UNIVERZA V LJUBLJANI

Fakulteta za računalništvo in informatiko

Tjaša Pernuš

Ocenjevanje esejev s strojnim učenjem

DIPLOMSKO DELO NA UNIVERZITETNEM ŠTUDIJU

Mentor

Prof. dr. Igor Kononenko

Ljubljana, 2016

Izjava o intelektualni lastnini diplomskega dela

Tema diplomskega dela

Za ocenjevanje esejev eksperti uporabljajo objektivne, pa tudi subjektivne kriterije. Zaradi subjektivnosti ocen se tudi ocene ekspertov med seboj razlikujejo. Avtomatsko ocenjevanje esejev uporablja učno množico že ocenjenih esejev, kjer se esej opiše z množico atributov, na osnovi katerih zatem klasifikacijski in regresijski algoritmi zgradijo funkcijo za ocenjevanje esejev. V diplomski nalogi primerjajte uspešnost različnih algoritmov strojnega učenja za klasifikacijo in za regresijo, kot so naključni gozdovi, SVM, usmerjene večnivojske umetne nevronske mreže in več različic globokih nevronskih mrež, na različnih podatkovnih množicah esejev z znanimi ocenami ljudi strokovnjakov. Za vsak algoritem poiščite ustrezne vrednosti parametrov ter poskusite tudi odstraniti redundantne in irelevantne atribute. Pri tem poleg standardnih mer uspešnosti klasifikacije in regresije uporabite tudi specializirane mere za uspešnost ocenjevanja esejev, kot sta utežena kappa in sosednje ujemanje, ter preizkusite spreminjanje klasifikacijskih napovedi v regresijske in obratno. Za globoke nevronske mreže preizkusite poleg običajnega nabora atributov tudi različne predstavitve esejev, ki potencialno nudijo več možnosti za ekstrakcijo informacije o kvaliteti eseja (kot npr. vektorje TF-IDF).

IZJAVA O AVTORSTVU
diplomskega dela

Spodaj podpisani/-a TJAŠA PERNUŠ, ,

sem avtor/-ica diplomskega dela z naslovom:

OCENJEVANJE ESEJEV S STROJNIM UČENJEM

S svojim podpisom zagotavljam, da:

- sem diplomsko delo izdelal/-a samostojno pod mentorstvom (naziv, ime in priimek)

  Prof. dr. Igor Kononenko

  in somentorstvom (naziv, ime in priimek)

  _____

- so elektronska oblika diplomskega dela, naslov (slov., angl.), povzetek (slov., angl.)
- ter ključne besede (slov., angl.) identični s tiskano obliko diplomskega dela

- soglašam z javno objavo elektronske oblike diplomskega dela v zbirki »Dela FRI«.

V Ljubljani, dne _____ Podpis avtorja/-ice: _____

Zahvala

Zahvaljujem se mentorju prof. dr. Igorju Kononenku ter as. Kaji Zupanc za nasvete in pomoč pri izdelavi diplomske naloge.

Kazalo

Seznam uporabljenih kratic in simbolov

- Naključni gozdovi – RF
- Naivni Bayes – NB
- Nevronska mreža NN
- Globoka nevronska mreža DNN
- Metoda podpornih vektorjev – SVM
- Avtomatsko ocenjevanje esejev – AEE
- Odločitveno drevo – DT
- Srednja kvadratna napaka – MSE
- Relativna srednja kvadratna napaka – RMSE
- Kvadratna utežena Kappa – QWK
- Klasifikacijska točnost - EA

Povzetek

Diplomska naloga se dotika področja strojnega učenja, bolj podrobno pa še področja globokih nevronskih mrež. Cilj naše diplomske naloge je bila primerjava različnih pristopov strojnega učenja pri avtomatskem napovedovanju ocen esejev ter oceniti uspešnost globokih nevronskih mrež v primerjavi z ostalimi modeli. Pri gradnji globokih nevronskih mrež je bila izvedena tudi gradnja globoke nevronske mreže z osnovnimi eseji, ki so razbiti na n-terke, vsaka n-terka pa je predstavljala posamezni atribut. Za primerjavo je bilo uporabljeno okolje R, kjer je bilo izvedeno testiranje in primerjava modelov. Izdelanih je bilo več različnih modelov istega tipa, nato pa za posamezni tip izbran najbolj uspešen, ki je bil nato uporabljen v končni primerjavi različnih tipov modelov.

Ključne besede: ocenjevanje esejev, strojno učenje, globoke nevronske mreže, nevronska mreža, Friedman-Nemenyi test

Abstract

The diploma thesis covers the field of machine learning. In more detail it covers the field of deep neural networks. The goal of our thesis was comparing different approaches of machine learning for building models for automated essay scoring and to evaluate the success of deep neural networks compared to other models. For building models we have used already extracted attributes, but for the deep neural network we have also used original essays, represented by the three attributes, that represent the relationships in a sentence. For comparing we have used the R environment, where we have built, tested and compared the models. Many different models of the same kind were built, from which the best was chosen for further comparison with models of different types.

Key words: essay scoring, machine learning, Deep neural network, neural network, Friedman-Nemenyi test

1 Uvod

V diplomski nalogi smo se dotaknili področja avtomatskega ocenjevanja esejev z modeli strojnega učenja.

Glavna motivacija diplomske naloge je bila, da preučimo, koliko se lahko zanesemo na strojne modele, ki nam pomagajo pri ocenjevanju esejev, in ali je globoka nevronska s svojo kompleksnostjo dejansko zmožna boljšega ocenjevanja esejev kot bolj preprosti modeli, kot npr. naključni gozdovi, metoda podpornih vektorjev itd.

Diplomska naloga vsebuje primerjavo pristopov strojnega učenja. Predvsem nas je zanimalo delovanje globokih nevronskih mrež. Za izbiro najboljših modelov smo uporabili devet različnih množic podatkov z že pridobljenimi atributi. Pri končni primerjavi smo za gradnjo globokih nevronskih mrež na podlagi n-terk uporabili le 4 množice, saj smo za njih pridobili podatkovne množice.

Na kratko smo predstavili metodo naključnih gozdov (Random Forest – RF), metodo podpornih vektorjev (Support Vector Machine – SVM), metodo Naivnega Bayes-a (naive Bayes - NB). Nekaj več pozornosti pa smo namenili nevronskim mrežam (Neural networks – NN), predvsem globokim nevronskim mrežam (Deep Neural Networks – DNN).

Za uvod smo se dotaknili področja avtomatskega ocenjevanja esejev, nato smo na kratko predstavili teorijo modelov, ki smo jih uporabljali v našem diplomskem delu. V nadaljevanju smo predstavili rezultate našega testiranja ter zaključili s povzetkom testiranja in analizo rezultatov.

2 Pregled področja avtomatskega ocenjevanja esejev

Esej je kratko besedilo, napisano na določeno temo. Za določene raziskovalce esej predstavlja obliko, s pomočjo katere najlažje ocenijo, koliko se je posameznik naučil o določeni temi. Esej posamezniku omogoči, da prikaže, kakšno širino znanja je pridobil o določeni temi. Samo branje in ocenjevanje esejev pa je časovno potratno, prav tako se poraja vprašanje o zanesljivosti in objektivnosti ocene. Iz vseh naštetih razlogov se je razvila ideja o avtomatskem ocenjevanju esejev (Automated essey evaluation – AEE).

Razvoj se je začel v zgodnjih 60. letih, zadnja leta pa se je hitro razširil. Leta 1966 je angleški profesor E. Page prvi predlagal strojno ocenjevanje in dal pobudo za začetke razvoja na tem področju. Leta 1973 je imel na voljo dovolj strojne moči, da je lahko razvil prvi sistem z imenom Project Essay Grade [1]. Rezultati sistema so bili presenetljivo dobri, saj je bila primerjava med ocenjevalcem in strojem bolj uravnotežena kot primerjava med dvema ocenjevalcema. Kljub temu pa sistem ni naletel na odobravanje.

Do leta 1990, ko so se razvila orodja za obdelavo jezika, sistemi za e-učenje, je AEE dobil podporo v izobraževanju [1]. Danes je AEE uporabljen v kombinaciji z ocenjevalci v več testih. V literaturi najdemo, da se uporablja predvsem pri angleških testih, kot so Graduate Record Examination, Test of English as a Foreign Language…

Glavni cilj je razvoj sistema, ki dosledno sledi potrebam učiteljev in njihovih študentov. Poleg prihranka časa in denarja sistemi AEE doprinesejo k višji stopnji povratne informacije o eseju. Prav tako pa lahko točno razberemo zakaj in kako je določeno besedilo dobilo oceno. S pomočjo AEE lahko posameznik trenira in izboljšuje svoje spretnosti v pisanju esejev, saj za isti esej sistem vedno vrne enako oceno.

Kot pri vseh računalniških sistemih se tudi tu pojavi vprašanje z zanesljivosti, točnosti. Točnost se predstavlja kot podpora oceni, ki jo sistem vrne. Zanesljivost pa se ocenjuje kot ocena ocenjevalcev, koliko zaupajo sami podpori ocene. Pri zanesljivost se pojavi vprašanje o zanesljivosti ocene, saj se ocena poda na podlagi analize besedila in ne dejanskega razumevanja. Avtorji sistemov AEE so predstavili rezultate pri ocenjevanju esejev v primerjavi z ocenami strokovnjakov in dokazali, da sistemi ponudijo zanesljivo oceno. Attali [2] zato predlaga seznam korakov, s katerimi lahko ocenimo zanesljivost sistema AEE:

- preverljive atribute – določiti elemente besedila, ki jih AEE lahko meri,
- analiza vpliva na oceno eseja, če združimo več atributov,

- možnost združitve človeškega in strojnega ocenjevanja v sistem.

3 Pregled uporabljenih metod

3.1 Naključni gozdovi (Random Forests - RF)

Naključni gozdovi so metoda ansambelskega učenja pri klasifikaciji. Prednost RF je, da naj bi izboljšala napoved DT. Glavna ideja RF temelji na konceptu, da se zgradi zaporedje odločitvenih dreves z izborom podmnožice najboljših atributov. Velikost naključnih gozdov je ponavadi vsaj 100 odločitvenih dreves.

Primarno se na učni množici oceni najboljše atribute. Na množici najboljših atributov se nato za vsako vozlišče v posameznem drevesu naključno izbere relativno majhno število atributov in zgradi odločitveno drevo. Vsako odločitveno drevo pa je zgrajeno na celotni učni množici [3]. Klasifikacija primera se nato izvede po principu glasovanja. Vsako odločitveno drevo poda svoj glas, iz vseh glasov pa se nato pridobi verjetnostna porazdelitev razredov.

S to metodo ansambelski algoritem doseže napovedno točnost, primerljivo z najboljšimi algoritmi strojnega učenja. Slaba stran algoritma je, da je razlaga odločitve otežena, saj je množica dreves nepregledna in zato nerazumljiva za uporabnika [3].

Pri diplomi smo zgradili naključne gozdove s 100 in 200 drevesi. Za nadaljnjo primerjavo med modeli smo uporabili model s 100 drevesi, saj je test Friedman-Nemenyi pokazal, da razlike niso statistično značilne.

3.2 Metoda podpornih vektorjev (Support vector machine - SVM)

SVM je matematični postopek, ki prepoznava vzorce. SVM razdeli primere v razrede, tako da so primeri iz nasprotnih razredov ločeni z jasno mejo in da je ta čim širša. Tako lahko nov primer klasificiramo v razred glede na to, na katero stran meje pade.

Vsak učni primer predstavimo s pomočjo vektorja v vektorskem prostoru, ki je mnogo dimenzionalen v primerjavi z originalnim atributnim prostorom. To preslikavo vhodnega prostora v vektorski prostor SVM doseže z uporabo jedrnih funkcij. Ideja metode SVM je, da v tem prostoru poišče hiper-ravnino, ki ločuje primere iz različnih razredov. Lego hiper-ravnine določajo najbližji vektorji (učni primeri), ki jim rečemo podporni vektorji [4].

Prednosti SVM:

- uspešen v večdimenzionalnem prostoru,
- uspešen tudi v primerih, kjer je dimenzionalnost večja od števila primerov,
- ko je hiper-ravnina določena, je klasifikacija hitra.

Slabosti SVM:

- parametre klasifikacije je težko razložiti,
- v primeru razširitve dimenzionalnosti imamo lahko težave pri kasnejšem zmanjšanju,
- najbolj uspešni so v dvorazrednih problemih, pri večrazrednih primerih je treba imeti posamezen klasifikator za posamezni razred,
- nastavitev parametrov algoritma SVM je netrivialna naloga.

Pri diplomi smo zgradili dva modela SVM, pri čemer smo spreminjali vrsto jedrne funkcije. Uporabili smo radialno in polinomsko funkcijo. Za gradnjo modelov smo posamezno podatkovno množico razdelili na 10 podmnožic in naredili 10-kratno prečno preverjanje. Za nadaljnjo primerjavo med modeli smo nato uporabili model z uporabo radialne funkcije, saj je imel boljšo uspešnost, kar je potrdil tudi test Friedman-Nemenyi (primerjava radialne s polinomsko funkcijo je dosegla vrednost p <0,05).

3.3 Naivni Bayes (Naive Bayes - NB)

Naivna Bayesova metoda je ena najpogostejše uporabljenih metod strojnega učenja. Je preprost klasifikator, ki ima temelje na Bayesovem teoremu s predpostavko o pogojni neodvisnosti vrednosti različnih atributov pri danem razredu. Naivno verjame, da je določena vrednost atributa pri danem razredu popolnoma neodvisna od katerekoli vrednosti drugega atributa [5]. Uporabna je na diskretnih atributih, v primeru zveznih atributov je le-te treba predhodno diskretizirati.

Sam koncept klasifikacije temelji na matematičnem izračunu pogojnih verjetnosti za nominalne atribute.

Kljub svoji naivnosti je metoda uspešna tudi v primerih, kjer je prisotna rahla odvisnost atributov. V primerih močne odvisnosti atributov pa metoda naivni Bayes odpove.

3.4 Nevronske mreže (Neural networks - NN)

Nevronske mreže so metoda strojnega učenja, ki naj bi najbolj posnemale delovanje človeških možganov. Njena sestava je zato definirana kot množica med seboj povezanih nevronov s t.i. sinapsami. Nevronska mreža je sestavljena iz plasti vhodnih nevronov, ki sprejmejo vhodne informacije in jih pošljejo naprej po sinapsah sosednjim nevronom. Po mreži se pošiljajo dražljaji, vsota dražljajev (signalov) pa določi, ali se bo določen nevron aktiviral ali ne.

Pri uporabi nevronskih mrež govorimo o treh fazah: fazi učenja, fazi testiranja ter fazi uporabe. V fazi učenja nevronska mreža na podlagi učnih primerov določi uteži na povezavah. Uteži na povezavah med nevroni so tiste, ki določajo lastnost nevronske mreže. S spremembo uteži, se nevronska mreža spremeni in posledično tudi izhod iz nevronske mreže. Nevronska mreža je zgrajena iz vhodnih nevronov, izhodnih nevronov ter vmesnih plasti nevronov. Zaradi teh skritih plasti nevronov je razlaga nevronskih mrež in njihove odločitve otežena. Same nevronske mreže se med seboj razlikujejo po zgradbi, vrsti učenja ter vrsti signalov.

Pri naši diplomi smo zgradili več modelov nevronskih mrež, ki so se med seboj razlikovali po število nevronov. Za gradnjo nevronske mreže smo uporabili knjižnico *nnet*, in njeno implementacijo usmerjene nevronske mreže (Feed-Forward neural network). Zgradili smo mrežo z 1 skritim nivojem, na katerem pa smo spreminjali število nevronov (5, 10, 20, 40 in 80 nevronov na skritem nivoju). Za gradnjo modelov smo posamezno podatkovno množico razdelili na 10 podmnožic in naredili 10-kratno prečno preverjanje. Uspešnost modelov smo nato ocenili in za nadaljnjo primerjavo med modeli uporabili najuspešnejšega.

Na podlagi testa Friedman-Nemenyi smo prišli do zaključka, da je najboljši model z 20 nevroni, saj je $p<0,05$ napram 10 nevronom, modeli z večjim številom nevronov pa ne kažejo nobenih dodatnih izboljšav. Rezultati testa so v tabeli 1.

Tabela 1: Vrednosti p po Friedman-Nemenyi testu za NN z različnim številom nevronov

	NN5	**NN10**	**NN20**	**NN40**
NN5	-	-	-	-
NN10	0.224	-	-	-
NN20	1.8e-06	0.046	-	-
NN40	2.7e-06	0.074	1.000	-
NN80	8.0e-05	0.328	0.996	0.998

3.5 Globoke nevronske mreže (Deep neural networks - DNN)

Globoka nevronska mreža je usmerjena nevronska mreža, ki ima več kot eno skrito plast nevronov. Torej je sestavljena iz plasti vhodnih nevronov, plasti izhodnih nevronov, med vhodno in izhodno plastjo pa obstaja več kot ena skrita plast nevronov [6]. Mreža je usmerjen graf, kjer je vsak skrit nevron povezan z več skritimi nevroni na nižji plasti. Vsaka skrita plast je torej nelinearna kombinacija nižjih plasti. Z optimizacijo mreže vsaka skrita plast dobi optimalne uteži, kar predstavlja optimalno nelinearno kombinacijo spodnje plasti. Z vsako plastjo, ki ima manj nevronov kot prejšnja plast, se dimenzionalnost napovedi manjša glede na prejšnjo plast.

Zaradi večjega števila plasti se število korakov za izračun poveča na število skritih nivojev plus ena. Prav zaradi večjega števila plasti lahko s pomočjo globokih nevronskih mrež rešimo tudi zelo težke nelinearne probleme [3].

Globoke nevronske mreže so zanimive, saj se jih lahko uporablja tako pri nadzorovanem kot tudi pri nenadzorovanem učenju. Pri nadzorovanem učenju se poskuša napovedati vektor Y na podlagi matrike vhodnih vektorjev X [7]. Pri nenadzorovanem učenju pa se poskuša napovedati matriko X na podlagi iste matrike X na vhodu. S tem mreža pridobi informacije o podatkih brez pomoči končnega razreda, na katerega ima po navadi vpliv človek. Naučene informacije se hranijo kot uteži na sinapsah. Posledica nenadzorovanega učenja je, da ima mreža enako število vhodnih in izhodnih nevronov.

Kot začetne uteži pri učenju globoka nevronska mreža uporabi uteži, ki jih je pridobila pri nenadzorovanem učenju. Tako je učenje nadzorovano učenje in popravlja uteži nenadzorovanega učenja.

Pri diplomi smo kreirali več globokih nevronskih mrež. Za gradnjo modelov smo posamezno podatkovno množico razdelili na 10 podmnožic in naredili 10-kratno prečno preverjanje. Izmed vseh implementacij smo najuspešnejšo nato naprej zgradili s 3 – 7 nivoji in nato spreminjali število nevronov na nivojih s tem, da je bilo število nevronov na vseh plasteh enako (5, 10, 20, 40, 80). Najuspešnejši model smo nato primerjali z ostalimi modeli.

Pri DNN smo kot vhodne podatke uporabili atribute, ki smo jih dobili v podatkovni množici, kasneje pa smo kot vhodne podatke uporabili n-terke esejev ter dele n-terk. Več o sami implementaciji je razloženo v poglavju 5.

3.5.1 Večnivojski Perceptron (Multilayer Perceptron)

Eden od predstavnikov globokih nevronski mrež je t.i. večnivojski Perceptron, ki ga tudi uporablja knjižnica, ki smo jo uporabili pri končni gradnji DNN.

Večnivojski preceptron je sestavljen iz več skritih plasti nevronov, ki so polno povezani z naslednjo plastjo.

Za samo učenje globokih mrež je Rumelhart leta 1986 s sodelavci razvil t.i. posplošeno pravilo delta, imenovano tudi pravilo vzvratnega razširjanja napake, ki se ga uporablja pri Perceptronu **Error! Reference source not found.**.

To pravilo omogoča učenje poljubno nivojsko sestavljene mreže. Osnovni princip pravila je sestavljen iz več korakov. Na začetku je mreža utežena z naključnimi utežmi. S tem preprečimo, da bi vsi nevroni na nekem skritem nivoju dobili enako utež. Po učnem primeru, ki ga pošljemo skozi vhod na izhod, se primerja razlika med dejanskim in želenim izhodom. S pomočjo razlike se izračunajo popravki. Na podlagi popravkov se najprej popravijo uteži med zadnjim in predzadnjim nivojem, nato pa se izračunajo želene vrednosti na predzadnjem sloju. Računanje se nato izvaja rekurzivno, vse do vhodnih nevronov. Zaradi računanja odvoda napake tudi pri nevronih iz skritih nivojev je potrebno, da je izhodna funkcija zvezna in zvezno odvedljiva. Za ta primer, se uporablja sigmoidna funkcija:

$$f(X) = \frac{1}{1 + e^{-X}}$$

DNN z veliko skritimi sloji in veliko enotami na sloj je zelo fleksibilen model, ki pa ima veliko parametrov. Zaradi velikosti je tudi težko optimizirati DNN v primeru, da imamo naključno izbrane začetne uteži. Kljub temu je DNN zmožen modelirati kompleksne in nelinearne odvisnosti med vhodom in izhodom.

Posplošeno pravilo delta ima tudi slabosti:

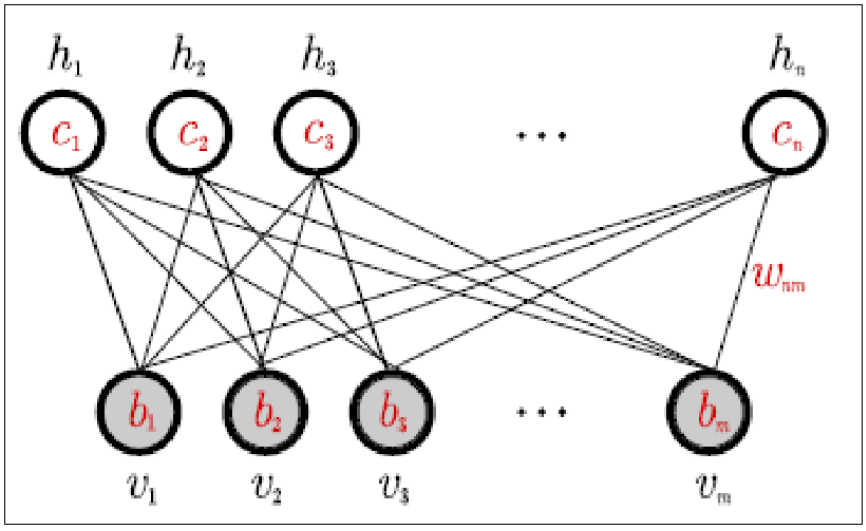
- ne konvergira vedno k optimalni rešitvi,

- empirično je treba določiti ustrezno število skritih slojev in število nevronov na vsakem sloju,
- zaradi prevelikega prilagajanja učni množici, je treba učenje ustaviti, ko začne napaka na neodvisni testni množici naraščati,
- potrebnih je veliko prehodov preko učnih primerov.

3.5.2 Boltzmannov model nevronske mreže (Restricted Boltzmann Machine – RBM)

Drugi predstavnik globokih nevronskih mrež so tudi t.i. RBM (Restricted Boltzmann Machine) nevronske mreže, ki temeljijo na Boltzamnnovem modelu nevronske mreže. [8]

RBM je sestavljen iz vidne in skrite plasti. Za razliko od usmerjene nevronske mreže so povezave med vidnimi in skritimi plastmi neusmerjene, vrednosti se lahko prenašajo naprej ali nazaj, hkrati pa je graf nevronske mreže polno povezan (slika 1). Vsak nevron na posamezni plasti je povezan z vsakim nevronom na naslednji plasti [9].



Slika 1 Neusmerjen graf Boltzmann-ove nevronske mreže [10].

RBM deluje v dveh fazah:

- pozitivna faza, ko se vhodni podatki *v* preslikajo na skrito plast, podobno kot pri usmerjeni mreži, kjer dobimo rezultat *h*,

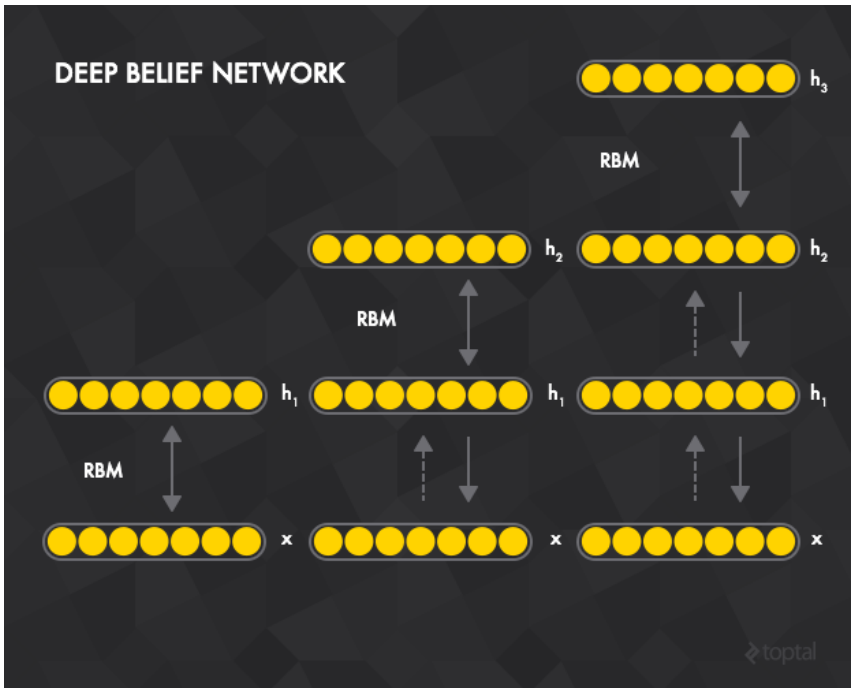
- negativna faza, ko se zaradi lastnosti RBM, da je neusmerjena mreža, h preslika nazaj na vidno plast, kjer dobimo rezultat v', ta pa se nato ponovno preslika na skrito plast in dobimo rezultat h',
- posodobitev uteži deluje po formuli:

$$w(t+1) = w(t) + a * \left(vh^T - v'^{h'^T}\right),$$

pri čemer je a hitrost učenja, v, v', h, h' in w pa so vektorji.

Pozitivna faza naj bi odražala notranjo predstavitev mreže na podlagi realnih podatkov, medtem ko negativna faza predstavlja poskus poustvariti vhodne podatke na podlagi notranje informacije. Cilj je, da se poskuša najbolj približati prvotnim vhodnim podatkom, kar pa se doseže s popravki uteži na povezavah.

RBM lahko zložimo tako, da tvori globoko nevronsko mrežo. V tem primeru skrita plast RBM t predstavlja vidno plast za RBM $t+1$ (slika 2).



Slika 2 RBM globoka nevronska mreža [10].

Vhodna plast prvega RBM je vhodna plast za celotno mrežo. Samo učenje pa deluje po naslednjih korakih:

- učenje prve RBM t=1 z uporabo kontrastne divergence (engl. contrastive divergence) z vsemi učnimi primeri,

- učenje druge RBM t=2, ki se začne z vhodom na vidno plast t=1, ki se nato preslika naprej v skrito plast t=1, kar pa nato sproži kontrastno divergenco za učenje t=2,
- postopek ponovimo za vse plasti.

Omenjeni model globoke nevronske mreže smo uporabili tudi pri naši gradnji modela globokih nevronskih mrež pri fazi, ko smo določali najuspešnejši model za končno primerjavo.

4 Opis podatkovnih množic

Uspešnost omenjenih pristopov smo preverili na področju avtomatskega ocenjevanja esejev. Kot podatkovno množico smo dobili devet množic podatkov s predhodno izbranimi atributi, ki opisujejo posamezni esej. V posamezni podatkovni množici je bilo predstavljenih od 723 do 1800 esejev z njegovimi atributi.

Ker ima vsaka podatkovna množica različen razpon končnega razreda, smo se odločili, da naredimo primerjavo individualno na vsaki množici posebej. Za izbiro najuspešnejšega modela znotraj enega pristopa smo uporabili celotni nabor atributov. Za končno primerjavo modelov pa smo atribute ocenili s pomočjo metode ReliefF in tako uporabili le prvih 9 najuspešnejših atributov (A1, A90, A3, GetissG, MoransI, A62, A63, A2, A82). Podrobnejši seznam in razlago atributov lahko najdemo v dodatku.

Pri implementaciji globokih nevronskih mrež smo poleg prejete podatkovne množice uporabili tudi originalne eseje. Za 4 podatkovne množice smo dobili originalne eseje in njihove n-terke. Na podlagi n-terk smo zgradili podatkovne množice ponavljajočih se n-terk. N-terke so sestavljene iz osebka, relacije in objekta. Zgradili smo podatkovno množico, kjer posamezna n-terka predstavlja posamezni atribut. N-terko smo kasneje razbili na tri atribute, kar nam je dalo novo podatkovno množico, s katero smo ponovno zgradili model.

Same n-terke iz esejev smo dobili že kot dodatno podatkovno množico. Pridobljene pa so bile s pomočjo aplikacije Open IE. Open IE je program, ki analizira povedi in vrne relacije v posamezni povedi. Samo poved razbije na 3 argumente (A, B in C), kjer C predstavlja relacijo med A in B [23].

Primer n-terke na enem od naših primerov:

Originalna poved:

A German dirigible was destroyed by fire.

N-terka:

A German dirigible; was destroyed; by fire

5 Implementacija

Celotni postopek testiranja smo izvedli v okolju R (version 3.2.2).

Med seboj smo primerjali gradnjo modelov na celotni množici in na okrnjeni množici atributov, saj s tem, ko izpustimo irelevantne atribute, lahko dosežemo večjo uspešnost modela.

Za gradnjo modelov smo posamezno podatkovno množico razdelili na 10 podmnožic in naredili 10-kratno prečno preverjanje. V vsaki iteraciji smo se iz 9 podmnožic učili in iz 1 podmnožice testirali. Rezultate točnosti smo nato povprečili čez vseh 10 iteracij. Tako smo dobili bolj zanesljivo oceno točnosti modela. Za razbitje množice na 10 podmnožic smo uporabili knjižnico cvTools [22].

Za končno primerjavo modelov smo uporabili samo najuspešnejše modele, ki smo jih nato primerjali med seboj. Spodaj so opisani modeli, ki so bili uporabljeni v končni primerjavi modelov.

Za gradnjo naključnih gozdov smo uporabili knjižnico randomForest [11], kjer je bilo uporabljenih 100 dreves. V vsakem vozlišču se naključno izbere $\sqrt{p}$ atributov (p je število atributov v učni množici), izmed katerih se nato izbere najboljši atribut. Samo vzorčenje primerov pa je narejeno z zamenjavo.

Za gradnjo modela podpornih vektorjev smo uporabili knjižnico e1071[18], pri čemer je bila uporabljena klasifikacija *c* ter radialna jedrna funkcija. SVM z uporabljeno klasifikacijo *c* pri učenju stremi k minimizaciji funkcije napake.

Za gradnjo modela Naivni Bayes smo uporabili knjižnico e1701 [18], kjer se predvideva Gaussova distribucija in neodvisnost atributov za napoved.

Za gradnjo nevronskih mrež smo uporabili knjižnico nnet [13]. Naša nevronska mreža ima 20 nevronov na skriti plasti. Število iteracij smo omejili na 1000, parameter decay pa smo nastavili na 0,1. Knjižnica omogoča tudi določitev uteži za posamezni primer, česar pa mi nismo uporabili. Za vsak primer je tako določena privzeta utež 1.

Za gradnjo globokih nevronskih mrež smo uporabili knjižnico neuralnet, ki ima implementiran model večnivojski perceptron [15]. Naša globoka nevronska mreža je sestavljena iz 7 skritih nivojev, pri čemer ima vsak skriti nivo po 40 nevronov. Število iteracij pa smo omejili na 500.

Pri gradnji globokih nevronskih mrež smo najprej s testiranjem modelu določili, koliko skritih nivojev je najbolj uspešno. S pomočjo testa Friedman-Nemenyi smo ugotovili, da je mreža s 7 skritimi nevroni najboljša, saj od vseh preostalih modelov kaže, da so razlike statistično značilne (tabela 2).

Tabela 2: Vrednosti p po Friedman-Nemenyi testu za DNN z različnim številom nivojev

	DNN_3_40	**DNN_4_40**	**DNN_5_40**	**DNN_6_40**
DNN_4_40	< 0.2e-15	-	-	-
DNN_5_40	< 0.2e-15	1.14289e-10	-	-
DNN_6_40	4e-14	0.02	5e-14	-
DNN_7_40	0.009	3.9e-14	< 0.2e-15	3.1057e-11

Nadalje smo testirali, koliko je optimalno število nevronov na skritih nivojih. Zgradili smo modele s 5, 10, 20, 40 in 80 skritimi nevroni na vsakem skritem nivoju. Vsak model smo zgradili z algoritmom *rprop+*, ponovitve učenja smo nastavili na 5, mejno vrednost odvoda funkcije napake smo nastavili na 0,01, število korakov pa smo omejili na 500. Izkazalo se je, da je model z 80 nevroni slabši od modela s 5 skritimi nevroni, od ostalih pa boljši. Model s 40 nevroni, pa je boljši od vseh predhodnih (tabela 3). Zato smo se odločili, da za nadaljnjo primerjavo modelov uporabimo model s 40 skritimi nevroni na plast.

Tabela 3: Vrednosti p po Friedman-Nemenyi testu za DNN s 7 nivoji in različnim številom nevronov

	DNN_7_5	**DNN_7_10**	**DNN_7_20**	**DNN_7_40**
DNN_7_10	0.882	-	-	-
DNN_7_20	0.930	1.000	-	-
DNN_7_40	0.48e-13	0.44e-13	0.56e-13	-
DNN_7_80	0.255	0.026	0.038	< 0.2e-15

Poleg naštetih knjižnic smo uporabili še knjižnice: Metrics [12] ter clusterSim [14].

Za ocenjevanje točnosti posameznih modelov smo uporabili več mer, in sicer kvadratna utežena Kappa (quadratic weighted kappa – Kw), natančno ujemanje (exact accuracy - EA), ki se najpogosteje uporabljata na področju avtomatskega ocenjevanja esejev. Ker smo gradili tudi regresijske modele, smo uporabili še srednjo kvadratno napako (mean squared error – MSE) in relativno srednjo kvadratno napako (relative mean square error – RMSE).

Same mere nam povedo, koliko točni so posamezni modeli. Seveda pa je treba preveriti, ali so razlike statistično značilne. Zato smo pri zaključkih uporabili mere za ocenjevanje več modelov na isti podatkovni množici, za kar smo uporabili Friedmanov test in Friedman-Nemenyi test.

5.1 Klasifikacijska točnost (exact accuracy - EA)

Klasifikacijsko točnost smo merili s primerjavo opazovane vrednosti in predvidene vrednosti razreda. Na podlagi vseh testnih primerov smo ocenili točnost. Če sta bili vrednosti enaki, smo vrnili vrednost 1, drugače 0. Vrednosti smo sešteli in delili s številom testnih primerov. Tako smo dobili točnost modela.

Mera ima določeno pomanjkljivost, saj primerja le točni razred, ne upošteva pa odmika. Če je predvidena vrednost 10 in je napovedana vrednost 10,5, je enako kaznovana kot, če dobimo napovedano vrednost 10,9. Hkrati pa se napovedani vrednosti 10,4 in 10,5 obravnavata povsem različno, kljub temu, da je njuna razlika majhna.

Zaradi svoje pomanjkljivosti mera tudi ni idealna za preverjanje kvalitete ocenjevanja esejev.

5.2 Kvadratna utežena Kappa (Quadratic weighted kappa - Kw)

Koeficient Kappa po Cohenu je mera zanesljivosti metode notranje konsistentnosti za nominalne merske lestvice. Definirana je kot:

$$\mathrm{K} = \frac{P_0 + P_t}{1 - P_t},$$

kjer so P_0 opazovane vrednosti skladnih parov in P_t teoretične vrednosti neskladnih parov. Leži na intervalu med [-1, 1] [19].

Težava koeficienta Kappa je, da vsa neujemanja obravnava enako. Izboljšana verzija koeficienta Kappa je kvadratna utežena Kappa *Kw*, ki se splošno uporablja kot mera zanesljivosti pri ocenjevanju esejev. Kw meri stopnjo ujemanja dveh ocenjevalcev (v primeru esejev sta to avtomatska podana ocena in ocena človeka). Pri izračunu se upošteva matrika opazovanih vrednosti, matrika predvidenih vrednosti ter matrika uteži. Matrika uteži po

diagonali predstavlja strinjanje in ima zato vrednosti enake 0. Vrednosti zunaj diagonale pa predstavljajo moč neujemanja med opazovano in predvideno vrednostjo.

Definirana je kot:

$$\mathrm{K} = 1 - \frac{\sum_{i=1}^{k}\sum_{J=1}^{k} w_{ij}x_{ij}}{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}m_{ij}},$$

kjer je k število razredov, w_{ij}, elementi v matriki uteži, x_{ij} element v matriki opazovanih vrednosti in m_{ij} element v matriki pričakovanih vrednosti [20]. Matrika uteži je izračunana na podlagi razlike med pričakovano in dobljeno vrednostjo razreda:

$$w_{i,j} = \frac{(i-j)^2}{(S-1)^2},$$

kjer S predstavlja število možnih klasifikacijskih razredov [21].

Matrika X je matrika reda S x S, kjer element x_{ij} predstavlja število esejev, ki so dobili oceno i po ocenjevalcu A in oceno j po ocenjevalcu B. Matrika M pa je matrika pričakovanih vrednosti reda S x S, kjer je posamezen element izračunan po naslednji formuli:

$$m_{ij} = \frac{H_{Ai} * H_{Bj}}{N},$$

kjer H_{Ai} , $i=1,\ldots,S$ predstavlja število esejev, ki jih je ocenjevalec A ocenil z oceno i, H_{Bj} , $i=1,\ldots,S$ število esejev, ki jih je ocenjevalec B ocenil z oceno j, N pa število ocenjenih esejev.

5.3 Srednja kvadratna napaka

Pri regresijskih problemih imamo opravka z zveznimi funkcijami, zato smo uporabili mero uspešnosti za avtomatsko zgrajene zvezne funkcije, t.i. *srednja kvadratna napaka*. Definirana je kot povprečni kvadrat razlike med napovedano vrednostjo $\hat{f}(i)$ in želeno vrednostjo $f(i)$:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(f(i) - \hat{f}(i)\right)^2.$$

Sama vrednost napake MSE je odvisna od dejanskega razpona možnih vrednosti funkcije.

5.4 Relativna srednja kvadratna napaka

Relativna srednja kvadratna napaka odpravlja pomanjkljivosti srednje kvadratne napake:

$$RMSE = \frac{N \text{ x } MSE}{\sum_i (f(i) - \bar{f}(i))^2},$$

kjer je $\bar{f} = \frac{1}{N}\sum_i f(i)$. RMSE je nenegativna in za sprejemljive hipoteze manjša od ena, vrednost RMSE = 1 lahko dosežemo s trivialno funkcijo, ki vedno vrne povprečno vrednost $\hat{f}(i) = \bar{f}$. Če je vrednost RMSE >1, je funkcija popolnoma neuporabna. Idealna funkcija je tista, pri kateri je RMSE=0.

5.5 Friedmanov test

Friedmanov test je neparametričen test za primerjavo razlik med dvema skupinama meritev. Test rangira uspešnost vsakega algoritma na vsaki domeni posebej. Friedmanov test dela na dveh predpostavkah, ki ju poskuša potrditi:

H_0 – Distribucija je enaka pri obeh meritvah

H_1 – Distribucija je različna glede na meritve

5.6 Friedmanov- Nemenyi test

Pri gradnji modelov je težko določiti, kateri model je najuspešnejši. Za primerjavo več modelov na isti podatkovni množici se pogosto uporablja Friedmanov-Nemenyi test, kjer primerjamo vsak algoritem z vsakim. Test potrdi značilnost razlike med dvema algoritmoma j_1 in j_2, če je razlika med povprečnima rangoma večja ali enaka kritični razliki CD:

$$\left|R_{j_1} - R_{j_2}\right| \geq CD = q_\alpha \sqrt{\frac{k(k+1)}{6D}},$$

pri čemer so q_α kritične vrednosti za test. Kritične vrednosti so predstavljene v dodatku.

6 Primerjava pristopov

Za končno primerjavo modelov smo uporabili le 4 podatkovne množice, saj smo za njih dobili tudi podatkovno množico n-terk. Modele smo zgradili na posamezni podatkovni množici posebej, rezultate točnosti pa smo povprečili čez vse podatkovne množice. Na podlagi povprečja in rezultatov primerjalnih testov smo nato naredili dodatne zaključke, kateri model se je bolje odrezal.

Skupaj smo imeli na voljo 7102 esejev (tabela 4). Izmed teh esejev jih je samo 5298, ki vsebujejo n-terko, ki se podvaja. Tako smo zgradili množico 5298 esejev, ki smo jih uporabljali pri gradnji modelov za končno primerjave pristopov. Modele smo zgradili na množici, ki je imela 126 atributov. Atributi so predstavljeni v dodatku.

Tabela 4: Seznam uporabljenih podatkovnih množic

Podatkovna množica	Št. esejev	Razpon razreda
3	1726	0-3
4	1771	0-3
5	1805	0-4
6	1800	0-4

Model, zgrajen na podlagi n-terk, je bil zgrajen na podatkovni množici, ki je vsebovala 2902 atributov (toliko različnih n-terk se je pojavilo v več kot enem eseju). Ker je n-terk veliko, lahko v dodatku najdete n-terke, ki se pojavljajo v več kot 20 esejih.

Kasneje smo n-terke tudi razbili na posamezne atribute (osebek, povedek predmet) in zgradili novo množico, kjer je vsak posamezen del n-terke nastopal kot atribut. Tako smo dobili množico z 2439 atributi. V dodatku lahko najdete atribute, ki jih najdemo več kot 20 krat.

6.1 Rezultati glede na točnost

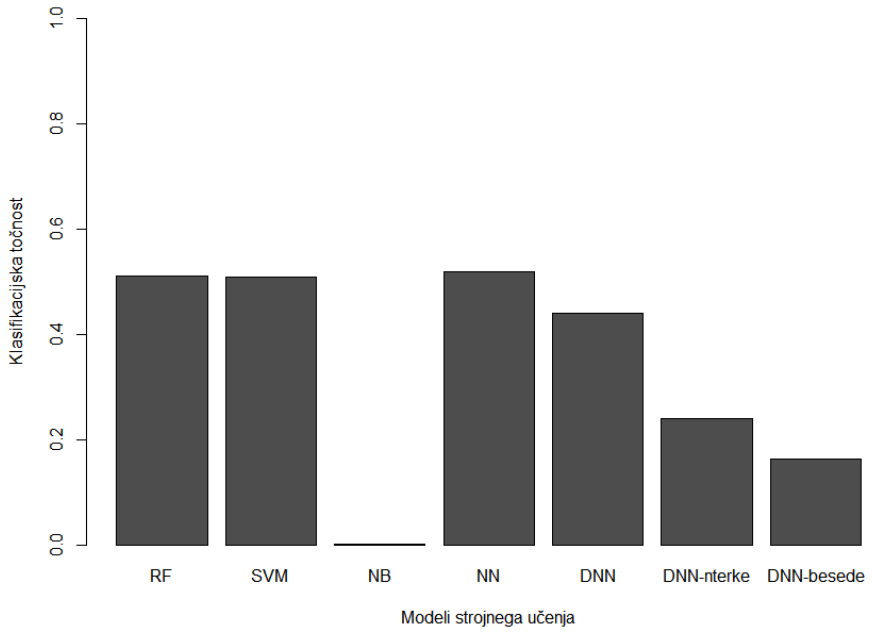
Pri vsakem modelu smo izračunavali 4 mere točnosti, in sicer klasifikacijsko točnost, kvadratno uteženo Kappo, srednjo kvadratno napako, ter relativno srednjo kvadratno napako. Ker sta

klasifikacijska točnost in kvadratna utežena Kappa mere točnosti za klasifikacijske probleme, smo pri tej meri naše regresijske napovedi zaokrožili k najbližji oceni.

Po klasifikacijski točnosti smo prišli do bolj slabih rezultatov, sledijo pa si po naslednjem vrstnem redu, od najboljšega k najslabšemu (slika 3): nevronska mreža, naključni gozdovi, metoda podpornih vektorjev, globoka nevronska mreža, globoka nevronska mreža na podlagi n-terk, globoka nevronska mreža na podlagi posameznih delov n-terke ter naivni bayes (tabela 5).

Tabela 5: Rezultati na podlagi klasifikacijske točnosti

RF	SVM	NB	NN	DNN	DNN-n-terke	DNN_besede
0,51038	0,50831	0,00170	0,51982	0,43922	0,23915	0,16365

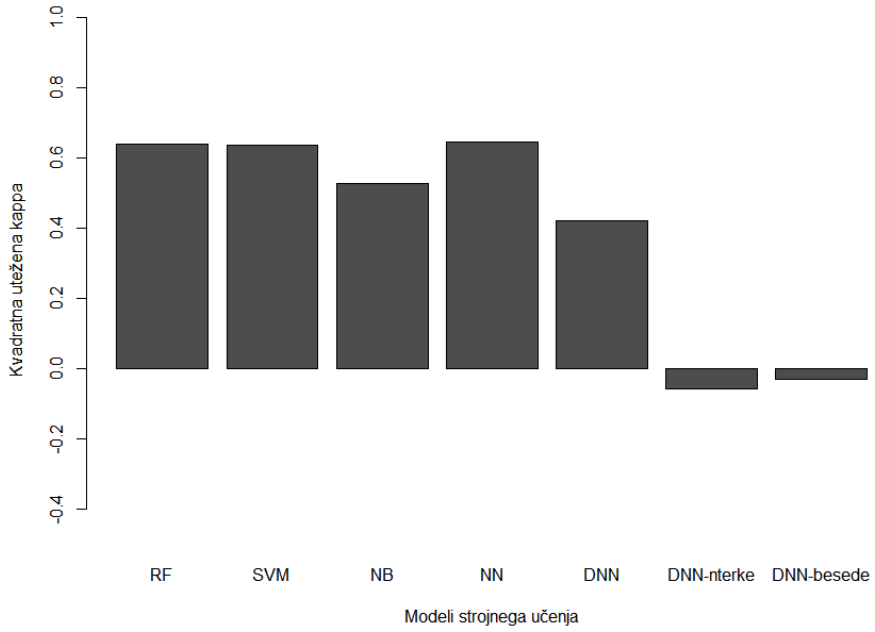


Slika 3: Grafični prikaz klasifikacijske točnosti

Pri meri kvadratna utežena kappa je prišlo do nekaj več razlik. Po rezultatih si sledijo po naslednjem vrstnem redu (slika 4): nevronska mreža, naključni gozdovi, metoda podpornih vektorjev, naivni bayes, globoka nevronska mreža, globoka nevronska mreža na podlagi posameznih delov n-terke, globoka nevronska mreža na podlagi n-terk (tabela 6).

Tabela 6:Rezultati na podlagi kvadrante utežene kappe

RF	SVM	NB	NN	DNN	DNN-n-terke	DNN_besede
0,64088	0,63524	0,52732	0,64445	0,42012	-0,05895	-0,03109

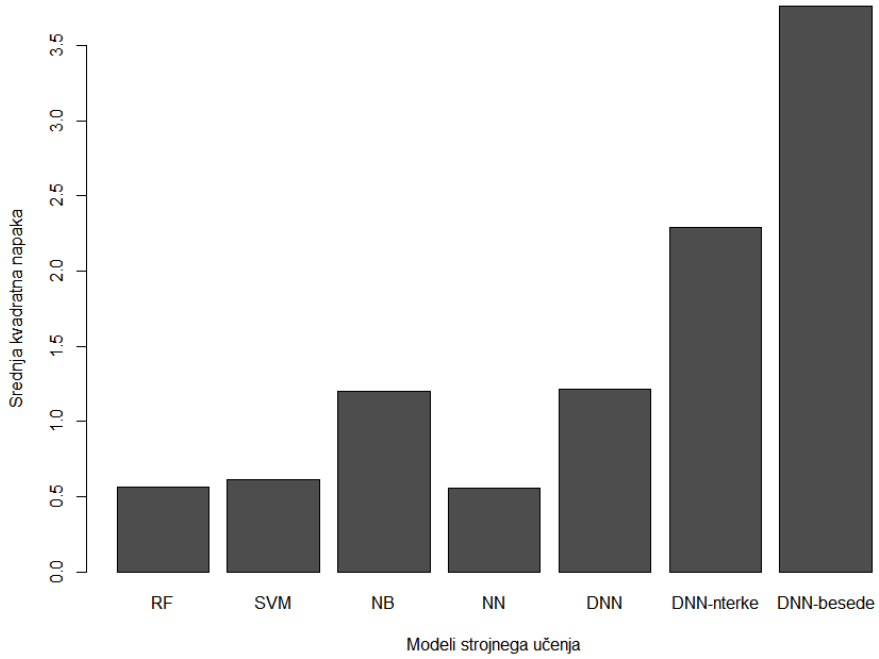


Slika 4:Grafični prikaz kvadratne utežene kappe

Pri meri srednja kvadratna napaka smo dobili razporeditev modelov po naslednjem vrstnem redu. Od najuspešnejšega do najmanj uspešnega si vrstijo(slika 5): naključni gozdovi, metoda podpornih vektorjev, nevronske mreže, naivni bayes, globoka nevronska mreža, globoka nevronska mreža na podlagi n-terk ter globoka nevronska mreža na podlagi posameznih delov n-terke (tabela 7).

Tabela 7: Rezultati na podlagi mere srednja kvadratna napaka

RF	SVM	NB	NN	DNN	DNN-n-terke	DNN_besede
0,56558	0,61497	1,1993	0,55423	1,21298	2,28907	3,76477

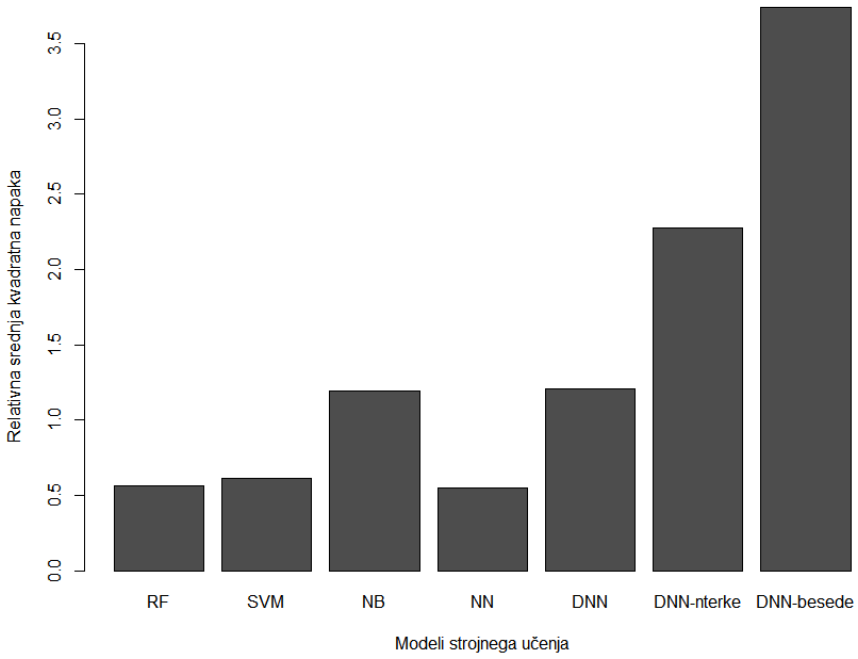


Slika 5: grafični prikaz srednje kvadratne napake

Pri relativni srednji kvadratni napaki smo ugotovili, da so le trije modeli dovolj uspešni za klasifikacijo novega primera. To so model, zgrajeni po metodi podpornih vektorjev, metodi naključnih gozdov in nevronske mreže. Ostali modeli imajo RMSE >1, kar pomeni, da so modeli bolj kot ne neuporabni (slika 6). Po vrstnem redu si ostali modeli sledijo: naivni bayes, globoka nevronska mreža, globoka nevronska mreža na podlagi n-terk ter globoka nevronska mreža na podlagi posameznih delov n-terke (tabela 8).

Tabela 8: Rezultati glede na mero relativna srednja kvadratna napaka

RF	SVM	NB	NN	DNN	DNN-n-terke	DNN_besede
0,56207	0,61115	1,19188	0,55079	1,20545	2,27486	3,74140



Slika 6:Grafični prikaz relativne srednje kvadratne napake

Rezultati MSE in RMSE so si zelo podobni, v kar smo se dodatno še poglobili. Po izračunu MSE za model, ki nam vedno vrne povprečno vrednost razreda (2,3), smo dobili vrednost 1,0062. Ker je RMSE definiran kot [MSE modela] / [MSE modela za povprečno vrednost], dobimo rezultate za RMSE skoraj enake kot za MSE.

6.2 Rezultati z uporabo Friedman-Nemenyi testa

Po rezultatih točnosti smo naredili še primerjavo modelov med seboj. S tem smo preverili, ali so razlike v izračunanih točnostih dejansko statistično reprezentativne, da lahko zaključimo, kateri model je najboljši. Za primerjavo več modelov na isti podatkovni množici smo uporabili Friedman-Nemenyi test.

Prvotno smo preverili Friedman-ov test, ki nam je vrnil vrednost $p < 0,05$. To pomeni, da so razlike dejansko statistično značilne.

Preverili smo, kako se modeli odrežejo še na Friedman-Nemenyi testu. Rezultati testa so predstavljeni v tabeli 9.

Tabela 9: Vrednosti p po Friedman-Nemenyi testu

	RF	**SVM**	**NB**	**NN**	**DNN**	**DNN-n-terke**
SVM	<0,2e-15	-	-	-	-	-
NB	<0,2e-15	<0,2e-15	-	-	-	-
NN	0,0021	<0,2e-15	<0,2e-15	-	-	-
DNN	<0,2e-15	0,992	<0,2e-15	<0,2e-15	-	-
DNN-n-terke	0,6671	<0,2e-15	<0,2e-15	0,2719	<0,2e-15	-
DNN-besede	<0,2e-15	<0,2e-15	<0,2e-15	<0,2e-15	<0,2e-15	<0,2e-15

Če sedaj uporabimo rezultate klasifikacijske točnosti v kombinaciji s Friedman-Nemenyi testom, lahko razberemo naslednje ugotovitve:

- primerjava DNN-nterke napram RF in NN, nam da vrednost p>0,05, kar pomeni, da razlike niso statistično signifikatne in ne moremo točno določiti, kateri model je boljši
- primerjava DNN napram SVM nam da vrednost p>0,05, kar pomeni, da razlike niso statistično signifikantne
- ostale primerjave modelov nam vrnejo s testom p<0,05, kar pomeni, da so razlike statistično signifikantne in lahko na podlagi točnosti določimo vrstni red modelov.

7 Sklepne ugotovitve

Za izbor najuspešnejših modelov znotraj posamezne skupine smo zgradili modele na celotni množici danih atributov ter s pomočjo testov določili izbor najboljših modelov. Tako smo prišli do pristopov, ki smo jih kasneje uporabili za končno primerjavo modelov.

Pred končno primerjavo modelov smo celotni atributni prostor ocenili z metodo ReliefF, s pomočjo katere smo izbrali najbolj pomembne atribute, ki vplivajo na določitev ocene eseja. Prav tako smo zgradili 2 novi podatkovni množici, ki imata atributni prostor bolj splošen. Prva množica ima kot atribute n-terke, ki so bile pridobljene s pomočjo Open IE. Drugo množico pa smo zgradili tako, da smo n-terke razbili na njene osnovne 3 dele in to uporabili kot atribute.

Po gradnji modelov smo uspešnost preverjali z več merami točnosti. V primeru, ko smo uporabili klasifikacijske točnosti (klasifikacijska točnost, utežena Kappa), smo pri regresijskih modelih razred zaokrožili k najbližjemu razredu. Pri obeh merah se je pokazalo, da je najboljša nevronska mreža, ki ji sledijo naključni gozdovi ter SVM in globoka nevronska mreža. Ostali trije modeli imajo točnost zelo nizko. Pri primerjavi SVM in globokih nevronskih mrežah nam je Friedman-Nemenyi test pokazal, da njune razlike v točnosti niso statistično značilne in zato ne moremo z gotovostjo reči, da je SVM boljši od DNN, kljub temu, da ima SVM po testih boljšo točnost.

Ker smo gradili regresijske modele, smo uporabili tudi mere točnosti, ki se uporabljajo v regresiji. Pri teh merah, smo dobili modele razvrščene po drugačnem vrstnem redu. Pokazalo se je, da je SVM najboljši, sledi RF, NN, NB ter na zadnjem mestu vse različice DNN. Kljub vsemu se je izkazalo, da so le trije zgrajeni modeli uporabni za klasifikacijo novih problemov.

Pri gradnji modelov smo poskušali spreminjati določene parametre, na podlagi katerih bi dobili boljše rezultate. Kljub vsemu pa se je izkazalo, da so bolj enostavni modeli (RF, SVM) vseeno boljši od kompleksnejših (DNN). Pri diplomski nalogi smo zgradili tudi globoko nevronsko mrežo na podlagi n-terk ter delov izločenih iz n-terk. Sama gradnja take DNN je bila časovno zahtevna, saj je bilo atributov veliko, rezultati DNN pa niso prinesli želenih ciljev. DNN model se je sicer uspešno zgradil, vendar pa je po vseh merah točnosti pristal na zadnjem mestu. Atributi, kot so n-terke esejev, nosijo premalo informacij, na podlagi katerih bi lahko določili oceno. Na podlagi atributov, kot so dolžina eseja, število različnih besed v eseju ipd., lahko strojni model lažje določi oceno eseja. Vsebinska struktura eseja se je izkazala za modele irelevantna, le-ta pa je lahko pri ocenjevalcih zelo pomembna.

Literatura

[1] Zupanc, K., Bosnić, Z. (2015), Advances in the Field of Automated Essay Evaluation, University of Ljubljana, Faculty of Computer and Information Science, Ljubljana, Slovenia

[2] ——, "Validity and Reliability of Automated Essay Scoring," in Handbook of Automated Essay Evaluation: Current Applications and New Directions, M. D. Shermis and J. C. Burstein, Eds. New York: Routledge, 2013, ch. 11, pp. 181–198.

[3] Kononenko, I., Kukar, M. (2007) Machine learning and data mining: Introduction to Principles and Algorithms. Horwood Publishing Chichester, UK

[4] W. S. Noble. What is a support vector machine? Nature biotechnology, 24(12):1565–1567, 2006.

[5] Naive Bayesian. Dostopno na: http://www.saedsayad.com/naive_bayesian.htm, Datum dostopa: Julij, 2016

[6] Hinton, G., Deng, L., Dong, Y., Dahl, G., Mohamed, A., , Navdeep, J., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., Kingsbury, B. (2012) Deep Neural Networks for Acoustic Modeling in Speech Recognition, Dostopno na: http://static.googleusercontent.com/media/research.google.com/sl//pubs/archive/38131.pdf, Datum dostopa: Julij, 2016

[7] Hall, P., How does deep learning work and how is it different from normal neural networks and/or SVM?, Dostopno na: http://www.quora.com/How-does-deep-learning-work-and-how-is-it-different-from-normal-neural-networks-and-or-SVM#, Datum dostopa: Junij, 2016

[8] Deep Belief Networks, Dostopno na: http://deeplearning.net/tutorial/DBN.html, Datum dostopa: maj, 2016

[9] Relić, I. (2015), Predstavljanje slika ograničenim Boltzmannovim neuronskim mrežama, Dostopno na: http://www.zemris.fer.hr/~ssegvic/project/pubs/relic15bs.pdf, Datum dostopa: maj, 2016

[10] Vasilev, I., A Deep Learning Tutorial: From Perceptrons to Deep Networks, Dostopno na: http://www.toptal.com/machine-learning/an-introduction-to-deep-learning-from-perceptrons-to-deep-networks, Datum dostopa: maj, 2016

[11] Liaw, A., Wienet, M. (2015). Package 'randomForest', Dostopno na: https://cran.r-project.org/web/packages/randomForest/randomForest.pdf, Datum dostopa: februar, 2016

[12] Hamner, B. (2012). Package 'Metrics',
Dostopno na: https://cran.r-project.org/web/packages/Metrics/Metrics.pdf, Datum dostopa: februar, 2016

[13] Ripley, B., Venables, W. (2016), Package 'nnet',
Dostopno na: https://cran.r-project.org/web/packages/nnet/nnet.pdf, Datum dostopa: februar, 2016

[14] Walesiak, M., Dudek, A. (2015) Package 'clusterSim', Dostopno na:https://cran.r-project.org/web/packages/clusterSim/clusterSim.pdf, Datum dostopa: februar, 2016

[15] Rong, X. (2016), Package 'deepnet',
Dostopno na: https://cran.r-project.org/web/packages/deepnet/deepnet.pdf, Datum dostopa: februar, 2016

[16] Therneau, T., Atkinson, B., Ripley, B. (2015), Package 'rpart',
Dostopno na: https://cran.r-project.org/web/packages/rpart/rpart.pdf, Datum dostopa: februar, 2016

[17] Robnik-Sikonja, M. & Savicky, P. (2014). Package ‘Corelearn’, Dostopno na: http://cran.rproject.org/web/packages/CORElearn/CORElearn.pdf, Datum dostopa: februar, 2016

[18] Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C., Lin, C. (2015), Package 'e1071', Dostopno na: https://cran.r-project.org/web/packages/e1071/e1071.pdf, Datum dostopa: februar, 2016

[19] Zemljič, B. (2010), Preučevanje kakovosti merjenja popolnih omrežij, Fakulteta za družbene vede, Ljubljana, Slovenija

[20] Fleiss, J. L., Cohen J. (1973), The equivalence of weighted kappa and the intraclass correlation coefficient as measure of reliability, Educational and Psychological Measurement, 33, Strani: 613-619

[21] Quadratic weighted Kappa, Dostopno na: https://www.kaggle.com/c/asap-aes/details/evaluation, Datum dostopa: februar, 2016

[22] Alfons, A. (2015), Package 'cvTools', Dostopno na: https://cran.r-project.org/web/packages/cvTools/cvTools.pdf, Datum dostopa: februar, 2016

[23] Open IE. Dostopno na: https://github.com/knowitall/openie, Datum dostopa: julij, 2016

Dodatek

Rezultati primerjave

Rezultati primerjave modelov SVM

	Radialna jedrna funkcija	Polinomska jedrna funkcija
EA	0,493717	0,434852
QWK	0,691727	0,571902
MSE	3,516407	4,45478
RMSE	0,448548	0,696538

Rezultati primerjave modelov NN z različnim številom nevronov

	5	10	20	40	80
EA	0,47043	0,47057	0,47078	0,47133	0,47180
QWK	0,65034	0,65122	0,65217	0,65333	0,65365
MSE	210,98990	211,06990	211,13530	211,16940	211,19100
RMSE	0,46778	0,46767	0,4674	0,46728	0,46718

Kritične vrednosti Friedmanov- Nemenyi test

Kritične vrednost za q_α pri več klasifikatorjih, pri čemer je α stopnja zaupanja.

	število klasifikatorjev								
α	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
0,05	1,960	2,343	2,569	2,728	2,850	2,949	3,031	3,102	3,164
0,010	1,645	2,052	2,291	2,459	2,589	2,693	2,780	2,855	2,920

Opis atributov

ID	ID eseja
A1	Number of char's

A2	Number of words
A3	Number of tokens
A4	Lexical diversity
A5	Number of long words
A6	Number of short words

A7	Most frequent word length
A8	Number of sentences
A9	Number of long sentences
A10	Number of short sentences
A11	Mode (modus) – word length
A12	Number of words divided by the number of sentences
A13	Number of characters divided by the number of sentences
A14	Number of ADV: adverb
A15	Number of CC: conjunction, coordinating
A16	Number of CD: numeral, cardinal
A17	Number of DT: determiner
A18	Number of EX: existential there
A20	Number of IN: preposition or conjunction, subordinating
A21	Number of JJ: adjective or numeral, ordinal
A22	Number of JJR: adjective, comparative
A23	Number of JJS: adjective, superlative
A24	Number of J: adjective
A26	Number of MD: modal auxiliary
A27	Number of NN: noun, common, singular or mass
A28	Number of NNP: noun, proper, singular
A29	Number of NNPS: noun, proper, plural
A30	Number of NNS: noun, common, plural
A31	Number of N: noun
A32	Number of P: prepositions
A33	Number of PAR: participle
A34	Number of PDT: pre-determiner
A35	Number of POS: genitive marker
A36	Number of POS: genitive marker
A37	Number of PRO: pronoun
A38	Number of PRP: pronoun, personal
A39	Number of PRP$: pronoun, possessive
A40	Number of RB: adverb
A41	Number of RBS: adverb, superlative
A42	Number of RP: particle
A44	Number of TO: "to" as preposition or infinitive marker
A46	Number of VB: verb, base form
A47	Number of VBD: verb, past tense
A48	Number of VBG: verb, present participle or gerund
A49	Number of VBN: verb, past participle
A50	Number of VBP: verb, present tense, not 3rd person singular
A51	Number of VBZ: verb, present tense, 3rd person singular
A52	Number of V: verb
A53	Number of WDT: WH-determiner
A54	Number of WP: WH-pronoun
A56	Number of WRB: Wh-adverb
A57	Number of spellchecking errors
A58	Number of different stemms
A59	Number of different lemmas
A60	Number of non-initial CAPS words
A61	Number of non-initial CAPS words
A62	Square root of number of words
A63	Fourth root of number of words
A64	Number of stopwords
A65	Stenford parser – height of the tree
A67	Word spell checker
A68	Word grammar checker
A69	Word variation index (48) - OVIX
A70	Nominal ratio (48) -NR
A71	Cosinusna korelacija z izvirnim besedilom
A72	Cosinusna korelacija z izvirnim besedilom

A73	Gunning Fog Index
A74	Flesch Reading Ease
A75	Flesch Kincaide Grade Level
A76	Dale-Chall readability formula
A77	Automated readability index
A78	SMOG
A79	Similarity – Text2Onto
A80	TF-IDF correlation
A81	Cosine correlation
A82	#words^5
A83	povprečna dolžina besed
A84	povprečna dolžina stavkov
A85	število različnih besednih zvrsti
A86	LIX
A90	Number of different words
A91	Type Token Ratio
A92	Guiraud's Index
A93	Yule's K
A94	Happax Legomena
A95	Score point value for which the max cosine corelation was obtained
A96	cosine corelation value (to best)
A97	the pattern cosine
A98	weighted sum of all cosine correlation values
distX	Povprečna razdalja med sosednimi točkami (X=2,3,5) v sematičnem prostoru TF-IDF, ki predstavljajo ne prekrivajoče dele eseja.
windowAvg14	Povprečna razdalja med sosednimi točkami v sematičnem prostoru TF-IDF, ki predstavljajo prekrivajoče dele eseja.
windowMin14	Najkrajša razdalja med sosednima točkama v sematičnem prostoru TF-IDF, ki predstavljajo prekrivajoče dele eseja.
windowMax14	Najdaljša razdalja med sosednima točkama v sematičnem prostoru TF-IDF, ki predstavljajo prekrivajoče dele eseja.
windowIndex14	Koeficient med najkrajšo in najdaljšo povezavo med sosednima točkama v sematičnem prostoru TF-IDF, ki predstavljajo prekrivajoče dele eseja.
diameter14	Najdaljša razdalja med katerimakoli točkama v sematičnem prostoru TF-IDF, ki predstavljajo prekrivajoče dele eseja.
AvgDistance14	Povprečna razdalja med vsemi točkami v sematičnem prostoru TF-IDF, ki predstavljajo prekrivajoče dele eseja.
dispAvg14	Povprečna oddaljenost točk od centroida
dispMin14	Najmanjša oddaljenost točke od centroida. = medoid
dispMax14	Največja oddaljenost točke od centroida
dispIndex14	Koeficient med najmanjšo in največjo oddaljenostjo točke od centroida.
SD14	Standard distance is the spatial equivalent of standard deviation
RD14	Relative Distance
MoransI	measure of spatial autocorrelation
GearysC	measure of spatial autocorrelation
GetissG	measure of spatial autocorrelation
DNN	Clark's measure of spatial relationship (distance to nearest neighbour)
NN	Distance to nearest neighbour
Gfun	G-function – Nearest neighbour analysis
det	Determinant of distance matrix
class	Ocean eseja

Uporabljene N-Terke (>20 ponovitev)

airships; flying too low; L:over urban areas
hydrogen; is; highly flammable
The mood; created; by the author in the memoir
the snows; melt;
they; come; back
The features of the setting; affect; the cyclist
they; did;
a ship; to tie up; T:ever
nothing; to do; with being a blood relative
I; will take; that test; T:again; T:then
The author; concludes; the story with this paragraph
we; considered; family
I; will be; grateful to my parents; T:always
The mood; created; by the author
they; had;
the country; loved; they
he; had;
The winds on top of the building; were shifting; due to violent air currents; T:constantly
a ship; to even approach; the area; T:ever
They; had;
I; will never forget; how my parents turned this simple house into a home
many immigrants; do;
endless celebrations; encompassed; both
Saeng; vowed silently;
family; had; nothing to do with being a blood relative
the hibiscus; is budding;
the mood; created; by the author
the back of the ship; would swivel; around
accident; could have been;
I; am; eternally grateful
the many things; learned; L:there; about how to love
I; adore; to this day
airships; flying;
I; think;
this hibiscus; is budding;
a much more courageous thing; could have done; T:ever
The builders; had;
he; has;
he; to die;
it; says;
They; came; selflessly
he; is;
The features of the setting; affected; the cyclist
The winds on top of the building; were constantly shifting; due to violent air currents
the builders; had;
The builders of the Empire State Building; faced; many obstacles; T:in attempting to allow dirigibles to dock there

The mood; created; by the author in this memoir
They; to give; their children; a better life
the mood; created; by the author in the memoir
the setting; affected; the cyclist
Most dirigibles from outside of the United States; used; hydrogen rather than helium
The greatest obstacle to the successful use of the mooring mast; was; nature
how important family and friends; are;
the author; created; L:in the memoir
a love of cooking; is; T:still; with me
he; says;
she; says;
a love of cooking; is; T:still; with me; T:today
many obstacles; faced; in attempting to allow dirigibles to dock there
the Empire State Building; was an existing law against; airships
the building; held; by a single cable tether
New York City; was the lack of; a suitable landing area
The stress of the dirigible's load and the wind pressure; to be transmitted; all the way to the building's foundation
the building's foundation; was; nearly eleven hundred feet below
the snow; melt;
Another obstacle; faced; the builders
I; 've told; them; that what they did was a much more courageous thing; T:often
I; will take; the test; T:again; T:then
My parents; kept; their arms and their door open to the many people; T:always
I; 've thanked; them; T:repeatedly
it; was constructed; T:ever
the author; concludes; the story with this paragraph
the innocence of childhood; formed; the backdrop to life; L:in our warm home; L:Here
I; to die;
spring; comes;
He; had;
The steel frame of the Empire State Building; to strengthened; to accommodate this new situation
dirigibles; could not moor; L:at the Empire State Building
the geese return and this hibiscus; is budding;
I; will never forget; that house or its gracious neighborhood or the many things
airships; from flying too low; L:over urban areas
people; not necessarily were clearly; in need
they; were; able to get back on their feet
Another obstacle; was; nature
Dirigibles; moored; L:in open landing fields

Uporabljene besedne zveze iz N-terk (>20 ponovitev)

he
I
they
is
was
had

created
she
it
The mood
faced
airships

the cyclist
hydrogen
L:over urban areas
flying too low
highly flammable
T:again

The author
concludes
they
says
melt
by the author in the memoir
the author
did
T:ever
will take
The features of the setting
the test
Saeng
back
affect
T:then
the snows
They
due to violent air currents
come
by the author
is budding
T:always
to do
the mood
family
nothing
used
to tie up
the builders
The winds on top of the building
you
affected
that test
came
the story with this paragraph
we
it
a ship
Another obstacle
nature
will never forget
do
airships
will be
considered
loved
the hibiscus
many obstacles
were shifting
T:constantly
dirigibles
He
with being a blood relative
would swivel
T:in attempting to allow dirigibles to dock there
were constantly shifting
the country
grateful to my parents
the area
hydrogen rather than helium
to even approach
the back of the ship
there
held
were
L:in the memoir
This
It
the setting
vowed silently

are
a better life
family
many immigrants
moored
L:there
the cyclist
The builders of the Empire State Building
to die
his parents
by a single cable tether
failed
to dock
has
them
T:still
for him
learned
the builders
around
would have
both
There
My parents
to be modified
could have been
with me
him
am
how my parents turned this simple house into a home
endless celebrations
encompassed
think
the story
to this day
to be transmitted
a love of cooking
was an existing law against
The obstacles
this
was destroyed
nothing to do with being a blood relative
everything
the building
the snow
to give
many obstacles
the obstacles
filled
affects
eternally grateful
accident
comes
"I
kept
The builders
their children
could have done
the many things
adore
a much more courageous thing
flying
about how to love
the air
dehydrated
Hydrogen
made
in attempting to allow dirigibles to dock there

all the way to the building's foundation
him
have
would be
to take
L:at the top of the building
me
That
people
this hibiscus
know
said
love
selflessly
the builders of the Empire State Building
She
went
've told
T:before
The cyclist
would add
was replaced
was the lack of
to strengthened
to be
would be dangling
T:often
stress
T:today
by the author in this memoir
have done
the German dirigible Hindenburg
Another problem
the Empire State Building
had thought
the dirigible
the features of the setting
Most dirigibles from outside of the United States
spring
nearly eleven hundred feet below
creates
very flammable
One obstacle
the story with that paragraph
L:at the Empire State Building
was no one in
sight
would make
of as strange
grew up
to get
The stress of the dirigible's load and the wind pressure
This law
left
safety
New York City
formed
the backdrop to life
he
water
the innocence of childhood
L:Here
in office
hot
states
The greatest obstacle to the successful use of the mooring mast
how important family and friends

the dirigibles
I will always be grateful to my parents for their love and sacrifice
T:now
L:in our warm home
could not moor
The setting
The steel frame of the Empire State Building
all
the geese
by short, rolling hills
to try
to accommodate this new situation
extended
a hand
to people
T:the spring
by fire in Lakehurst
Dirigibles
T:repeatedly
their arms and their door open to the many people
in need
home
a suitable landing area
to allow dirigibles to dock there
happy
from flying too low
together
the building's foundation
the winds on top of the building
L:in open landing fields
get
loves
that what they did was a much more courageous thing
starts

got
highly flamable
talks
by Joe Kurmaskie
ended
the author
the setting
it illegal for a ship to ever tie up to the building or even approach the area
L:in
the snow melts and the geese return
does
've thanked
would pick
mixing with the aromas of the kitchen
needed
to beat down
to give their children a better life
a dirigible
was riding
was constructed
to get back
L:above pedestrians on the street
A thousand-foot dirigible
could drop
mean
that house or its gracious neighborhood or the many things
safe
Most dirigibles from outside the United States
lived
concluded
started
The mood created by the author in the memoir

in
can see
by fire
no water
thought
clean
illegal
on
could not simply drop
shows
L:in this memoir
something
the geese return and this hibiscus
T:in June
to pass
to have
to say
carter
not necessarily were clearly
the spring
able to get back on their feet
L:the Empire State Building
was a law against
Flat road
was running
was traveling
extremely flammable
Narciso
takes
The architects
were many obstacles that
all of these cultures
to drink
budding
"The sun
"There