkov zelo neenakomerna in ker imajo nekateri razredi zelo malo pripadajočih sporočil, sva morala pri deljenju na učno in testno množico paziti, da je bilo v obeh množicah vsaj nekaj sporočil iz vsakega razreda. Pri tem sva si pomagala s stratificiranim vzorčenjem (angl. stratified sampling). 70 % sporočil sva uporabila za učno množico, preostalih 30 % pa za testno.

Table 4. Kdo je največ prispeval h končnemu odgovoru

Vprašanje	Pseudonim	Podobnost
1.1	dig-01	0,26076
1.2	pim-06	0,15764
1.3	dig-04	0,0
1.4	dig-08	0,08012
1.5	dig-10	0,25847
2.1	dig-13	0,13365
2.2	dig-14	0,40414
2.3	dig-16	0,15382
2.4	dig-19	0,0
2.5	dig-24	0,16821
3.1	pim-02	0,25516
3.2	blank	0,20895
3.3	pim-09	0,31015
3.4	pim-11	0,12908
3.5	pim-14	0,31100
4.1	pim-16	0,0
4.2	pim-20	0,05845
4.3	pim-23	0,44649
4.4	pim-25	0,05385
5.1	edf-01	0,15841
5.2	edf-08	0,34438
5.3	edf-05	0,27309
6.1	edf-10	0,07543
6.2	edf-12	0,22879
6.3	edf-15	0,10225

4.3 Izvedba eksperimentov

Za klasifikacijo na podlagi ročno pridobljenih značilk in TF-IDF vektorjev sva uporabila MLP. Tako za izračun TF-IDF vektorjev, kot za MLP sva uporabila funkcije iz knjižnice scikit-learn. Pri MLP-ju sva uporabila optimizator adam, nevronska mreža ima en sam skriti nivo z 20 nevroni. Parametre algoritma sva nastavila tako, da sva algoritem iterativno izboljševala s spreminjanjem posameznih paramatrov.

Za naprednejši metodi z uporabo BERT pa sva uporabila knjižnico transformers. Pri obeh metodah sva uporabila prednatreniran model in izvedla fino nastavljanje (angl. fine tune). Pri klasičnem BERT sva uporabila model “bert-base-uncased”, pri DistilBERT pa “distilbert-base-cased”.

Pri uporabi podobnosti besedil sva pri pristopu s TF-IDF vektorji uporabila scikit-learn, pri pristopu z BERT vektorskimi vložitvami pa knjižnico SentenceTranformer.

5. Rezultati

5.1 Ročno pridobljene značilke

S celotnim naborom ročnih značilk, ki je predstavljen v 3.1, sva pridobila rezultate, ki so prikazani v tabeli 7. V delu zbirke CREW so kar štirje razredi od katerih klasifikator ni pravilno razvrstil niti enega sporočila. To so razredi “content question”, “general comment”, “assignment instructions” in “opening statement”. Zelo slabo so klasificirana tudi sporočila

Table 5. Kosinusne podobnosti med sporočili iz zavihka CREW, končnimi odgovori in besedili knjig.

Razred	Končni odgovor	Knjiga
assignment instructions	0,10	0,15
assignment question	0,05	0,04
content discussion	0,20	0,19
content question	0,19	0,15
discussion wrap-up	0,05	0,08
emoticon/non-verbal	0,00	0,00
external material	0,11	0,10
feedback	0,08	0,07
general comment	0,08	0,09
greeting	0,01	0,01
incomplete/typo	0,00	0,00
logistics	0,08	0,10
opening statement	0,05	0,07
response	0,03	0,04

Table 6. Kosinusne podobnosti med sporočili iz zavihka Discussion in besedili knjig.

Razred	Knjiga
content discussion	0,31
content question	0,17
general comment	0,13
incomplete/typo	0,00
response	0,02

razreda “assignment question”. To da je klasifikacija sporočil obeh razredov z vprašanji slaba, ni bilo pričakovano, saj sva značilki o prisotnosti vprašaja in frekvenci besed, ki se začnejo na w ali h namenila prav za te razrede. Pri tem velja omeniti, da so bila tri sporočila razreda “content question” klasificirana kot “assignemnt question”, kar je verjetno posledica poenostavljenega prepoznavanja vprašanj. Značilka o prisotnosti vprašaja je vseeno vsaj malo koristna, saj v primeru, ko ni uporabljena ni pravilno razvrščeno nobeno sprocilo iz razredov vprašanj in makro F1 se zniža iz $0,392$ na $0,328$.

Sporočila razredov “general comment”, “assignment instructions” in “logistics”, ki imajo sicer relativno veliko vzorcev, so bila večinoma napačno razvrčena v razred “content discussion”, ki je daleč največji. To je do neke mere pričakovano, saj se glede na lastnosti, ki jih upoštevajo uporabljene značilke bistveno ne razlikujejo.

V delu zbirke Discussion only je le en razred, v katerem ni pravilno klasificirano niti eno sporočilo. Presenetljivo je ta razred ravno “content question”. Značilki, ki upoštevata prisotnost emoji-ev v tem delu zbirke poslabšata klasifikacijo. Prav tako klasifikacijo poslabša značilka o prisotnosti klicaja. Brez teh značilk se namreč vrednost makro mere F1 poveča na $0,746$ in pravilno je razvrščeno vsaj po eno sporočilo iz vsakega razreda.

Ugotovila sva, da korenjenje in lematizacija nimata pozitivnega učinka na uspešnost napovedovanja, enako velja tudi

Table 7. Rezultati evalvacije algoritma, ki deluje na podlagi celotnega nabora ročno pridobljenih značilk.

	CREW	Discussion
accuracy	0,533	0,838
macro precision	0,437	0,629
macro recall	0,395	0,619
macro F1	0,392	0,609

za odstranjevanje stop besed (angl. stop words). Razlog za to je morda, da so sporočila kratka in je zato pomembna vsaka informacija, ki jo lahko uporabimo, medtem, ko je pri daljših besedilih pomembneje prodobiti bistvo brez podrobnosti.

5.2 TF-IDF

Z uporabo TF-IDF vektorjev se je klasifikacija sporočil CREW v primerjavi z ročnimi značilkami izboljšala glede na vse uporabljene mere. Glede na konfuzijsko matriko se je najizraziteje izboljšala klasifikacija sporočil razreda “logistics”. Močno pa se je izboljšala še klasifikacija v razrede “incomplete/typo”, “content question” in “assignment instructions”.

Za nekatere razrede se je uspešnost klasifikacije nekoliko poslabšala. Primera takih razredov sta “emoticon/non-verbal”, ki sva mu namenila posebni značilki za prisotnost emotikonov in razred “external material”, ki je tudi imel namensko značilko prisotnost spletne povezave.

Za sporočila iz dela Discussion only, se je izboljšala le točnost, predvsem zaradi več pravilno razvrščenih sporočil večinskega razreda “content discussion”. Glede na ostale mere, se je uspešnost klasifikacije v tem delu zbirke poslabšala. Vsi rezultati pridobljeni z uporabo TF-IDF vektorjev so prikazani v tabeli 8.

Table 8. Rezultati evalvacije algoritma, ki uporablja TF-IDF vektorje.

	CREW	Discussion
accuracy	0,613	0,892
macro precision	0,516	0,531
macro recall	0,457	0,600
macro F1	0,446	0,561

Čeprav sva predvidela uporabo kombinacije unigramov in bigramov, sva ugotovila, da je delovanje boljše če se uporabijo le unigrami.

Zmanjševanje dolžine TF-IDF vektorjev, oziroma zmanjševanje nabora značilk se ni izkazalo za učinkovito, je pa manjše izboljšanje prineslo pravilo, da se mora beseda pojaviti vsaj v dveh sporočilih, da se doda kot značilka.

5.3 BERT

Vse uporabljene mere za evalvacijo klasifikacije sporočil CREW, so se v primerjavi z uporabo TF-IDF vektorjev dodatno izboljšale. Izboljšala se je tudi mera F1 za večino posameznih razredov. Ostal je samo še en razed iz katerega

ni bilo pravilno klasificirano niti eno sporočilo. To je razred "opening statement", sporočila tega razreda so bila klasificirana kot "response" ali "logistics".

Za razreda "assignment instructions" in "external material" je bila klasifikacija glede na mero F1 boljša z uporabo TF-IDF vektorjev. Za "external material" je bila celo še boljša ob uporabi ročnih značilk.

Pri sporočilih iz dela zbirke Discussion only, so se v primerjavi z uporabo TF-IDF vektorjev zvišale vrednosti vseh mer, razen makro priklica, ki se ni spremenil. Še vedno pa je bila klasifikacija z uporabo ročnih značilk na tem delu zbirke bolj uspešna. Vsi rezultati za oba dela zbirke so prikazani v tabeli 9.

Table 9. Rezultati evalvacije algoritma BERT

	CREW	Discussion
accuracy	0,698	0,889
macro precision	0,610	0,575
macro recall	0,521	0,600
macro F1	0,541	0,587

5.4 DistilBERT

Predvidevala sva, da bo model DistilBERT deloval bolje od klasičnega BERT, saj bi poenostavljena arhitektura lahko delovala bolje na uporabljeni zbirki, ki vsebuje relativno malo podatkov.

Za sporočila iz dela zbirke CREW se ta hipoteza ni izkazala za pravilno, saj je bila klasifikacija glede na vse mere slabša od klasičnega modela BERT. Vsi rezultati so prikazani v tabeli 10. Presenetljivo se je občutno poslabšala klasifikacija za razred "content discussion", ki ima daleč največ sporočil. Priklic za ta razred je bil boljši celo ob uporabi ročnih značilk, vrednost mere F1 pa se je v primerjavi s klasičnim modelom BERT znižala iz $0,873$ na $0,753$. Kar 17 sporočil tega razreda je bilo napačno razvrščenih v razred "general comment". Vrednost mere F1 se v primerjavi s klasičnim modelom BERT ni izboljšala za noben razred.

Tako kot s klasičnim modelom BERT, tudi ob uporabi modela DistilBERT ni bilo pravilno razvrščeno nobeno sporočilo razreda "opening statement", dodatno pa ni bilo pravilno razvrščeno še nobeno sporočilo razreda "assignment question".

Za razliko od dela zbirke CREW se je za del Discussion only uspešnost klasifikacije ob uporabi modela DistilBERT izboljšala. Vrednost mere makro F1 se je iz $0,587$ doseženih s klasičnim BERT zvišala na $0,723$. To je tudi edini naprednejši pristop s katerim sva dosegla boljši rezultat, kot ob uporabi ročnih značilk, vendar to velja le za uporabo celotnega nabora ročnih značilk. Z zmanjšanim naborom sva namreč dosegla vrednost mere makro F1 $0,746$.

5.5 Uporaba podobnosti

Vsi rezultati do sedaj so bili pridobljeni z običajnimi metodami klasifikacije besedil. V nadaljevanju so predstavljene še ugotovitve ob poskusu izboljšanja rezultatov z uporabo podobnosti sporočil s skupnim končnim odgovorom in vsebino knjige. Način uporabe podobnosti je predstavljen v razdelku 3.6.

Table 10. Rezultati evalvacije algoritma DistilBERT

	CREW	Discussion
accuracy	0,519	0,917
macro precision	0,516	0,781
macro recall	0,352	0,700
macro F1	0,389	0,723

Splošna ugotovitev je, da uporaba podobnosti na preizkušen način ne izboljša delovanja klasifikacije. To velja tako za podobnost sporočila z vsebino knjige, kot tudi za podobnost s končnim skupnim odogovorom.

Rezultati dveh izjem pri katerih se uspešnosti klasifikacije ob uporabi podobnosti ni poslabšala, so prikazani v tabelah 11 in 12. V 11 so rezultati, ki so bili pridobljeni brez in z uporabo podobnosti iz dela zbirke CREW, za primer ko je bil uporabjen MLP klasifikator na podlagi TF-IDF vektorjev. Podobnost je bila izračunana na podlagi BERT vektoskih vložitev. Vidimo lahko, da se je dejansko izboljašala le mera makro F1, točnost je ostala popolnoma enaka, makro preciznost in makro priklic sta se celo nekoliko znižala. To pomeni, da so bila sporočila le nekoliko drugače razporejena po razredih, število pravilno klasificiranih je ostalo enako. Uporaba podobnosti je pri razvrščanju nekaterih sporočil pomagala, pri drugih pa škodovala.

Table 11. Rezultati brez in z uporabo podobnosti sporočil CREW s končnim skupnim odgovorom za MLP klasifikator, ki deluje na podalgi TF-IDF vektorjev. Podobnost je bila izračunana na podlagi BERT vektorskih vložitev.

	Brez podobnosti	S podobnostmi
accuracy	0,613	0,613
macro precision	0,516	0,511
macro recall	0,457	0,454
macro F1	0,446	0,462

V tabeli 12 je prikazan primer, kjer je uporaba podobnosti dejansko izboljšala klasifikacijo glede na vse uporabljene mere. Gre za klasifikacijo spročil iz dela zbirke Discussion only z uporabo TF-IDF vektorjev in MLP klasifikatorja. V tem primeru je bila tudi podobnost izračunana na podlagi TF-IDF vektorjev.

5.6 Časovna zahtevnost

Ker bi se algoritem za klasifikacijo sporočil v aplikaciji IMapBook lahko uporabljal tudi za avtomatsko spremljanje pogovorov in obveščanje učitelja oziroma moderatorja, sva preverila tudi, kakšne so časovne zahtevnosti posameznih algoritmov.

Časovno zahtevnost sva preverila na testni množici dela zbirke CREW, torej na 212 sporočilih. Teste sva izvajala na

Table 12. Rezultati brez in z uporabo podobnosti sporočil Discussion only z vsebino knjige za MLP klasifikator, ki deluje na podalgi TF-IDF vektorjev. Tudi podobnost je bila izračunana na podlagi TF-IDF vektorjev.

	Brez podobnosti	S podobnostmi
accuracy	0,892	0,919
macro precision	0,531	0,781
macro recall	0,600	0,700
macro F1	0,561	0,723

prenosnem računalniku, za čas izvajanja pa sva upoštevala le čas potreben za izračun vektorjev in klasifikacijo. Branje iz Excel datoteke in nalaganje modela v primeru pristopov z modelom BERT sva izključila. Vsak algoritem sva zagnala trikrat in izračunala povprečje. Časovne zahtevnosti za klasifikacijo posameznega spročila so prikazane v tabeli 13.

Table 13. Časovne zahtevnosti testiranih algoritmov v milisekundah. Prikazan je čas za klasifikacijo enega sporočila

Ročne značilke	TF-IDF	BERT	DistilBERT
0,287	0,217	58,771	16,170

Najhitejša je klasifikacija zuporabo TF-IDF vektorjev in MLP. Razlog da je hitrejša od algoritma, ki uporablja ročne značilke, je verjetno v neoptimizirani kodi za izračun ročnih značilk.

Algoritma, ki uporabljata modele BERT sta veliko počasnejša, vendar še vedno popolnoma dovolj hitra, da bi omogočala sprotno klasifikacijo sporočil. Pričakovano je klasifikacija z modelom DistilBERT hitrejša od klasičnega modela BERT.

6. Zaključek

V današnjem času nevronskih mrež smo navajeni zbirk podatkov, ki imajo tudi po več tisoč vzorcev za vsak razred. Recept za uspešno klasifikacijo je v takšnih primerih pogosto znan že v naprej - uporaba kompleksne nevronske mreže, ki dobro izkoristi množico podatkov. Zbirke podatkov, kot je bila uporabljena v tej seminarski nalogi pa nas prisilijo, da se bolj posvetimo podatkom, ki so na voljo in da poskusimo ugotoviti, kako jih čim bolje uporabiti na podlagi specifičnih lastnosti.

V delu zbirke Discussion only, ki vsebuje manj sporočil, se je najbolje izkazal pristop z uporabo nabora ročno pridobljenih značilk, vrednost mere makro F1 je bila $0,749$. V delu zbirke CREW pa smo lahko videli moč modela BERT, ki je bil po zaslugi prednatreniranega modela in nakoliko večjega nabora sporočil sposoben klasificirati bolje od preprostejših pristopov in dosegel mero makro F1 $0,541$.

Rezultati za del zbirke Discussion only so precej boljši, kar je pričakovano, glede na to, da je bilo v tem delu zbirke le pet razredov (po odstranjevanju razredov z zelo malo primeri), v delu zbirke CREW pa kar 14.

V prihodnosti bi se bilo dobro posvetiti še uporabi sosledja sporočil. V okviru te seminarske naloge sva se posvetila klasifikaciji posameznih sporočil neodvisno od okolice, vendar je verjetno v spletnih pogovorih prisotno smiselno soslednje, ki bi ga z metodami strojenga učenja lahko izkoristili.

Če bi imeli na voljo sporočila iz več skupin in pogovore o več različnih knjigah, bi bilo morda bolje uporabiti drugačno delitev podatkov na učno in testno množico. Uporabili bi lahko tehniko, ko iz zbirke podatkov izvzememo en primer (angl. leave one subject out - LOSO). Tako bi za učno množico uporabili vsa sporočila razen tistih, ki pripadajo eni skupini ali tistih, ki se nanašajo na eno knjigo. Tako bi lahko pripravili klasifikator, ki je čim manj odvisen od podatkov o specifični knjigi.

Eden od naslednjih korakov bi bila sigurno tudi praktična uporaba klasifikacije sporočil za analizo pogovorov v aplikaciji IMapBook. En od možnih načinov uporabe bi bil pomoč učetelju pri spremljanju učencev. S pomočjo avtomatske klasifikacije bi lahko sproti zaznali, če bi učenci začeli govoriti o temah, ki s knjigo in vprašanjem niso povezane. Učitelj bi bil o tem lahko obveščen in tako ob primernem trenutku učence usmeril nazaj na pogovor o knjigi. Druga možnost uporabe pa bi bila analiza pogovora, ko bi bil ta že zaključen. Na ta način bi dobili povzetek oziroma analizo pogovora v katerem bi bili podatki o tem, kateri učenec je bil najbolj aktiven in čigava sporočila so bila najpogosteje povezana s knjigo, vprašanjem in končnim odgovorom.

References

[1] Ge Song, Yunming Ye, Xiaolin Du, Xiaohui Huang, and Shifu Bie. Short text classification: A survey. *Journal of multimedia*, 9(5):635, 2014.

[2] Soroush Vosoughi and Deb Roy. Tweet acts: A speech act classifier for twitter. *arXiv:1605.05156v1*, 2014.

[3] Marta R. Costa-juss'a, Esther Gonzalez, Asuncion Moreno, and Eudald Cumalat. Abusive language in spanish children and young teenager's conversations: datapreparation and short text classification with contextual word embeddings. *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, page 1533–1537, 2020.

[4] Justin D. Weisz. Segmentation and classif ication of online chats. *cs.cmu.edu*, 2006.

[5] Raymond Cheng. Bert text classification using pytorch, Jul 2020.

[6] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.