Detekcija spojlera u književnim kritikama upotrebom rekurentnih neuronskih mreža i attention mehanizma

Marko Vuković
Fakultet tehničkih nauka
Univerzitet u Novom Sadu
21000 Novi Sad
marko.vukovic9@uns.ac.rs

***Apstrakt*—Spojler predstavlja informaciju koja unapred otkriva sadržaj književnog, filmskog ili televizijskog dela. Kako društvene mreže beleže sve veći broj aktivnih korisnika, javlja se problem i sa rasprostranjenosti spojlera u njihovim okvirima. Psihološka istraživanja pokazala su da različiti ljudi različito reaguju na spojlere. Određene društvene mreže nude mogućnost svojim korisnicima da sadržaj koji postavljaju označe posebnim spojler oznakama, i tako daju mogućnost drugim korisnicima da sami odluče žele li takav sadržaj da vide ili ne. Međutim, praksa je pokazala da samo mali broj korisnika zapravo koristi ove mehanizme. Iz tog razloga pojavljuje se potreba za modelom koji će automatski moći da detektuje spojlere u sadržaju kritika, kako bi takvi sadržaji mogli biti sakriveni. Cilj ovog rada je bio da prilagodi HAN arhitekturu za klasifikaciju dokumenata na problem detekcije spojlera, uzevši u obzir to da klasifikacija kritika na klase sadrži spojler i ne sadrži spojler nije jasna kao neki tipičniji primeri tog problema, pre svega zbog subjektivnosti. Uzevši ograničenost skupa podataka i računarskih resursa, dobijeni su solidni rezultati koji ukazuju na to da HAN arhitektura može da se primeni na opisani problem, i da se uz proširenje skupa podataka i raspolaganjem većim računarskim resursima može obučiti model koji uz izuzetno visoku tačnost detektuje spojlere.**

Ključne reči—klasifikacija teksta; bidirekciona rekurentna neuronska mreža; GRU; attention mehanizam; spojler

I. UVOD

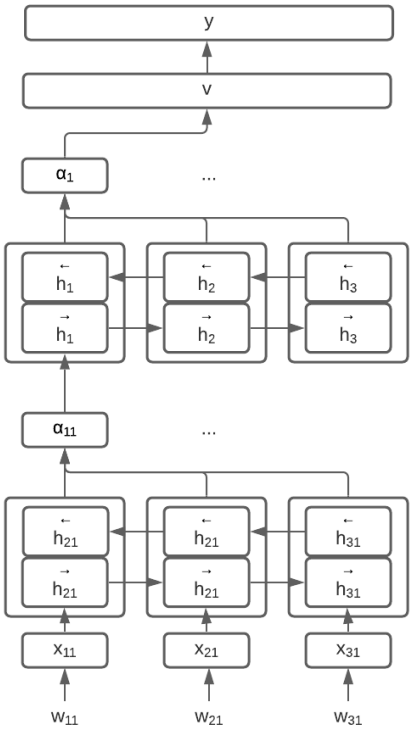
Spojler (eng. *spoiler*) predstavlja informaciju koja unapred otkriva sadržaj (najčešće rasplet) književnog, filmskog ili televizijskog dela i tako utiče na smanjeno interesovanje čitalaštva ili gledališta, a samim tim i na prihode vlasnika prava [1]. Rastom popularnosti društvenih mreža uvećala se i rasprostranjenost spojlera, pa se postavlja pitanje da li i u kom obliku spojleri treba da postoje u okviru društvenih mreža. Mnoga psihološka istraživanja, uključujući i [2] pokazala su da spojleri različito utiču na različite tipove ličnosti, odnosno na ljude sa različitom izraženošću pojedinih potreba. Eksperimentalno je pokazano da osobe koje imaju nisku potrebu za kognicijom (eng. *need for cognition*) - odnosno potrebu za strukturiranjem situacija na smislen i integrativan način, tj. osobe koje u procesu razumevanja sveta ne teže samostalnom traženju novih informacija i razmišljanju o njihovom značenju i međusobnim odnosima, već se u tom procesu pouzdaju u druge ljude - preferiraju uživanje o tzv. *spojlovanom* sadržaju (uživanje u sadržaju nakon čitanja spojlera). S druge strane, osobe koje imaju visoku potrebu za afektima (eng. *need for affect*) - odnosno potrebu da ne izbegavaju emotivne situacije, tj. osobe koje žele da dožive i razumeju svoje emocije i emocije drugih ljudi – više uživaju u sadržaju koji nije spojlovan. Iz ovog proizilazi da mesto za spojlere u okviru društvenih mreža postoji, ali da takav sadržaj treba da bude prilagođen pripadnicima obema prethodno navedenih grupa. Najčešći mehanizam koje društvene mreže nude kako bi se ovaj problem rešio je mogućnost označavanja komentara ili kritike oznakom sadrži spojler (eng. *spoiler tag*), bilo autora prilikom pisanja, ili drugih korisnika prilikom čitanja, te se ovakav sadržaj potom skriva i prikazuje samo onim korisnicima koji izraze želju da ga takvom sadržaju pristupaju. U praksi se ovo nije pokazalo kao idealno rešenje, jer veoma mali broj korisnika zapravo koristi ove opcije. Zbog toga se u poslednjih nekoliko godina ispituju računarski pristupi rešavanja prethodno opisanog problema, pre svega predlaganjem modela koji vrše automatsku detekciju spojlera. Istraživanja su vršena pomoću različitih tehnika modelovanja: upotrebom tzv. *topic* modela [3], zatim korišćenjem modela *Support Vector Machine* [4][5], ali i upotrebom *deep learning* metoda, pre svega rekurentnih neuronskih mreža sa LSTM (*Long Short Term Memory*) ćelijama [6]. Klasifikacija teksta opsežno je istražena tema u literaturi. Arhitektura koja se posebno dobro pokazala u rešavanju ovog problema je HAN (*Hierarchical Attention Network* – hijerarhijska mreža sa slojem pažnje) [7] koja je zasnovana na bidirekcionim rekurentnim neuronskim mrežama sa GRU (*Gated recurrent unit*) ćelijama i *Attention* mehanizmom. Svrha ovog istraživanja je postavljanje problema detekcije spojlera kao problem klasifikacije dokumenta (književne kritike) na dve disjunktne klase (sadrži spojler i ne sadrži spojler) upotrebom modifikovane HAN arhitekture. Pronađeno je srodno istraživanje [8], koje modifikuje HAN arhitekturu odbacivanjem drugog *attention* mehanizma mreže i direktno koristi vektorske reprezentacije pojedinačnih rečenica kritika za detekciju spojlera. Ovo istraživanje razlikuje se od prethodno navedenog po tome što koristi sve slojeve HAN mreže, za dobijanje reprezentacije cele kritike književnog dela,

a zatim vrši klasifikaciju na osnovu dobijene vektorske reprezentacije cele kritike.

U nastavku, biće detaljnije opisano rešenje prethodno opisanog problema. U poglavlju II opisan je skup podataka, u poglavlju III opisana je primenjena metodologija (arhitektura modela) i diskutovani su dobijeni rezultati. Poglavlje IV sumarizuje rad, iznosi postignute rezultate i navodi moguće pravce za buduća istraživanja.

II. Opis skupa podataka

Za potrebe implementacije arhitekture i rešavanja polaznog problema upotrebljen je skup podataka *Goodreads Dataset* [8][9], odnosno njegov podskup *Review subset with parsed spoiler tags* [10]. Skup podataka sadrži približno 1.380.000 kritika sa društvene mreže i servisa za evidentiranje knjiga *Goodreads* [11], za 25.475 različitih knjiga od 18.982 različita korisnika. Polazni skup podataka sadržao je sledeća obeležja: obeležje `user_id` koje predstavlja identifikaciono obeležje korisnika servisa *Goodreads* – odnosno autora kritike, `book_id` koje predstavlja identifikaciono obeležje knjige na koju se kritika odnosi, `review_id` predstavlja identifikaciono obeležje same kritike, `timestamp` predstavlja vremensku odrednicu koja se odnosi na kreiranje kritike, `rating` predstavlja opciono obeležje koje sadrži brojčanu ocenu (od 1 do 5) koju je korisnik dodelio knjizi, `review_sentences` predstavlja složeno obeležje koje sadrži niz rečenica kritike uz oznaku da li pojedinačna rečenica sadrži spojler ili ne i `has_spoiler` koje predstavlja obeležje koje sadrži informaciju o tome da li kritika sadrži makar jednu rečenicu koja sadrži spojler. Od navedenih obeležja korišćena su obeležja `review_sentences` kao obeležje koje služi za obučavanje i `has_spoiler` kao ciljna labela (obeležje koje se predviđa). Uočeno je nekoliko problema sa izabranim skupom podataka. Prvi problem je taj da je skup podataka u velikoj meri nebalansiran, pri čemu kritike sa spojlerima predstavljaju tek 6.5% od celokupnog skupa, (89.627 kritika koje sadrže makar jednu rečenicu koja sadrži spojler od ukupno 1.378.033), odnosno tek 0.3% rečenica sadrži spojler u odnosu na ukupan broj rečenica (odnosno 569.724 rečenica koje sadrže spojler od ukupnog broja od 17.672.655 rečenica). Drugi problem leži u veličini skupa podataka, koja se pokazala kao prevelika za ograničene računarske resurse dostupne pri obučavanju modela. Oba prethodno navedena problema rešena su na sledeći način: doneta je odluka da se model trenira tako da vrši detekciju spojlera na nivou kritike (naspram detekcije na nivou rečenice) i da se upotrebom mehanizma *random undersampling* – nasumično odbace primeri dominantne klase (u ovom slučaju to je klasa – ne sadrži spojler), kao i manji broj primera manjinske klase. Rezultat ovog procesa je skup podataka koji sadrži 100.000 kritika, pri čemu su podjednako prisutni primeri i jedne i druge klase. Mehanizam *random undersampling* izabran je zbog toga što se u literaturi pokazao kao solidno rešenje, posebno kada se u obzir uzme njegova jednostavnost [12][13]. Skup podataka na kraju je podeljen na trening, validacioni i test skup, koji predstavljaju 80%, 10% i 10% polaznog skupa, odnos često korišćen u literaturi, a posebno u radu koji uvodi arhitekturu korišćenu u ovom rešenju [7].



Slika 1- Šematski prikaz arhitekture rešenja

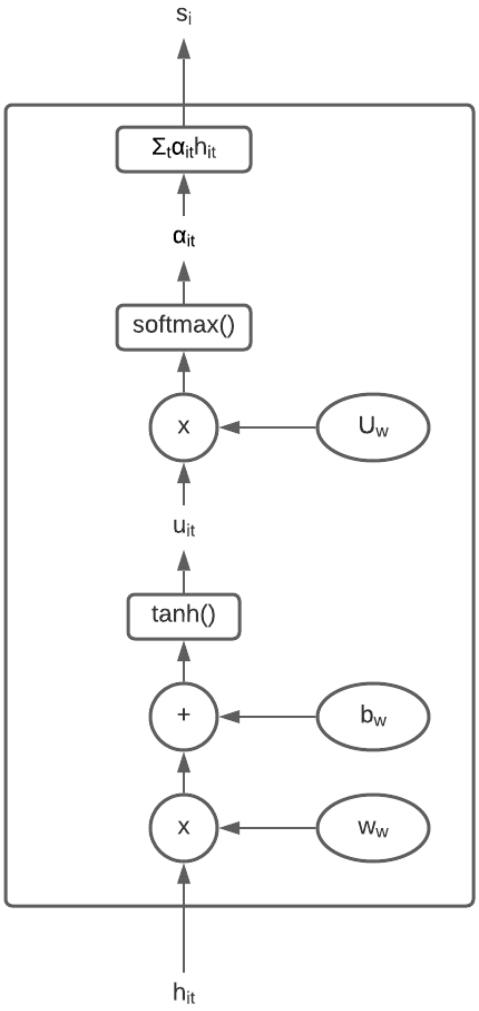
III. Metodologija

Problem detekcije spojlera formulisan je kao problem binarne klasifikacije kritike na dve klase – sadrži spojler i ne sadrži spojler.

Formiran je predikcioni model po uzoru na HAN model [7], uz nekoliko izmena radi prilagođavanja arhitekture posmatranom problemu i skupu podataka. Slika 1 sadrži šemu arhitektureSlika 1- Šematski prikaz arhitekture rešenja rešenja, a u nastavku je dato detaljno objašnjenje svakog segmenta arhitekture.

Ulazni sloj mreže predstavlja *embedding* sloj koji služi za transformaciju reči u vektorski prostor sa kojim je moguće dalje raditi. Ovakav sloj vrši transformaciju reči u n-dimenzioni vektor, pri čemu rastojanje između vektora bliskih reči treba da bude malo, dok rastojanje između vektora reči čije značenje nije slično treba da bude veliko. Za rešavanje ovog problema izabran je algoritam GloVe (*Global Vectors for Word Representation*) [14] i prethodno obučen model nad skupovima podataka *Wikipedia 2014* i *Gigaword 5*, koji sadrži 400.000 različitih reči i njihove transformacije u 100-dimenzione vektore.

Naredni sloj arhitekture predstavlja bidirekciona rekurentna neuronska mreža sa GRU ćelijama. Svrha ovog sloja arhitekture je da enkodira kontekst reči *r* u okviru rečenice. Sve reči prolaze kroz mrežu u oba smera: od prve ka poslednjoj ($\overrightarrow{h_{it}}$) i od poslednje ka prvoj ($\overleftarrow{h_{it}}$). Reprezentacija konteksta reči *i* rečenice *t* dobija se konkatenacijom ova dva vektora $h_{it} = [\overrightarrow{h_{it}}; \overleftarrow{h_{it}}]$.



Slika 2- Šematski prikaz arhitekture prvog attention sloja

Sledi prvi *attention* sloj arhitekture. Oba *attention* sloja imaju istu strukturu koju prikazuje Slika 2. Prvi *attention* sloj služi da istakne najznačajnije reči u okviru svake rečenice. Očekuje se da će se težine ovog sloja obučiti tako da pažnju skreću na reči koje otkrivaju radnju. Ulazni vektor množi se sa novim težinskim vektorom i dodaje mu se novi *bias*, kako bi se dodatno poboljšala reprezentacija reči. Zatim se novodobijeni vektor propušta kroz funkciju *tanh*, kako bi se njegova vrednost normalizovala na skup [-1,1] (1). Zatim se dobijeni vektor množi sa novim težinskim vektorom i propušta kroz funkciju *softmax* (2), čija je formula data u (3). Na kraju se sumiraju proizvodi reprezentacija reči dobijenih u (2) α_{it} sa ulaznim vektorima h_{it} za svaku reč u rečenici (4) kako bi se dobio vektor s_i koji predstavlja vektorsku reprezentaciju rečenice.

$$u_{it} = \tanh(W_w h_{it} + b_w) \quad (1)$$

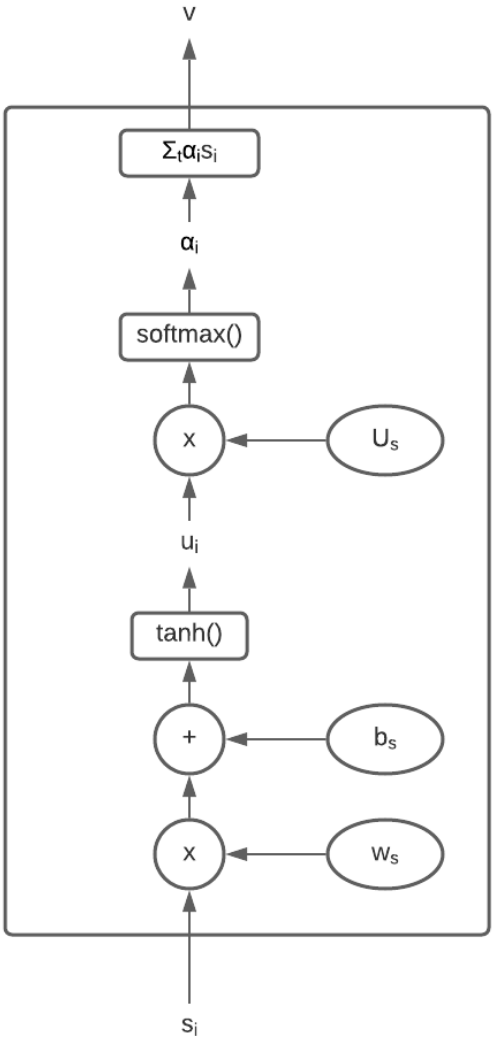
$$\alpha_{it} = \text{softmax}\,(U_w u_{it}) \quad (2)$$

$$\alpha_{it} = \frac{e^{U_w u_{it}}}{\sum_t e^{U_w u_{it}}} \quad (3)$$

$$s_i = \sum_t \alpha_{it}\, h_{it} \quad (4)$$

Prethodno opisani deo arhitekture služi za transformaciju konkretnih reči jedne rečenice u njenu vektorsku predstavu. Deo arhitekture opisan u nastavku ne razlikuje se od prethodno opisanog sloja u strukturi, već samo u tome što kao ulazne podatke prima vektorske reprezentacije rečenica i njih transformiše u vektorsku reprezentaciju cele kritike.

Prvi sloj drugog dela arhitekture predstavlja bidirekciona rekurentna neuronska mreža sa GRU ćelijama. Svrha ovog sloja arhitekture je da enkodira kontekst rečenica *s* u okviru kritike. Sve reprezentacije rečenica dobijene u prethodnom delu



Slika 3- Šematski prikaz arhitekture drugog attention sloja

arhitekture prolaze kroz mrežu u oba smera: od prve ka poslednjoj ($\overrightarrow{s_i}$) i od poslednje ka prvoj ($\overleftarrow{s_i}$). Reprezentacija konteksta rečenice *i* u kritici dobija se konkatenacijom ova dva vektora $h_i = [\overrightarrow{h_i}; \overleftarrow{h_i}]$.

Sledi drugi *attention* sloj arhitekture (Slika 3), koji služi da istakne najznačajnije rečenice u okviru kritike. Očekuje se da će se težine ovog sloja obučiti tako da pažnju skreću na najznačajnije rečenice sa stanovišta odavanja radnje. Ulazni vektor množi se sa novim težinskim vektorom i dodaje mu se novi *bias*, kako bi se dodatno poboljšala reprezentacija rečenice. Zatim se novodobijeni vektor propušta kroz funkciju *tanh*, kako bi se njegova vrednost normalizovala na skup [-1,1] (5). Zatim se dobijeni vektor množi sa novim težinskim vektorom i propušta kroz funkciju *softmax* (6). Na kraju se sumiraju proizvodi reprezentacija rečenica dobijenih u (6) α_i sa ulaznim vektorima s_i za svaku rečenicu u kritici (7) kako bi se dobio vektor v koji predstavlja vektorsku reprezentaciju cele kritike.

$$u_i = \tanh(W_s s_i + b_s) \quad (5)$$

$$\alpha_i = \frac{e^{U_s u_i}}{\sum_T e^{U_s u_i}} \quad (6)$$

$$v = \sum_T \alpha_i \, s_i \quad (7)$$

Izlazni sloj arhitekture je potpuno povezan sloj koji kao ulaz prima vektorsku reprezentaciju kritike, a kao izlaz daje verovatnoće za svaki od ishoda: jeste spojler i nije spojler.

Kao funkcija greške korišćena je funkcija *cross entropy*, korak napredovanja (eng. *learning rate*) iznosio je 0.01, i upotrebljen je algoritam napredovanja (eng. *optimizer*) ADAM, dok je za analizu performansi upotrebljena tačnost (eng. *accuracy*).

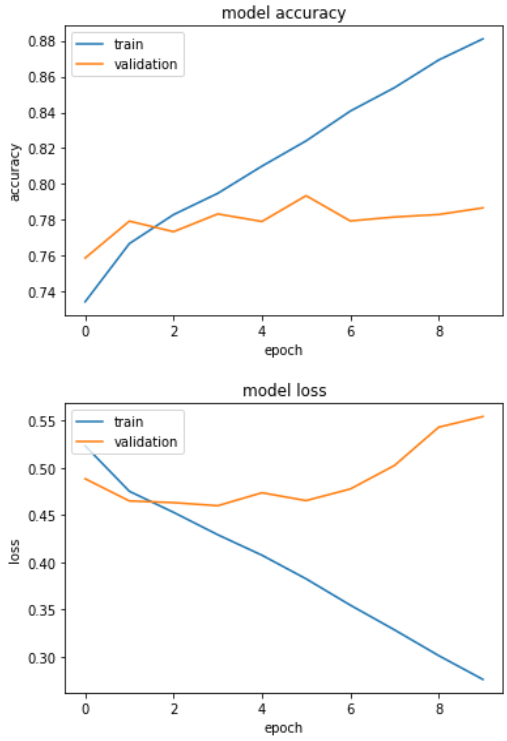
Parametri koji se odnose na ulazni skup podataka (maksimalan broj rečenica u kritici, maksimalan broj reči u rečenici, maksimalan broj reči na nivou svih kritika) birani su na osnovu statističkih podataka o skupu podataka.

Planirano je optimizovanje hiperparametara modela upotrebom algoritma *Grid Search*, ali se od toga odustalo zbog ograničenih računarskih resursa, i hiperparametri su izabrani na osnovu literature, pre svega na osnovu radova [7][8].

Dobijeni su sledeći rezultati (Slika 4): tačnost na test skupu (skupu koji model prethodno nije video) iznosila je 78%, što je prilično dobar rezultat kada se u obzir uzme relativno mali skup podataka i preostala ograničenja nametnuta ograničenim računarskim resursima. Primetno je da je model na granici *overfitting*-a, pre svega zbog upotrebe relativno malog skupa podataka, dok se nešto veće vrednosti funkcije gubitka (eng. *loss*) mogu pripisati nemogućnosti preciznog definisanja pojma spojler.

U nastavku će biti predstavljene i prodiskutovane neke od tipičnih predstavnika uočenih grupa sličnih grešaka.

Listing 1 ilustruje prvu tipičnu grupu grešaka koje se javljaju, a to su *relevatorne* reči (reči tipične za odavanje radnje



Slika 4- Prikaz rezultata treniranja

i spojlere, kao što su smrt, ubistvo i slične) koje zapravo ne odaju samu radnju. Autor ove kritike daje osnovne informacije o liku, ali *attention* mehanizam modela daje prevelik značaj reči smrt (eng. *death*) zbog čega model na kraju daje pogrešnu predikciju.

book explores life of joanna, a hospice volunteer and her relationships. Three times divorced and faces death on a regular basis .

Listing 1- Primer greške zbog upotrebe relevatornih reči koje zapravno ne odaju samu radnju

Listing 2 ilustruje primer greške koji je sličan prethodnom tipu grešaka. Autor ove kritike veoma često navodi reči kraj i završetak (eng. *end* i eng. *ending*). Razlog zbog kog se ovaj primer navodi kao posebna grupa grešaka je zbog ilustracije problema definisanja samog pojma spojler. Iako ovaj komentar nije označen kao komentar koji sadrži spojler, mnogi korisnici sigurno bi ga smatrali spojlerom, zbog toga što autor govori o svojim osećanjima povodom određenih događaja u knjizi, pre svega povodom završetka. Često je moguće na osnovu nečijih osećanja povodom završetka knjige pretpostaviti sam rasplet, posebno u slučaju knjiga koje izlaze u nastavcima, kada su čitaoci već upoznati sa likovima i zapletom, pa ova grupa neispravnih predikcija ne mora neminovno predstavljati grešku.

it is always bittersweet to see beloved series come to an end.
...
it has since became my favorite ending.

Listing 2- Primer upotrebe reči poput završetak i kraj

Primećeno je i da je izvesan broj grešaka posledica načina na koji je skup podataka formiran. Naime, skup podataka je formiran preuzimanjem kritika knjiga sa društvene mreže i servisa za evidentiranje knjiga *Goodreads*. *Goodreads* nudi korisnicima mogućnost ostavljanja komentara o knjigama, ali i mogućnost da svaki komentar označe oznakom sadrži spojler. Uočeno je da određen broj prijavljenih grešaka u predikciji zapravo posledica toga da je autor pogrešno označio svoj komentar kao da komentar ne sadrži spojler, dok on zapravo sadrži ili obrnuto.

IV. Zaključak

U ovom radu predstavljen je model za detekciju spojlera u književnim kritikama zasnovan na tehnikama mašinskog učenja – rekurentnim neuronskim mrežama i *attention* mehanizmu. Problem je značajno rešavati zbog toga što se broj korisnika društvenih mreža svakodnevno uvećava, a samim tim i broj komentara i kritika koji oni objavljuju. Uprkos tome što su mnogi servisi za evidentiranje i ocenjivanje filmova, serija i knjiga implementirali mehanizme za označavanje spojlera u sadržajima, praksa je pokazala da se krajnji korisnici njima ne služe u dovoljnoj meri. Ovaj rad predstavlja model za automatsko detektovanje spojlera, koji se zasniva na široko rasprostranjenom i korišćenom modelu HAN (*Hierarchical attention network*). Ovaj model sastoji se od dve bidirekcione rekurentne neuronske mreže sa dva sloja *attention* mehanizma, koji postepeno kreiraju vektorsku reprezentaciju dokumenta (kritike dela). Model je obučavan na skupu podataka sa kritikama prikupljenih sa društvene mreže i servisa za evidentiranje knjiga *Goodreads*. Uzevši u obzir ograničenost računarskih resursa koja je nametnula druga ograničenja u implementaciji, uključujući i smanjenje polaznog skupa podataka, dobijeni su prihvatljivi rezultati tačnosti od oko 80%. Ova implementacija marginalno je bolja u odnosu na modele koje nisu zasnovane na tehnikama *deep learning*-a, ali je u svetu rekurentnih neuronskih mreža tek može da posluži kao polazni model (*benchmark*) i osnova za dalji razvoj.

Mogući pravci daljeg razvoja rešenja pre svega jesu treniranje nad celokupnim skupom podataka, ali i poboljšanjem tog skupa podataka, a zatim pronalaženje boljeg rešenja problema nejednake zastupljenosti klasa za klasifikaciju. Jedno od mogućih poboljšanja svakako je i pronalaženje ili formiranje skupa podataka sa pismenijim kritikama (zbog problema sa lematizacijom i *embedding*-om) i preciznije i pouzdanije anotiranim komentarima. Dodatno je moguće kombinovati ovaj model sa drugim tehnikama za klasifikaciju teksta, kao što su tf-idf (*term frequency-inverse document frequency*) statistike, koje bi doprineti poboljšanju performansi. Na kraju, moguće je kombinovati *Goodreads* skup podataka sa nekim drugim skupovima podataka, zbog toga što kritike knjiga, filmova i serija poseduju neke svoje specifičnosti i obrasce.

Literatura

[1] NDNV. (2019). Rečnik aktuelnih termina u medijskoj industriji. [online] Available at: http://www.ndnv.org/2019/07/11/recnik-aktuelnih-termina-u-medijskoj-industriji/

[2] Rosenbaum, J.E. and Johnson, B.K., 2016. Who's afraid of spoilers? Need for cognition, need for affect, and narrative selection and enjoyment. Psychology of Popular Media Culture, 5(3), p.273.

[3] Guo, S. and Ramakrishnan, N., 2010, August. Finding the storyteller: automatic spoiler tagging using linguistic cues. In Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010) (pp. 412-420).

[4] Jeon, S., Kim, S. and Yu, H., 2016. Spoiler detection in TV program tweets. Information Sciences, 329, pp.220-235.

[5] Shiratori, Y., Maki, Y., Nakamura, S. and Komatsu, T., 2018, September. Detection of football spoilers on twitter. In International Conference on Collaboration Technologies (pp. 129-141). Springer, Cham.

[6] Ueno, A., Kamoda, Y. and Takubo, T., 2019. A spoiler detection method for japanese-written reviews of stories. International Journal of Innovative Computing Information and Control, 15(1), pp.189-198.

[7] Yang, Z., Yang, D., Dyer, C., He, X., Smola, A. and Hovy, E., 2016, June. Hierarchical attention networks for document classification. In Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies (pp. 1480-1489).

[8] Wan, M., Misra, R., Nakashole, N. and McAuley, J., 2019. Fine-Grained Spoiler Detection from Large-Scale Review Corpora. arXiv preprint arXiv:1905.13416.

[9] Wan, M. and McAuley, J., 2018, September. Item recommendation on monotonic behavior chains. In Proceedings of the 12th ACM Conference on Recommender Systems (pp. 86-94).

[10] UCSD Book Graph - Reviews. [online] Available at: https://sites.google.com/eng.ucsd.edu/ucsdbookgraph/reviews?authuser=0/.

[11] Goodreads. [online] Available at: https://www.goodreads.com/.

[12] Van Hulse, J., Khoshgoftaar, T.M. and Napolitano, A., 2009, August. An empirical comparison of repetitive undersampling techniques. In 2009 IEEE international conference on information reuse & integration (pp. 29-34). IEEE.

[13] Van Hulse, J., Khoshgoftaar, T.M. and Napolitano, A., 2007, June. Experimental perspectives on learning from imbalanced data. In Proceedings of the 24th international conference on Machine learning (pp. 935-942).

[14] Pennington, J. (2014). GloVe: Global Vectors for Word Representation. [online] Stanford.edu. Available at: https://nlp.stanford.edu/projects/glove/.