Prepoznava lažnih novic s pomočjo jezikovnega modela BERT

Urban Potočnik, Brin Soko, Andrej Košir

Univerza v Ljubljani, Fakulteta za elektrotehniko, Tržaška 25, 1000 Ljubljana, Slovenija
E-pošta: up8707@student.uni-lj.si

Fake news detection using large language model BERT

In recent years, the prevalence of fake news on the internet has significantly increased, posing a serious threat to an informed public and democratic processes. This paper presents an approach to automatic fake news detection using BERT, a state-of-the-art language model for natural language processing. We provide a comparison of various methods from the literature, highlight the strengths and weaknesses of the models, and analyse their robustness in the presence of linguistic manipulation. For experimental evaluation, we use the public two large language models BERT and DistilBERT. Our model achieves high accuracy and demonstrates potential for further application in detecting misleading content in online media.

1 Uvod

S širjenjem družbenih omrežij in digitalnih medijev je širjenje lažnih novic postalo globalen izziv, ki vpliva na javno mnenje, volitve, javno zdravje in družbeno stabilnost. Ob tem se je močno povečala količina lažnih novic (ang. fake news), ki namenoma širijo zavajajoče ali popolnoma izmišljene informacije.

Avtomatsko odkrivanje lažnih novic je postalo pomembno raziskovalno področje v okviru umetne inteligence, saj ročno preverjanje ni več izvedljivo zaradi obsega podatkov in hitrosti širjenja.

V tem članku obravnavamo problem samodejne detekcije lažnih novic s pomočjo sodobnih metod strojnega učenja. Posebno pozornost namenimo uporabi modela BERT (angl. Bidirectional Encoder Representations from Transformers) in njegove variante DistilBERT, enega najsodobnejših jezikovnih modelov za obdelavo naravnega jezika (NLP). Problem zastavimo kot binarno klasifikacijo, kjer želimo posamezno novico uvrstiti med resnične ali lažne.

Ker so slike pogosto del lažnih novic (npr. manipulirane fotografije, slike izven konteksta), je njihova vključitev v proces klasifikacije pomembna za izboljšanje točnosti. Zato se zadnja leta razvijajo večmodalni pristopi, ki združujejo besedilne in slikovne informacije v skupen model za detekcijo dezinformacij.

Članek je organiziran na naslednji način: v drugem poglavju predstavimo pregled trenutnega stanja na področju. Tretje poglavje opisuje uporabljene podatke, model in vrednotenje uspešnosti. V četrtem poglavju predstavimo eksperimentalne rezultate, v petem pa sledijo zaključki, diskusija in predlogi za prihodnje delo.

2 Trenutno stanje na področju

Zaradi eksponentne rasti spletnih medijev, blogov in predvsem družbenih omrežij, kot so Facebook, Twitter in Reddit, v zadnjem desetletju, je zaznavanje lažnih novic postalo ena ključnih raziskovalnih tem v računalništvu, družboslovju ter informacijskih znanostih. S pojavom velikih količin vsebin, ki jih generirajo uporabniki in niso nujno preverjene, se je povečalo tudi širjenje dezinformacij in manipulativnih vsebin. Razvoj učinkovitih metod za njihovo zaznavanje ni le tehnični izziv, temveč ima pomembne posledice tudi za družbo, demokracijo in informacijsko pismenost. V znanstveni literaturi se pojavlja širok spekter metod, ki vključujejo klasične pristope strojnega učenja, napredne modele globokega učenja ter integracijo zunanjih informacijskih virov in podatkov o širjenju vsebin. V tem poglavju predstavljamo ključne raziskave s področja zaznavanja lažnih novic, njihove metodološke pristope, uporabljene algoritme ter dosežene rezultate.

2.1 Klasični pristopi strojnega učenja

Eni izmed prvih poskusov avtomatiziranega zaznavanja lažnih novic so temeljili na ekstrakciji značilnosti iz besedilnih vsebin, kot so pogostost določenih besed, dolžina stavkov, raba zaimkov, čustvena obarvanost in druge stilistične ter lingvistične značilnosti. Na teh značilnostih so raziskovalci gradili klasične klasifikatorje, kot so Naivni Bayes, logistična regresija, naključni gozdovi ter podporni vektorski stroji (SVM), ki so se izkazali za hitro izvedljive in razložljive.

Na primer, Rashkin et al. [1] so uporabili besedilne značilnosti za razlikovanje med verodostojnimi, lažnimi in satiričnimi novicami. Uporabili so podatkovni niz, ki je vključeval članke z različnih spletnih virov (npr. *NYTimes*, *Breitbart*, *InfoWars*) in zgradili klasifikator na osnovi logistične regresije. Njihovi rezultati so pokazali natančnost med 60 % in 80 % glede na vrsto vsebine in razpoložljiv kontekst, kar kaže, da tudi enostavni modeli lahko dosegajo solidne rezultate, zlasti če so podatki dobro strukturirani.

Podobno sta Pérez-Rosas et al. [2] izvedla analizo lingvističnih značilnosti na več različnih podatkovnih nizih in dosegla F1-mere med 0.74 in 0.84. Avtorja sta izpostavila, da določeni lingvistični vzorci, kot je raba absolutnih trditev ali emocionalno nabitih besed, pogosto korelirajo z lažnimi vsebinami. Klasični pristopi se še danes uporabljajo kot osnova za hitre rešitve ali kot del hibridnih sistemov, ki vključujejo več nivojev obdelave.

2.2 Pristopi globokega učenja

Razvoj globokega učenja je pomembno izboljšal sposobnosti modelov pri razumevanju kompleksnih jezikovnih struktur, semantike in konteksta. Namesto ročne ekstrakcije značilnosti ti modeli samodejno iz besedila učijo reprezentacije, kar omogoča boljše razločevanje subtilnih razlik med verodostojnimi in zavajajočimi vsebinami.

Zhang et al. [3] so uporabili konvolucijske nevronske mreže (CNN) in rekurentne modele z dolgoročnim pomnilnikom (LSTM) za zaznavanje lažnih novic. Njihova analiza je pokazala, da CNN dobro zaznava lokalne vzorce v kratkih tekstih, medtem ko LSTM bolje modelira dolgoročne semantične odnose in kontekst. LSTM je dosegel boljše rezultate v primerih, kjer je bilo pomembno razumevanje narativa, ne le posameznih stavkov.

Eden najbolj vplivnih modelov v zadnjem času je BERT, ki sta ga razvila Devlin et al. [4]. BERT uporablja dvostransko pozornost (bi-directional attention), kar omogoča, da vsaka beseda v stavku ≫vidi≪ celoten kontekst, tako pred kot po sebi. Model je bil pred-učen na velikih korpusih in se nato dodatno učil na nalogi detekcije lažnih novic. BERT je postal de-facto standard za številne NLP naloge, saj pogosto presega rezultate klasičnih pristopov in tudi drugih globokih arhitektur.

Singhania et al. [5] so razvili arhitekturo 3HAN, ki uporablja hierarhično pozornost na treh nivojih: stavčni, odstavčni in dokumentni. Na podatkovnem naboru FakeNewsNet so dosegli 91 % natančnost, kar potrjuje učinkovitost večnivojske analize vsebine. Takšna arhitektura je še posebej uporabna pri obravnavi daljših besedil, kjer razumevanje globalne strukture pomembno prispeva k pravilni klasifikaciji.

2.3 Uporaba podatkovnih metapodatkov in zunanjih virov

Zgolj obdelava besedila pogosto ni dovolj za zanesljivo zaznavanje dezinformacij, saj lahko lažne novice jezikovno posnemajo verodostojne članke. Zato raziskovalci vključujejo tudi druge vire informacij, kot so metapodatki (datum, avtor, vir novice), družbeni kontekst (komentarji, delitve, odzivi) in pretekla zanesljivost vira.

Shu et al. [6] so predstavili večmodalni pristop, ki vključuje vsebine člankov, informacije o avtorju, število delitev in komentarjev na družbenih omrežjih ter zgodovino verodostojnosti vira. Njihov model ≫evidence-aware≪ združuje podatke iz različnih virov z uporabo mehanizmov združevanja pozornosti (attention fusion), kar izboljša zmožnost detekcije v kompleksnih primerih. Ta pristop se je posebej izkazal pri zaznavanju novic, kjer se vsebina ponavlja ali preoblikuje, vendar ohranja strukturo širjenja in odzive publike.

2.4 Zaznavanje s pomočjo socialnih grafov

Dezinformacije se pogosto širijo po določenih omrežjih uporabnikov, kar odpira možnost za uporabo grafskih metod za njihovo zaznavanje. Socialna omrežja imajo lastnosti majhnega sveta in visoko stopnjo homofiličnosti, kar pomeni, da se dezinformacije širijo v skupinah z visoko povezanostjo.

Lajali et al. [7] so uporabili grafsko konvolucijsko mrežo (GCN), ki modelira strukturo širjenja vsebin znotraj omrežja. Vsak uporabnik in vsaka delitev predstavljata vozlišče, povezave med njimi pa robove grafa. Na ta način lahko model zaznava vzorce, ki so značilni za viralne lažne novice, kot so hitro širjenje znotraj majhnega števila klik ali odmevnih komor. Rezultati so pokazali, da ta pristop presega tekstovne metode v primerih, ko je jezikovno razlikovanje težko ali zavajajoče.

2.5 Opomba k pristopom odkrivanja lažnih novic

Literatura jasno kaže, da je zaznavanje lažnih novic večplastna naloga, ki vključuje tako razumevanje vsebine kot tudi širšega konteksta. Klasični klasifikatorji, kot so logistična regresija ali SVM, so hitro izvedljivi in enostavni za interpretacijo, vendar pogosto zanemarjajo kontekst. Globoki modeli, kot so LSTM, CNN in BERT, omogočajo visoko natančnost, vendar zahtevajo več podatkov in računske moči. Pristopi, ki vključujejo metapodatke, socialne signale ali grafske strukture, prispevajo k večji robustnosti sistemov, saj upoštevajo širšo sliko širjenja informacij. Kombinacija teh metod – t. i. večmodalni in hibridni pristopi – se vse bolj uveljavlja kot najbolj obetavna smer za prihodnji razvoj natančnih in prilagodljivih sistemov za detekcijo dezinformacij.

3 Materiali in metode

3.1 Podatki

Pri razvoju sistema za zaznavanje lažnih novic smo uporabili podatkovni nabor Fakeddit[8], ki je dostopen v javnem repozitoriju na GitHubu. Gre za obsežen nabor objav in naslovov z družbenega omrežja Reddit, ki vključuje oznake o verodostojnosti vsebine. Vsak zapis vsebuje izvirni naslov novice, očiščeno verzijo naslova ter oznako, ki določa, ali je novica resnična ali lažna. Podatkovni nabor ne vsebuje eksplicitnih političnih primerov, zato model temelji predvsem na jezikovnih vzorcih in logičnih sklepih, ne pa na preverjanju dejstev.

Podatkovna množica vsebuje učno množico z 804 378 primeri in testno množico z 84 654 primeri. Vsebuje podatkovna polja `title`: izvirni naslov novice, `clean_title`: očiščeni naslov brez posebnih znakov, `2_way_label`: binarna oznaka (0 – resnična novica, 1 – lažna novica). Primer podatkov je {(Scientists discover new planet in solar system, 0), (Aliens landed in New York and took over the city, 1), (Government announces new tax reforms, 0), (Chocolate cures cancer, experts claim, 1)}. Za nadaljnjo obdelavo smo iz polj

`title` in `clean_title` oblikovali novo besedilno polje `combined_text`, ki je služilo kot vhod v model. Manjkajoči vnosi so bili predhodno odstranjeni. Podatki zajemajo različna tematska področja (npr. znanost, tehnologija, zabava), kar prispeva k večji raznolikosti in robustnosti modela.

3.2 Algoritem za prepoznavo lažnih novic na osnovi modela BERT

BERT je bil osrednji model za klasifikacijo. Gre za dvosmerni jezikovni model, ki na podlagi arhitekture transformerjev omogoča globoko razumevanje konteksta v besedilu, saj upošteva tako prejšnje kot naslednje besede. Postopek uporabe modela vključuje: 1.Tokenizacijo besedila z `BertTokenizer` iz knjižnice `transformers`, 2. Pretvorbo tokeniziranega besedila v numerične vektorje, 3. Učenje modela `BertForSequenceClassification` za binarno klasifikacijo novic (resnične ali lažne).

Za robustno oceno modela smo uporabili K-kratno navzkrižno validacijo (K-Fold Cross-Validation, K=2). Ta tehnika razdeli podatke na K delov, kjer se model večkrat trenira in testira na različnih podmnožicah podatkov, kar omogoča zanesljivejše meritve uspešnosti.

3.3 Algoritem DestilBERT

Za besedilni del je bil uporabljen DistilBERT, predtrenirani jezikovni model, ki predstavlja kompresirano različico modela BERT. DistilBERT temelji na arhitekturi transformatorjev in je rezultat postopka znanega kot knowledge distillation.

3.4 Algoritem CLIP

Za slikovni del objav smo uporabili model CLIP (Contrastive Language–Image Pretraining), ki združuje slikovno in besedilno razumevanje v skupen latentni prostor. CLIP je bil predtreniran na velikih količinah parov besedilo–slika, kar mu omogoča ustvarjanje semantičnih vektorskih predstavitev slik, ki jih je mogoče neposredno primerjati z besedilnimi opisi. V eksperimentu smo pridobili vektorske predstavitve slik prek CLIP-ovega vizualnega enkoderja, ki smo jih nato vključili v nadaljnjo analizo in klasifikacijo.

3.5 Merjenje uspešnosti rešitve

Za merjenje uspešnosti rešitve smo uporabili standardne metrike za klasifikacijske modele, ki omogočajo oceno natančnosti in zanesljivosti napovedi. Uporabljene metrike so naslednje: **Natančnost** (angl. Accuracy): delež pravilno napovedanih primerov glede na skupno število primerov, **Metrika F1**: harmonično povprečje med priklicem in preciznostjo, še posebej uporabno pri neuravnoteženih podatkovnih nizih, **Preciznost** (angl. Precision): delež pravilno napovedanih pozitivnih primerov glede na vse napovedane pozitivne primere, **Priklic** (angl. Recall): delež pravilno napovedanih pozitivnih primerov glede na vse dejanske pozitivne primere.

Izračun teh metrik smo izvedli z uporabo funkcij iz knjižnice `scikit-learn`. Metrike smo uporabili tako med validacijo (K-kratna navzkrižna validacija) kot pri testiranju na ločenem testnem naboru.

4 Eksperimentalni rezultati

4.1 Model BERT

Eksperimentalne nastavitve so bile naslednje: Model BERT za binarno klasifikacijo, K=2 kratna navzkrižna validacija, hiperparametri: Velikost paketa (batch size): 32, Število epoh: 5, Največja dolžina zaporedja: 128, in Evalvacijski korak: 500. Rezultati, povzeti po ključnih metrikah, so prikazani v spodnji tabeli.

Tabela 1: Rezultati modela BERT

Metrika	Vrednost
Natančnost	0.878
Preciznost	0.865
Priklic	0.894
Metrika F1	0.879

Rezultati kažejo, da model dosega visoko natančnost in dobro uravnoteženost med preciznostjo in priklicem, kar pomeni, da je uspešen tako pri pravilni identifikaciji resničnih kot tudi lažnih novic.

Kljub temu pa so se pojavile nekatere omejitve modela, predvsem pri razumevanju širšega konteksta in specifičnega znanja o svetu, kar se kaže v nelogičnih napovedih za nekatere primere. Model v določenih primerih napačno ocenjuje očitno neresnične trditve kot resnične, kar jasno kaže na potrebo po dodatnem učenju na bogatejših in bolj raznolikih podatkih. Prav tako se model težje spopada z zaznavanjem ironičnih, sarkastičnih ali domiselnih izrazov, kar vpliva na natančnost ocen v kompleksnejših besedilih.

Novica	Napoved
Drinking bleach cures COVID-19	Fake, 96.56%
President signed budget bill	Real, 36.15%
Aliens landed in NYC	Real, 0.27%
Stock market new high	Fake, 91.20%
Chocolate healthier than veggies	Fake, 99.01%
City approved new park plan	Fake, 60.83%
China to invade Taiwan soon	Fake, 88.56%
Slovenia borders Austria, Hungary, Croatia and Italy	Fake, 88.46%
Slovenia borders Ecuador	Fake, 93.83%
Joe Biden won 2024 election	Real, 0.74%

Tabela 2: Nekaj napovedi modela za ročno vnesena besedila, kjer lahko takoj opazimo. Naveden napoved Real ali Fake pomeni razvrstiev avtomatskega klasifikatorja, naveden % zanesljivost te napovedi.

4.2 Model DistilBERT

Model smo učili 10 epoh z optimizacijskim algoritmom Adam in funkcijo izgube BCEWithLogitsLoss.

Tabela 3: Uspešnost modela na testni množici

Metrika	Vrednost
Natančnost	0,835
Točnost	0,879
Priklic	0,830
Metrika F1	0,832

Rezultati kažejo, da je model uspešno kombiniral tekstovne in vizualne informacije za prepoznavanje lažnih novic, s čimer je dokazano, da večmodalni pristop omogoča učinkovito klasifikacijo.

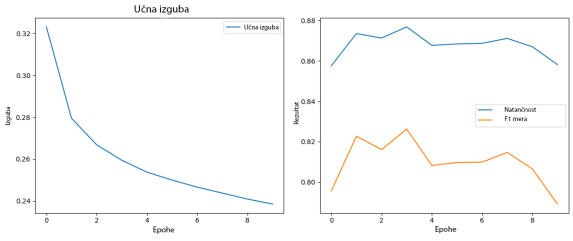
Primerjava modelov BERT in DistilBERT kaže, da je model BERT malo uspešnejši kot model DistilBERT, ki je manj požrešen za računalniške vire. Razlika je tako majhna, da ne pričakujemo, da bi v praksi pomenila opazno razliko v prepoznavanju lažnih novic.

5 Zaključki in diskusija

V okviru raziskave smo razvili in preizkusili model za prepoznavo lažnih novic na osnovi modela BERT. Eksperimentalni rezultati so pokazali, da model dosega visoko natančnost in uravnotežene metrike F1, preciznosti ter priklica, kar potrjuje njegovo zmožnost učinkovite klasifikacije novic na testnih podatkih. Kljub temu pa so se v nekaterih primerih pojavile omejitve, predvsem zaradi pomanjkanja širšega konteksta in specifičnega svetovnega znanja, kar je vodilo do napačnih napovedi.

Opazili smo, da model obstoječe podatke dobro izkorišča, vendar se njegova robustnost zmanjša pri obravnavi ironičnih, sarkastičnih ali kontekstualno zahtevnih primerov. Poleg tega ročno vneseni primeri, ki izstopajo iz učno-podatkovnega nabora, razkrivajo potrebo po dodatnem učenju na bogatejših in bolj raznolikih podatkih, ki vključujejo širše svetovno znanje.

Napredek po epohah kaže naslednja slika, izdelana je za model DistiBERT. Opazimo, da so najboljši rezultati doseženi dovolj hitro in da nadaljevanje učenja modela nima smisla.



Slika 1: Rezultati učenja modela DistilBERT po epohih

Pridobljene izkušnje kažejo, da je pri reševanju problema prepoznavanja lažnih novic ključnega pomena kombinacija zmogljive arhitekture modela, kakovostnih in raznovrstnih podatkov ter ustreznih metod validacije. Prihodnje delo bi lahko usmerili v uporabo še bolj naprednih modelov, kot so RoBERTa ali GPT, ter v integracijo zunanjih podatkovnih virov za izboljšanje semantičnega razumevanja.

Možnosti izboljšav so med drugim:

- Uporaba zmogljivejših večmodalnih Transformer modelov bi lahko še izboljšala rezultate.
- Dodatno čiščenje in balansiranje podatkov bi lahko vplivalo na izboljšanje rezultatov.
- Identifikacija primerov, kjer model dela največ napak, bi pomagala pri bolj ciljanem izboljševanju.
- Analiza napak v smislu jezikovnih karakteristik tekstov, ki so razvrščeni pravilno in tistih, ki so razvrščeni napačno.

Na splošno raziskava poudarja, da je avtomatska prepoznava lažnih novic tehnično zahtevna naloga, ki presega zgolj klasifikacijo besedila in zahteva globlji vpogled v kontekst ter širše svetovno znanje.

Zahvala. Raziskavo je podprl program P2-0246 ICT4QoL - Informacijske in komunikacijske tehnologije za kakovost življenja.

Literatura

[1] Hannah Rashkin, Maarten Sap Choi, Emily Allaway, Noah A. Smith, and Yejin Choi. Truth of varying shades: Analyzing language in fake news and political fact-checking. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2921–2927, 2017.

[2] Vera Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. Automatic detection of fake news. *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401, 2018.

[3] Jing Zhang, Lizhen Cui, Yanjie Fu, and Xiaofei Gouza. Fake news detection with deep diffusive neural network. *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 465–474, 2018.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*, pages 4171–4186, 2019.

[5] Swapna Singhania, Neha Fernandez, and S. Roy Rao. 3han: A deep neural network for fake news detection. *arXiv preprint arXiv:1705.06906*, 2017.

[6] Kai Shu, Suhang Wang Mahudeswaran, and Huan Liu. Beyond news contents: The role of social context for fake news detection. *Proceedings of WSDM*, pages 312–320, 2019.

[7] Mohammad Lajali, Moataz Qasem, and Michele Fazzolari. Fake news detection using graph convolutional networks and social context. *Expert Systems with Applications*, 213:118991, 2023.

[8] Kai Nakamura, Sharon Levy, and William Yang Wang. r/fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. 2019. Accepted at LREC 2020.