Zaznavanje sentimenta v novicah z globokimi nevronskimi mrežami

Andraž Pelicon

Odsek za tehnologije znanja
Institut "Jožef Stefan"
Jamova cesta 39, 1000 Ljubljana
Andraz.Pelicon@ijs.si

Povzetek

Področje analize sentimenta v novicah se v zadnjem času vse pogosteje uporablja predvsem v okviru napovedovanja gibanja finančnih trgov, vendar je za slovenski jezik še dokaj slabo raziskano. V okviru te raziskave smo zasnovali arhitekturo na osnovi nevronskih mrež, ki za klasifikacijo uporablja kombinacijo samodejno generiranih značilk s pomočjo slojev s povratno zanko in TF-IDF obtežitev. Modeli, ki uporabljajo omenjeno arhitekturo, dosegajo primerljive rezultate z že obstoječimi modeli in so sposobni učinkovitega učenja na korpusih v velikosti okrog 10.000 dokumentov.

1. Uvod

Mnenja in subjektivni odnos ali sentiment do določenega subjekta so ljudem zelo pomembni pri vsakodnevnem odločanju in sodobne tehnologije omogočajo hitro in učinkovito izmenjavo mnenj kjerkoli po svetu. Na družbenih platformah kot sta Twitter in Facebook ljudje neprestano izražajo mnenja o različnih izdelkih ter svoj odnos do raznih tem in oseb, od politikov do podjetnikov. Spletne trgovine omogočajo, da kupci o določenem izdelku oddajo svoje mnenje in oceno, kar pomaga tako drugim potencialnim kupcem kot trgovini sami, da lahko lažje oceni priljubljenost prodajanih izdelkov. Novičarske platforme omogočajo objavo komentarjev, kjer bralci izražajo svoja mnenja glede prebranih novic in subjektivne poglede na obravnavane teme. Prek vseh teh kanalov se tako proizvedejo ogromne količine prosto dostopnih besedil, ki jih je praktično nemogoče ročno pregledati. Zato se je že precej zgodaj pojavilo zanimanje za izdelavo modelov in sistemov, ki bi s pomočjo tehnik strojnega učenja bili sposobni samodejno analizirati mnenja v krajših in daljših oblikah besedil.

Najbolj zgodnji modeli, ki so se ukvarjali z analizo mnenj, so se osredotočali predvsem na samodejno prepoznavanje mnenj o posameznih izdelkih v ocenah in komentarjih uporabnikov, kjer je mnenje uporabnika dokaj jasno izraženo v besedilu in kjer mnenje poleg besedila izraža tudi številčna ali znakovna ocena. Z eksplozijo družbenih omrežij je za posamezna podjetja postalo zanimivo sledenje priljubljenosti svoje znamke v javnih objavah uporabnikov na spletu. Kmalu pa se je zanimanje začelo usmerjati tudi na vrste besedil, ki naj bi načeloma nevtralno podajale informacije in ne bi izražale nobenega mnenja avtorja ali njegovega subjektivnega pogleda na obravnavano tematiko, kot so na primer novice. Analiza sentimenta v novicah se pogosto uporablja na novicah iz finančne domene, kjer številni raziskovalci poskušajo na podlagi analiziranega sentimenta napovedati gibanja na finančnih trgih. Van de Kauter et al. (2015) navajajo, da imajo novice neposreden vpliv na finančne trge, saj dobre novice načeloma spodbujajo, slabe novice pa omejujejo rast na trgu.

Večinoma se analiza mnenj in sentimenta zastavi kot klasifikacijski problem, kjer se posamezna besedila klasificira v eno od predhodno definiranih kategorij. Najbolj razširjene tehnike strojnega učenja na tem področju so metode podpornih vektorjev, vendar se v zadnjem času tudi na tem področju pojavlja vse več modelov globokih nevronskih mrež, ki dosegajo primerljive rezultate.

V zvezi s količino podatkov, ki je potrebna za učenje sposobnega modela, ni enotnega mnenja, saj je slednje precej odvisno od samega problema in velikosti modela. Kljub temu empirični rezultati iz posameznih študij kažejo, da se lahko na podlagi večje količine podatkov modeli strojnega učenja, predvsem modeli globokega učenja, naučijo bolj informativnih značilk, ki pomagajo pri boljši klasifikaciji. Yang et al. (2016) so razvili arhitekturo globokih nevronskih mrež za klasifikacijo daljših dokumentov. Poročali so o izboljšanju rezultatov na vseh šestih preskusnih zbirkah podatkov, pri čemer sta dve manjši zbirki podatkov vsebovali več kot 300.000 primerov, tri večje podatkovne zbirke so vsebovale več kot milijon primerov, največja pa več kot tri milijone primerov. Devlin et al. (2018) so na podlagi variante nevronskih mrež, tako imenovane pozornosti (ang. attention), razvili velik jezikovni model, imenovan BERT. V članku poročajo, da je bil model naučen na dveh sicer neoznačenih korpusih, BookCorpus, ki vsebuje 11.038 knjig v celoti in 800 milijonov besed, in korpusu člankov angleške Wikipedije, ki vsebuje 2.500 milijonov besed. Pridobivanje tako velikih količin podatkov, predvsem označenih, je pogosto zelo drag in zamuden postopek, zato je učenje uspešnih klasifikacijskih modelov na omejenih virih aktualen problem.

Raziskovalni prispevki našega dela so naslednji. Razvili smo nov sistem za zaznavanje senitmenta v novicah, ki temelji na arhitekturi globokih nevronskih mrež in pokazali, da dosega primerljive rezultate z dosedanjimi modeli, ki temeljijo na tradicionalnih metodah strojenga učenja. Nadalje smo pokazali, da lahko naš nevronski model učinkovito učimo tudi na manjših podatkovnih množicah reda do 10.000 primerov.

2. Namen članka

Čeprav se analiza sentimenta v novicah vse pogosteje izvaja za večje jezike, je to področje za slovenščino zaenkrat še precej neraziskano. Edini modeli, ki po naših informacijah obstajajo, so modeli, naučeni na podlagi metode podpornih vektorjev. Cilj te raziskave je, da bi razvili model globokih nevronskih mrež, ki bi uspešno napovedoval sentiment v novicah, kot ga dojema povprečen slovenski bralec. Pri tem si kot prvo hipotezo postavljamo trditev, da je mogoče z modeli nevronskih mrež doseči primerljivo klasifikacijsko točnost kot z drugimi tehnikami strojnega učenja.

Za naš problem trenutno še ni na voljo velike količine kakovostno označenih podatkov. Tako bomo poskušali naše modele naučiti na korpusu, ki vsebuje okrog 10.000 primerov novic. Naša druga hipoteza je, da je tudi s tako omejeno količino podatkov mogoče izdelati model globokih nevronskih mrež, ki uspešno modelira dani problem.

3. Pregled raziskav

Analiza sentimenta, v literaturi imenovana tudi analiza mnenj, je večdisciplinarno področje na preseku računalništva in jezikoslovja, ki spada v širše področje obdelave naravnega jezika (NLP). Ukvarja se z analizo in prepoznavanjem odnosa, čustev in mnenj ljudi do določenega objekta. Zajema skupek tehnik iz računalniške obdelave naravnega jezika, s katerimi lahko iz besedila izvlečemo subjektivne informacije (Beigi et al., 2016).

Tradicionalno se sentiment v besedilih razpoznava v odnosu do objekta sentimenta, tj. predmeta, pojma ali osebe, na katero se izraženi sentiment nanaša (Mejova, 2009). Primeri objektov sentimenta v tvitih so izdelki določenega podjetja, v ocenah filmov pa filmi.

Čeprav je tako razmerje med sentimentom v jezikovni strukturi in objektom sentimenta precej razširjeno, pa se nekateri avtorji problema lotijo z druge perspektive. Namesto vprašanja, kakšno mnenje ima avtor do nekega objekta v besedilu, se sprašujejo, kakšen sentiment zaznavajo bralci v določenem besedilu. Primer takega razumevanja vloge sentimenta najdemo v Lin et al. (2008). Naša raziskava se osredotoča na slednjo perspektivo sentimenta, in sicer na nivoju dokumenta.

Na področju analize sentimenta v novicah avtorji pogosto kombinirajo pristope z leksikoni sentimenta in algoritmi strojnega učenja. Taj et al. (2019) so sestavili svoj korpus iz novic, objavljenih na portalu BBC, ter s pomočjo TF-IDF obtežitev v vsaki novici poiskali ključne besede. Nato so izključno na podlagi podatkov o sentimentu ključnih besed iz leksikona sentimenta poskušali analizirati sentiment v posameznih novicah. V okviru tehnik strojnega učenja se za učenje modelov najpogosteje uporablja metoda podpornih vektorjev. Kaur in Kaur (2015) sta s pomočjo metode podpornih vektorjev analizirala sentiment v novicah, napisanih v jeziku pandžabi, pri čemer sta besedila obdelala s standardnimi metodami predobdelave, in sicer tokenizacijo, odstranjevanjem neinformativnih besed in krnjenjem. Pogosto se v kombinaciji s strojnim učenjem uporabljajo tudi leksikoni sentimenta za podajanje dodatnih informacij o sentimentu v model. Lin et al. (2008) so analizirali sentiment v kitajskih novicah, kjer so uporabili metodo podpornih vektorjev, pri čemer so kot vhod uporabili več vrst značilk, ki so jih pridobili s kombinacijo predobdelave besedil v korpusu in informacij iz leksikona sentimenta. Kot značilke so uporabili bigrame pismenk, posamezne besede, metapodatke o posamezni novici, pripone besed in sentiment posameznih besed, ki so ga pridobili v leksikonu sentimenta. Li et al. (2014) so podobno metodo uporabili za preverjanje vpliva novic na donosnost delnic. Naloga modelov je bila predvideti gibanje cen delnic, pri čemer je bila učna množica sestavljena iz novic, kot napovedni cilj pa so uporabili historične cene delnic iz istega petletnega obdobja, v katerem so izšle novice iz učnega korpusa. Da bi v model vnesli podatke o sentimentu, so s pomočjo leksikona sentimenta sestavili matriko sentimenta za celotno besedišče. Nato so matriko sentimenta uporabili za preslikavo TF-IDF matrike, ki so jo sestavili na podlagi pogostosti izrazov v korpusu novic, v nov vektorski prostor. Dobljeno matriko so nato uporabili za učenje modela s pomočjo metode podpornih vektorjev. Kumar et al. (2017) so novicam poskušali določiti sentiment kot zvezno vrednost v intervalu med -1 in 1, pri čemer je vrednost -1 predstavljala najbolj negativen sentiment, vrednost 1 pa najbolj pozitiven sentiment. Tudi oni so za učenje uporabili metodo podpornih vektorjev s to razliko, da so problem opredelili kot regresijski problem. Za napoved sentimenta so kot značilke uporabili uni- in bigrame besed, obtežene s TF-IDF utežmi, značilke, izdelane na podlagi informacij iz leksikona sentimenta, in 300-dimenzijske vektorske vložitve besed word2vec.

V zadnjem času se tudi na področju analize sentimenta vedno bolj pojavljajo modeli globokega učenja, ki temeljijo na nevronskih mrežah. Mansar et al. (2017) so za analizo sentimenta uporabili konvolucijske nevronske mreže, ki se sicer pogosteje uporabljajo v robotskem vidu. S pomočjo konvolucijskih plasti so pridobili notranje predstavitve posamezne novice iz učnega korpusa in jih nato združili z oceno sentimenta posamezne novice, ki so jo pridobili s preprostim modelom VADER, ki temelji na pravilih. Tako pridobljene značilke so nato uporabili kot vhod v končni klasifikator, ki je temeljil na polnopovezani nevronski mreži. Razen tokenizacije se metod predobdelave besedil niso posluževali. Njihov model je bil leta 2017 najboljši v eni od nalog na letnem mednarodnem tekmovanju o računalniški semantični analizi SemEval. Moore in Rayson (2017) sta za namene analize sentimenta iz naslovov novic izdelala dva modela, enega na podlagi metode podpornih vektorjev z ročno generiranimi značilkami, drugega pa na podlagi dvosmernih nevronskih mrež s povratno zanko z dolgo-kratkoročno pomnilno celico (ang. bidirectional recurrent neural networks with long-short term memory ali BILSTM). V svojem eksperimentu poročata, da je njihov nevronski model dosegel za 4–6 odstotkov višjo točnost v primerjavi z metodo podpornih vektorjev.

Kljub temu, da številne od navedenih raziskav uporabljajo predhodno naučene besedne vektorske vložitve za predstavitev vhodnih besedil, Yu et al. (2017) navajajo, da slednje načeloma ne zajamejo dovolj podatkov o sentimentu, saj nosijo samo informacijo o kontekstu posamezne besede. Tako imamo lahko besede, ki se uporabljajo v zelo podobnih kontekstih, izražajo pa povsem nasproten

sentiment. Vzemimo na primer protipomenki *dober* in *slab*. Besedi izražata povsem nasproten sentiment, vendar ker sta obe pridevnika in se pojavljata v zelo podobnih kontekstih, imata zelo podobno vektorsko vložitev. Vnašanja informacije o sentimentu v vektorske vložitve so se lotili tako, da so vzeli javno dostopne predhodno naučene vektorske vložitve in jih na podlagi informacij iz leksikona sentimenta prilagodili. Pri tem so uporabili leksikon E-ANEW, v katerem je sentiment kodiran z realnim številom na intervalu od 1 do 9, kjer 1 predstavlja najbolj negativen, 5 najbolj nevtralen in 9 najbolj pozitiven sentiment. Za vsako besedo v korpusu so poiskali njenih deset najbližjih sosedov glede na kosinusno razdaljo njihovih vektorskih vložitev. Nato so besede v tej množici razvrstili po sentimentu, in sicer od bolj do manj podobne glede na absolutno razliko v sentimentu iz leksikona sentimenta med besedo iz množice in ciljno besedo. Ko je bil seznam besed sestavljen, so vektorsko vložitev ciljne besede posodobili tako, da je bila bliže vektorskim vložitvam na vrhu seznama, ki so ji bile po sentimentu podobne, in dlje od vektorskih vložitev besedam na dnu seznama, ki so se po sentimentu od ciljne besede razlikovale. Da bi preprečili preveliko spremembo izvirnega vektorskega prostora in posledično izgubo kontekstualnih informacij, so dodali omejitev, ki je določala, koliko se lahko posodobljeni vektor ciljne besede največ razlikuje od izvirnega vektorja. Metodo posodabljanja vektorjev so preizkusili na dveh predhodno naučenih javno dostopnih vektorskih vložitvah Word2vec in GloVe. Posodobljene in izvirne vektorske vložitve so nato uporabili na problemih binarne in petrazredne klasifikacije sentimenta, pri čemer so modele učili na podatkih iz korpusa SST (Socher et al., 2013). V primerjavi rezultatov avtorji poročajo o 1,5-odstotni izboljšavi za vektorje GloVe in 1,7-odstotni izboljšavi za vektorje Word2vec na problemu binarne klasifikacije ter o 1,5-odstotni izboljšavi za oboje vektorje na problemu petrazredne klasifikacije.

Področje analize sentimenta v novicah v slovenskem jeziku je zaenkrat še precej neraziskano. Bučar et al. (2018) so na korpusu novic, ki so ga sami sestavili in tudi javno objavili, naučili več modelov na podlagi metod naivnega Bayesa, podpornih vektorjev in k najbližjih sosedov, pri čemer so kot značilke uporabljali n-grame besed s TF-IDF obtežitvami. Najbolj učinkovite modele, ki so bili naučeni za razlikovanje sentimenta glede na tri kategorije, negativno, nevtralno in pozitivno, smo v tej raziskavi uporabili kot osnovo za primerjavo naših naučenih modelov.

4. Metodologija dela

V tem poglavju predstavimo metodološki pristop, ki smo ga uporabili za izvedbo eksperimenta. Najprej opišemo podatkovno množico, ki smo jo uporabili za učenje modelov in njihovo evalvacijo. Nato opišemo arhitekturo nevronskih mrež, ki smo jo uporabili za naše modele, ter postopek učenja vseh petih modelov analize sentimenta v novicah. Poglavje zaključimo z opisom postopka validacije naučenih modelov.

4.1. Opis podatkov

Za namene razvoja modelov za zaznavanje sentimenta v slovenskih novicah smo uporabili korpus slovenskih novic SentiNews (Bučar et al., 2018), ki je sestavljen iz 10.427 novic z ekonomsko, finančno in politično vsebino. Novice so bile zbrane na petih najbolj branih slovenskih spletnih novinarskih portalih, in sicer 24ur, Dnevnik, Finance, Rtvslo in Žurnal24, v obdobju med 1. septembrom 2007 in 31. decembrom 2013. Novice je šest označevalcev ročno označilo na treh stopnjah granularnosti, in sicer na nivoju besed, povedi in celotnega dokumenta. Med označevanjem so morali označevalci sentiment v novicah označiti s stališča povprečnega slovenskega bralca, pri čemer so odgovorili na vprašanje, kakšne občutke so doživljali po branju novice. Za učenje modelov smo uporabili končne oznake, ki posamezno novico v korpusu označijo kot negativno, nevtralno ali pozitivno. Označeni del korpusa tako vsebuje 3337 negativnih, 5425 nevtralnih in 1665 novic, kar kaže na dokaj drastično neuravnoteženost kategorij. Ker so mere strinjanja med označevalci najvišje pri oznakah na ravni dokumenta, smo se odločili, da bomo za namene nadzorovanega učenja modela uporabili oznake na tej ravni.

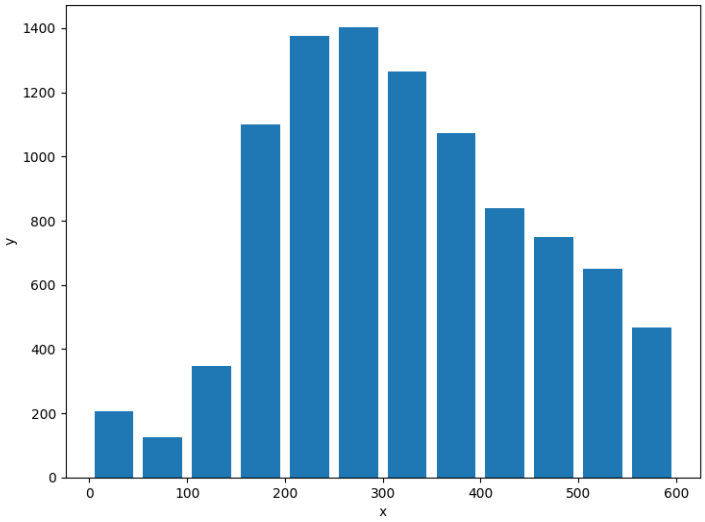
Po podrobnejšem pregledu podatkov smo ugotovili, da so določeni primeri v korpusu izjemno kratki. Načeloma pri teh primerih večinoma ne gre za celotno novico, temveč samo za podpise pod slikami, ki so sestavljeni iz ene do treh povedi. Takih priemrov je sicer malo, vendar se po sami zgradbi in informativnosti močno razlikujejo od ostalih primerov. Da v modele ne bi vnašali nepotrebnega šuma, smo se odločili, da bomo zato iz korpusa odstranili vse primere, ki so sestavljeni iz manj kot petih povedi. Takih primerov je bilo v korpusu 307, tako da je naš končni korpus vseboval 10.120 označenih primerov.

4.2. Zasnova eksperimenta

Za modeliranje sentimenta novic smo naučili klasifikacijski model na osnovi nevronskih mrež. Model kot vhod tako prejme besedilo kot celoto in ga v nadaljevanju obdela v dveh obdelovalnih cevovodih. V prvem obdelovalnem cevovodu smo vsako novico najprej tokenizirali. Nato smo vsako besedo, ki se pojavi v korpusu, preslikali v ustrezno vektorsko vložitev. V našem primeru smo uporabili predhodno naučene 300-dimenzionalne vektorske vložitve Fasttext (Grave et al., 2018). Tako smo dobili matriko velikosti R^{dxm}, kjer je d število besed v novici, m pa dimenzija vektorske vložitve.

Novice v korpusu so različne dolžine, za zagotavljanje računsko učinkovitih matričnih operacij med učenjem modela pa moramo nevronski mreži s povratno zanko posredovati vhodne podatke enake dolžine. Zato smo posamezne novice po potrebi skrajšali ali podaljšali. Posamezno novico smo podaljšali tako, da smo konec matrike zapolnili z ničelnimi vektorji, ki ne predstavljajo besed. Dolžino novic smo glede na porazdelitev dolžin novic v korpusu, predstavljeni na Sliki 1, določili na 500 besed.

Predobdelane novice smo tako poslali v sloj s povratno zanko z LSTM celicami s 120-dimenzionalnimi skritimi plastmi. V vsakem koraku obdelave sloj s povratno zanko prejme kot vhod vektorsko vložitev posameznega elementa vhodne sekvence in jo preslika v 120-dimenzionalno notranjo vektorsko predstavitev. Za namene regularizacije smo na sloju LSTM uporabili izpust neposredno na nevronih kot



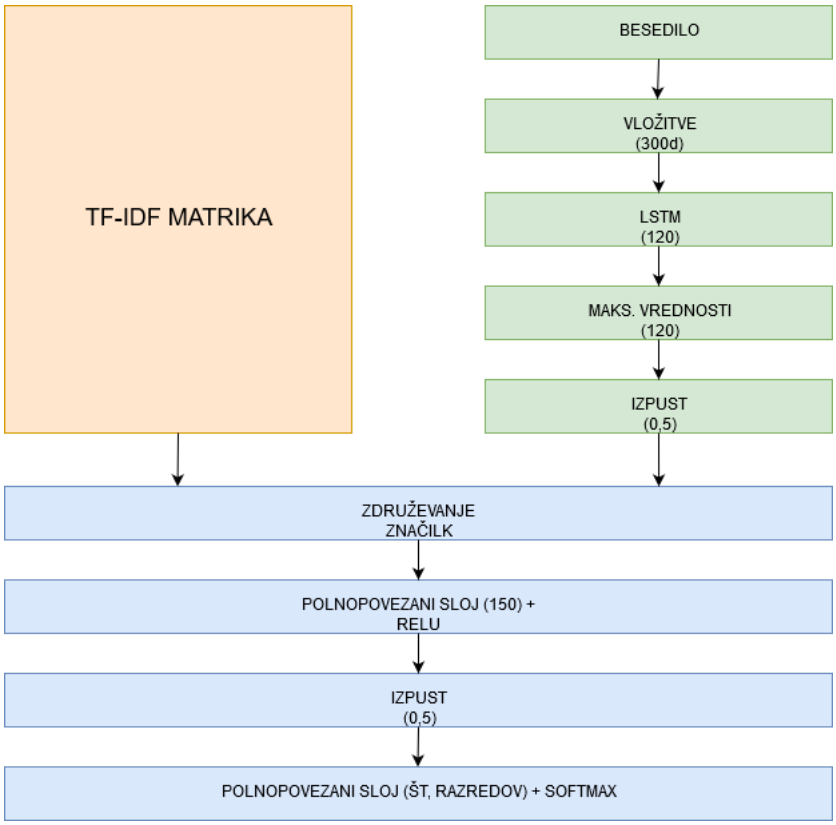
Slika 1: Porazdelitev dolžin besedil v korpusu glede na število besed

tudi na povratnih zankah. Parameter p, ki določa verjetnost izpusta posameznega nevrona ali zanke, smo za obe vrsti izpusta nastavili na 0,4. Na izhodu iz sloja s povratno zanko smo nato uporabili operacijo globalne izbire maksimalnih vrednosti (ang. global max pooling) prek vseh notranjih vektorskih predstavitev, ki jih mreža generira pri obdelavi sekvence. Na ta način smo želeli, da model med samodejno generiranimi notranjimi predstavitvami besedila uporabi samo tiste, ki najbolje predstavljajo vhodno besedilo. Na ta način smo dodatno zmanjšali dimenzionalnost vhoda, tako da smo izbrali zgolj najbolj diskriminativne značilke.

Tudi v drugem obdelovalnem cevovodu smo vsako vhodno besedilo iz učne množice najprej tokenizirali. Nato smo izdelali dve matriki s TF-IDF obtežitvami. V prvi matriki smo obtežili vse besedne N-grame dolžine od 1 do 5, pri čemer smo dodali omejitev, da v končno matriko vključimo zgolj tiste N-grame, ki se v učni množici pojavijo vsaj petkrat. Na tak način smo nekoliko zmanjšali dimenzionalnost matrike, pri čemer smo predpostavljali, da tak poseg odstrani vse izjemno redke besede in besedne zveze, ki ne dodajo veliko teže pri klasifikaciji, ter besede in besedne zveze, ki so morda zatipkane. Vse tako pridobljene N-grame smo obtežili s sublinearno obtežitvijo. V drugi matriki smo s TF-IDF utežmi obtežili vse N-grame, sestavljene iz znakov dolžine od 1 do 7. Tudi v tem primeru smo iz končne matrike odstranili vse N-grame, ki se pojavljajo manj kot 5-krat, preostale N-grame pa obtežili s sublinearno obtežitvijo. Nato smo obe matriki združili v eno TF-IDF matriko dimenzije R^{nxm}, kjer n predstavlja število besedil v učni množici, m pa število značilk. Matriki, ki smo jih pridobili iz obeh obdelovalnih cevovodov, smo nato združili in posredovali polnopovezani nevronski mreži, ki skrbi za končno klasifikacijo. Mreža je sestavljena iz ene skrite plasti in ene izhodne plasti. Skrita plast vsebuje 150 nevronov, kot aktivacijska funkcija pa je uporabljena funkcija Relu.

Kot regularizacijsko metodo smo tudi na tem sloju uporabili izpust z vrednostjo parametra p 0,5. Izhodno plast sestavljajo trije nevroni, torej po en nevron za vsako ciljno oznako. Na izhodnem sloju uporabimo aktivacijsko funkcijo softmax, ki vrne verjetnostno porazdelitev vhodnega parametra glede na ciljne oznake. Kot napovedani razred nato vzamemo tisto ciljno oznako z največjo verjetnostjo. Za bolj celovit pregled je opisana arhitektura grafično predstavljena na Sliki 2.

Opisano arhitekturo si deli vseh pet modelov, ki smo jih naučili v sklopu te raziskave. Modeli se med seboj razlikujejo v predobdelavi vhodnih podatkov in uporabljenih vektorskih vložitvah besed. Pri modelu LSTM_BOW smo vhodna besedila minimalno obdelali, in sicer smo zgolj vse velike črke v celotni novici spremenili v male. Pri modelu LSTM_BOW+TPROC so bila vhodna besedila nekoliko temeljiteje predobdelana, in sicer smo iz besedil odstranili vse številke, neinformativne besede in ločila. Pri modelih LSTM_BOW+REF in LSTM_BOW+TPROC+REF smo vektorske vložitve na podlagi metode, opisane v poglavju 3., obogatili z informacijami o sentimentu besed. Sentimente slovenskih besed smo pridobili v leksikonu sentimenta JOB (Bučar et al., 2018). Leksikon JOB je bil sestavljen na istem korpusu kot smo ga uporabili za učenje našega modela. Vendar so informacije iz leksikona bile pridobljene s pomočjo oznak na nivoju posameznih povedi besedil v korpusu, medtem ko smo za učenje naših modelov uporabili oznake na nivoju besedil. Modela se razlikujeta po predobdelavi besedil, in sicer je bila za model LSTM_BOW+TPROC+REF uporabljena temeljitejša predobdelava, ki je bila uporabljena tudi za model LSTM_BOW+TPROC. Pri zadnjem modelu LSTM_BOW+REF2 so bila besedila enako minimalno obdelana kot pri modelu LSTM_BOW, uporabljene pa so bile s sentimentom obogatene vektorske vložitve. Za razliko od ostalih modelov, kjer se vektorske vložitve med učenjem niso posodabljale, smo jih v tem primeru med sa-



Slika 2: Arhitektura modelov nevronskih mrež.

mim učenjem modela dodatno prilagajali. S tem smo upali, da bodo na podlagi podatkov iz učnega korpusa zajele dodatne informacije, ki so specifične za podani problem, kar bi pozitivno vplivalo na zmogljivost modela.

Izvirno podatkovno zbirko smo razdelili na učno in testno množico v razmerju 80:20. Vse modele smo naučili s pomočjo stohastičnega gradientnega spusta z velikostjo paketov 32. Uteži modelov smo med učenjem prilagajali z algoritmom ADAM (Kingma in Ba, 2014), pri čemer je bila začetna učna stopnja nastavljena na 0,001. Mdele smo učili 10 epoh, pri čemer smo v vsaki epohi model učili na celotni učni množici. Uspešnost naučenih modelov smo na testni množici ocenili s pomočjo dveh standardnih mer, in sicer klasifikacijsko točnostjo in povprečno oceno F1.

Tabela 1: Primerjava rezultatov modelov, naučenih v sklopu te raziskave, in osnovnih modelov.

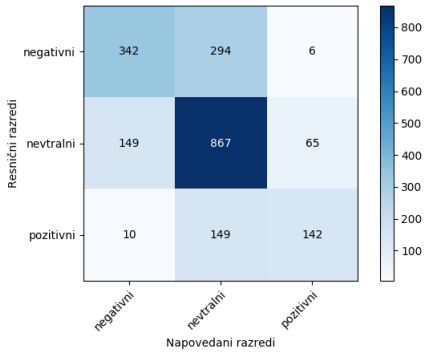
Model	Točnost	F1
SVM (poročano v (Bučar et al., 2018))	66.5	63.4
NBM (poročano v (Bučar et al., 2018))	64.3	65.9
SVM (ponovitev poizkusa)	62.7	59.0
NBM (ponovitev poizkusa)	53.8	24.57
LSTM_BOW	66.5	61.6
LSTM_BOW+TPROC	66	61.1
LSTM_BOW+TPROC+REF	65.9	61
LSTM_BOW+REF	66.4	62.1
LSTM_BOW+REF2	66.7	62.5

5. Rezultati

V tem poglavju predstavljamo rezultate vseh petih naučenih modelov. Kot osnovo za primerjanje zmogljivosti modelov smo uporabili modela iz Bučar et al. (2018). Tudi ta modela sta naučena za prepoznavanje sentimenta v novicah na isti podatkovni zbirki, pri čemer model na podlagi metode podpornih vektorjev (SVM) dosega višjo klasifikacijsko točnost, model, naučen na podlagi metode naivnega Bayesa (NBM), pa dosega višjo oceno F1.

Za testiranje modelov so avtorji v izvirnem članku izvedli 5 ponovitev 10-kratnega križnega preverjanja. Zaradi znatno višje računske zahtevnosti naših modelov, enakega načina testiranja nismo mogli ponoviti, zato predpostavljamo, da neposredna primerjava med dvema študijama ni povsem primerljiva. Zato smo modela na podlagi metod podpornih vektorjev in naivnega Bayesa, kot sta opisana v izvirnem članku, znova naučili na naši razdelitvi podatkovne zbirke. Na naši razdelitvi sta oba modela dosegla nižje rezultate. Rezultate modelov nevronskih mrež in modelov, ki smo jih vzeli za osnovo, predstavljamo v Tabeli 1.

V primerjavi z osnovnimi modeli, ki so bili naučeni na naši razdelitvi podatkovne zbirke, dosegajo vsi modeli nevronskih mrež opazno boljše rezultate tako po klasifikacijski točnosti kot po oceni F1. V primerjavi z modeli iz izvirnega članka je zmogljivost naučenih modelov nevronskih mrež primerljiva s poročanimi rezultati. Modela, ki najslabše klasificirata, LSTM_BOW+TPROC in LSTM_BOW+TPROC+REF, se glede na točnost povsem približata SVM-ju, medtem ko po oceni F1 zaostajata za približno 2 %. Dober rezultat doseže najenostavnejši model



Slika 3: Matrika zamenjav najuspešnejšega nevronskega modela LSTM_BOW+REF2.

LSTM_BOW, ki je po natančnosti izenačen s SVM-jem, izboljša pa tudi oceno F1 v primerjavi z našima najšibkejšima modeloma. Najboljše rezultate dosegata nevronska modela s popravljenimi vektorskimi vložitvami LSTM_BOW+REF in LSTM_BOW+REF2. Glede na klasifikacijsko točnost sta oba modela povsem izenačena s SVM-jem, pri čemer ga drugi model celo preseže za 0,2%.

Zanimiva je tudi primerjava modelov nevronskih mrež med sabo. Kljub temu, da se metode predobdelave besedila smatrajo kot zelo učinkovite pri učenju klasifikacijskih modelov, se je v naši raziskavi izkazalo ravno nasprotno. Ne samo, da pri modelih LSTM_BOW+TPROC in LSTM_BOW+TPROC+REF, ki uporabljata bolj izrazito predobdelavo vhodnih besedil, ni opaziti izboljšav, zdi se tudi, da sama predobdelava besedila rezultate celo nekoliko poslabša. Vnašanje informacij o sentimentu v vektorske vložitve je imelo nekoliko boljši vpliv. Oba modela, ki sta uporabljala popravljene vektorske vložitve, LSTM_BOW+REF in LSTM_BOW+REF2, sta bila po klasifikacijski točnosti primerljiva z osnovnim SVM-jem, v primerjavi z ostalimi modeli nevronskih mrež pa sta izboljšala oceno F1 za 0,5-1 %.

S pregledom matrik zamenjav posameznih naučenih modelov (glej Sliko 3) pridobimo boljši vpogled v napake, ki jih delajo modeli pri napovedovanju sentimenta. V članku smo vključili zgolj matriko zamenjav najuspešnejšega nevronskega modela LSTM_BOW+REF2, vendar matrike zamenjav vseh nevronskih modelov, ne glede na razlike v predobdelavi podatkov, kažejo dokaj konsistentno sliko. Modeli najmanj napak delajo pri klasifikaciji nevtralnih primerov, nekoliko slabše se odrežejo pri klasifikaciji negativnih primerov, najslabše pa klasificirajo pozitivne primere. Ta razporeditev je konsistentna z neuravnoteženostjo primerov v uporabljenem korpusu podatkov, ki v primerjavi z nevtralnimi primeri vsebuje skoraj dvakrat manj negativnih primerov in kar trikrat manj pozitivnih primerov. Nadalje modeli ne delajo veliko napak med dvema skrajnima sentimentoma, negativnim in pozitivnim. Pri vseh modelih je skupno število takih napak na testni množici namreč manjše od 20. Večino napak tako modeli naredijo pri klasifikaciji negativnih in pozitivnih primerov v nevtralne. To je še posebej očitno pri pozitivnem razredu, ki je tudi edini razred, pri katerem vsi modeli več primerov klasificirajo napačno kot pravilno. Iz teh rezultatov lahko torej sklepamo, da je razpoznavanje skrajnih sentimentov (pozitivnega in negativnega) v novicah dokaj preprosta naloga za nevronske modele, precej težje pa se odločajo pri vmesnem nevtralnem razredu. Glede na to, da so informacije v novicah podane veliko objektivneje kot v nekaterih drugih vrstah besedil, na primer ocenah izdelkov, lahko sklepamo, da vsebujejo tudi manj besed, ki izražajo močan negativen ali pozitiven sentiment. Vsled pomanjkanja takih očitnih nosilcev sentimenta, modeli tako pogosto klasificirajo novice v vmesni nevtralni razred, čeprav so morda te novice pri bralcih izzvale čustveno reakcijo.

Upoštevati moramo tudi, da je vsak učni (in testni) primer označen s konsenzom šestih označevalcev. Poleg strinjanja označevalcev na celotni podatkovni zbirki, bi bilo zato vredno raziskati nestrinjanje tudi na nivoju posameznih razredov. Če bi se izkazalo, da je nestrinjanje med označevalci precej veliko za novice, ki spadajo v nevtralni razred, bi lahko slabšo sposobnost modela na takih primerih pripisali tudi šumu v podatkih. Na podlagi tega bi lahko sklepali, da je konsistentno označevanje nevtralnih novic trd oreh tudi za ljudi.

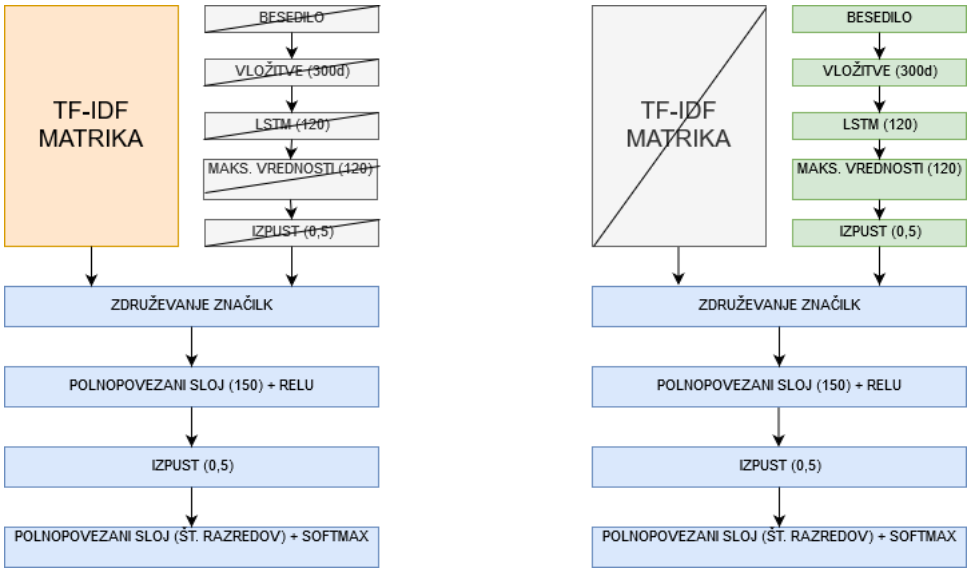
6. Primerjava napovedne učinkovitosti posamezne skupine značilk

V zadnjem delu tega članka bomo predstavili še rezultate primerjave napovedne učinkovitosti posamezne skupine značilk nevronskih modelov. Vsi naučeni modeli si delijo isto arhitekturo z dvema obdelovalnima cevovodoma. En cevovod je sestavljen iz nevronskih mrež s povratno zanko z dolgoročnimi in kratkoročnimi pomnilnimi celicami, ki iz vhodnih besedil samodejno izluščijo značilke, drugi cevovod pa vhodna besedila predstavi v obliki TF-IDF matrike. Napovedno učinkovitost posamezne vrste značilk smo preverjali tako, da smo iz arhitekture odstranili en obdelovalni cevovod in model na učni množici naučili samo s pomočjo preostalih značilk. Za vsak model smo nato izračunali klasifikacijsko točnost in oceno F1 ter rezultate primerjali. Shemo eksperimenta predstavlja Slika 4.

Rezultati kažejo, da so TF-IDF uteži za podani problem nekoliko bolj učinkovite, saj model s TF-IDF utežmi doseže za 3,7 % višjo klasifikacijsko točnost in za 5,8 % višjo oceno F1, kot če značilke generira nevronska mreža s povratno zanko. Opazimo lahko tudi, da sta oba modela zelo primerljiva z modeli iz glavne študije, pri čemer model, ki uporablja zgolj TF-IDF uteži kot značilke celo preseže modele iz glavne študije za 0,5 % po oceni F1.

7. Zaključek

V raziskavi smo se osredotočili na problem samodejnega zaznavanja sentimenta v slovenskih novicah, problema, ki zaenkrat še ni bil podrobno raziskan. Cilj naloge je bil razvoj modela, ki bi uspešno napovedal sentiment v novicah, kot ga dojema povprečen slovenski bralec. Problem smo zastavili kot klasifikacijski problem, kjer smo poskušali novice kategorizirati v eno od treh kategorij: negativno, pozitivno in nevtralno. V ta namen smo razvili arhitekturo nevronskih mrež, ki posamezno novico predstavi na dva načina, in sicer s pomočjo vektorskih vložitev in kot



Slika 4: Priprava arhitekture modelov za primerjavo vpliva značilk na klasifikacijo.

Tabela 2: Rezultati primerjave napovedne učinkovitosti posamezne vrste značilk.

Značilke	Predobdelava besedila	Točnost	F1
Samodejno generirane z LSTM plastmi	Sprememba velikih črk v male	61,4 %	57,1 %
TF-IDF matrika	Sprememba velikih črk v male	65,1 %	62,9 %

matriko n-gramov s TF-IDF utežmi. Za učenje modelov smo uporabili korpus novic SentiNews, ki vsebuje 10.427 novic, ročno označenih s sentimentom. Skupaj smo naučili pet modelov z enako arhitekturo, ki so se rahlo razlikovali v predstavitvi besedil. Kot osnovo za primerjavo rezultatov smo vzeli rezultate modelov iz Bučar et al. (2018), ki so po naših informacijah tudi edini modeli, ki modelirajo ta fenomen na slovenskih besedilih. Modeli v izvirnem članku so bili evalvirani s petkratno ponovitivjo 10-kratne križne validacije, česar zaradi računske zahtevnosti nevronskih modelov nismo mogli ponoviti. Da bi zagotovili primerljivost rezultatov, smo modele, kot so opisani v izvirnem članku, tudi sami naučili na naši razdelitvi učne in testne množice. Modele nevronskih mrež smo nato primerjali tako z rezultati iz ponovljenega poskusa kot z rezultati modelov iz izvirnega članka.

Rezultati raziskave so vzpodbudni, saj so naši modeli rezultate iz ponovljene študije presegli. Poleg tega so dosegli povsem primerljivo klasifikacijsko točnost z najboljšimi osnovnimi modeli iz izvirnega članka ter se jim po oceni F1 povsem približali. Tako smo potrdili prvo hipotezo, da so lahko nevronske mreže konkurenčne tudi na področju analize sentimenta, kjer navadno prevladujejo bolj tradicionalne metode strojnega učenja, predvsem metoda podpornih vektorjev. Hkrati smo potrdili tudi drugo hipotezo, in sicer, da se lahko modeli na osnovi nevronskih mrež, ki navadno zahtevajo večje količine podatkov, uspešno učijo tudi na manjših korpusih. Vse naše modele smo naučili na korpusu, ki vsebuje zgolj okrog 10.000 novic.

Za konec smo opravili še pregled vpliva posamezne vrste značilk na rezultate klasifikacije. Študijo smo opravili tako, da smo naučili dva modela, pri vsakem pa izločili eno od vrst značilk. Rezultati so pokazali, da so n-grami s TF-IDF utežmi nekoliko bolj diskriminativni za klasifikacijo kot samodejno generirane značilke plasti s povratno zanko, pri čemer model s tako vrsto predstavitve besedil celo preseže modele iz glavne študije glede na oceno F1.

Verjamemo, da arhitektura, uporabljena v tej raziskavi, še ni dosegla mej zmogljivosti. Za začetek bi bilo treba hiperparametre modela dodatno optimizirati. Treba je namreč pripomniti, da smo vrednosti hiperparametrov izbrali glede na vrednosti, ki v praksi dosegajo ugodne rezultate, niso pa nujno optimalne za dani problem. S pomočjo tehnike mrežnega ali naključnega iskanja bi lahko poskusili te vrednosti dodatno optimizirati.

Trenutni nevronski modeli so naučeni in preizkušeni na novicah z ekonomsko, finančno ali politično vsebino, ni pa iz trenutnih rezultatov razvidno, kako bi se ti modeli obnašali pri klasifikaciji novic z drugačno vsebino. Preizkus modelov na novicah iz drugih domen bi bolje ocenil možnost prenosa in uporabe naših modelov na različnih vrstah novic, vendar ker taka označena podatkovna zbirka po naših informacijah ne obstaja, prepuščamo tak preizkus modelov za nadaljnje raziskave.

Najnovejši nevronski klasifikacijski modeli so osnovani na predhodno naučenih jezikovnih modelih, med katerimi sta najbolj znana modela BERT in XLM. Ti modeli so pomembno izboljšali rezultate na številnih nalogah na področju računalniške obdelave naravnega jezika, zato bi bila njihova uporaba za modeliranje sentimenta v slovenskih novicah zanimivo izhodišče za nadaljnje raziskave.

8. Zahvala

Avtor članka se zahvaljuje doc. dr. Petri Kralj Novak in izr. prof. dr. Biljani Milevi Boshkoski za vse nasvete, vodenje in mentorstvo, ki sta mu jih nudila med potekom te raziskave.

9. Literatura

Ghazaleh Beigi, Xia Hu, Ross Maciejewski in Huan Liu. 2016. An overview of sentiment analysis in social media and its applications in disaster relief. V: *Sentiment analysis and ontology engineering*, str. 313–340. Springer.

Jože Bučar, Martin Žnidaršič in Janez Povh. 2018. Annotated news corpora and a lexicon for sentiment analysis in slovene. *Language Resources and Evaluation*, 52(3):895–919.

Jacob Devlin, Ming-Wei Chang, Kenton Lee in Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin in Tomas Mikolov. 2018. Learning word vectors for 157 languages. *arXiv preprint arXiv:1802.06893*.

Gagandeep Kaur in Kamaldeep Kaur. 2015. Sentiment analysis on punjabi news articles using svm. *Int. J. Sci. Res., 6 (8)*, str. 414–421.

Diederik P Kingma in Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Abhishek Kumar, Abhishek Sethi, Md Shad Akhtar, Asif Ekbal, Chris Biemann in Pushpak Bhattacharyya. 2017. Iitpb at semeval-2017 task 5: Sentiment prediction in financial text. V: *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, str. 894–898.

Xiaodong Li, Haoran Xie, Li Chen, Jianping Wang in Xiaotie Deng. 2014. News impact on stock price return via sentiment analysis. *Knowledge-Based Systems*, 69:14–23.

Kevin Hsin-Yih Lin, Changhua Yang in Hsin-Hsi Chen. 2008. Emotion classification of online news articles from the reader's perspective. V: *2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, zvezek 1, str. 220–226. IEEE.

Youness Mansar, Lorenzo Gatti, Sira Ferradans, Marco Guerini in Jacopo Staiano. 2017. Fortia-fbk at semeval-2017 task 5: Bullish or bearish? inferring sentiment towards brands from financial news headlines. *arXiv preprint arXiv:1704.00939*.

Yelena Mejova. 2009. Sentiment analysis: An overview. *University of Iowa, Computer Science Department*.

Andrew Moore in Paul Rayson. 2017. Lancaster a at semeval-2017 task 5: Evaluation metrics matter: predicting sentiment from financial news headlines. *arXiv preprint arXiv:1705.00571*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng in Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. V: *Proceedings of the 2013 conference on empirical methods in natural language processing*, str. 1631–1642.

Soonh Taj, Baby Bakhtawer Shaikh in Areej Fatemah Meghji. 2019. Sentiment analysis of news articles: A lexicon based approach. V: *2019 2nd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*, str. 1–5. IEEE.

Marjan Van de Kauter, Diane Breesch in Véronique Hoste. 2015. Fine-grained analysis of explicit and implicit sentiment in financial news articles. *Expert Systems with applications*, 42(11):4999–5010.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola in Eduard Hovy. 2016. Hierarchical attention networks for document classification. V: *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, str. 1480–1489.

Liang-Chih Yu, Jin Wang, K Robert Lai in Xuejie Zhang. 2017. Refining word embeddings for sentiment analysis. V: *Proceedings of the 2017 conference on empirical methods in natural language processing*, str. 534–539.