Propagandos atpažinimas lietuviškame tekste naudojant transformeriais pagrįstus, iš anksto apmokytus daugiakalbius modelius

Paulius Zaranka, Gražina Korvel

Vilniaus universitetas, Duomenų mokslo ir skaitmeninių technologijų institutas, Akademijos g. 4, LT-08412 Vilnius
paulius.zaranka@mif.vu.lt

Santrauka. Didėjant informacijos kiekiui ir jos svarbai visuomenėje atsiranda vis didesnis poreikis automatinių įrankių, gebančių atpažinti propagandą. Dėl geopolitinės situacijos Lietuvos valstybė gali būti ypatingai pažeidžiama propagandinių mechanizmų, o automatinis jos atpažinimas lietuviškuose tekstuose yra nepakankamai ištyrinėta sritis. Šio darbo tikslas – išbandyti 3 pagrindinius transformeriais pagrįstus, iš anksto apmokytus daugiakalbius modelius propagandos atpažinimui. Sprendžiamas binarinis klasifikavimo uždavinys, priskiriant tekstui propagandinio arba nepropagandinio teksto klasę. *LitLat*, *XLM-R* ir *mBERT* modeliai adaptuoti apmokant ekspertų suanotuotu duomenų rinkiniu. Nors geriausią, 88,5 % F1 statistikos įvertį pavyko pasiekti adaptavus *LitLat* iš anksto apmokytą modelį, kiti šiame darbe adaptuoti modeliai pasiekia panašius rezultatus.

Raktiniai žodžiai: propagandos atpažinimas; daugiakalbiai modeliai; transformeriai; iš anksto apmokyti modeliai; modelių adaptavimas.

1 Įvadas

Šiuolaikinėje informacijos eroje politiniai procesai pasaulyje yra stipriai veikiami propagandos [1]. Propagandiniai mechanizmai formuoja visuomenės požiūrį, elgseną ir veiksmus per sistemingą įvairios informacijos skleidimą bei sąmoningą faktų manipuliavimą. Šiais laikais internetinė erdvė yra vienas svarbiausių kanalų, per kuriuos tokia informacija plinta [2]. Nors įvairiuose kontekstuose propaganda gali turėti įvairią – tiek teigiamą, tiek neigiamą – konotaciją, dabar iš tam tikrų šaltinių sklindanti ši informacija tampa visuotiniu iššūkiu.

Šių reiškinių neigiamas poveikis visuomenėms tampa vis akivaizdesnis, todėl šiuo metu yra aktyviai vystomi tyrimai automatinio propagandos at-

DOI: https://doi.org/10.15388/LMITT.2024.22

pažinimo srityje [3][4]. Pasitelkiant natūralios kalbos apdorojimo ir giliojo mokymosi metodus kuriami modeliai, gebantys spręsti su propaganda susijusias problemas. Šis darbas tyrinėja propagandos aptikimą tekste. Pažangiausi tokio tipo tyrimai ieško būdų efektyviai spręsti uždavinį, kurio tikslas – tekste nustatyti teksto fragmentus, kuriuose yra propagandos technikų. Toks uždavinys dažnai skirstomas į du dalinius uždavinius: pirmojo uždavinio – Fragmento Identifikavimo (angl. *Span Identification*) – tikslas yra tekste atpažinti konkrečius propagandos fragmentus; antrojo uždavinio – Technikos Klasifikacijos (angl. *Technique Classification*) – tikslas – teksto fragmentui, kuris yra nustatytas kaip propagandinis, priskirti jame naudojamas technikas iš propagandos technikų sąrašo [3]. Šiame darbe išbandomi 3 pagrindiniai transformeriais pagrįsti, iš anksto apmokyti daugiakalbiai modeliai, apmokyti ir lietuvių kalba, sprendžiant paprastesnį, binarinį propagandos klasifikavimo uždavinį.

2 Transformeriais pagrįsti kalbos modeliai

Transformeriais pagrįsti kalbos modeliai jau kurį laiką dominuoja natūralios kalbos apdorojimo tyrimų srityje. Šių modelių neuroninių tinklų architektūra, leidžianti efektyviai užkoduoti ir atkoduoti žodinę informaciją [5], leido jiems pasiekti pažangiausių rezultatų daugelyje sričių, įskaitant klasifikavimą, įvardintų objektų atpažinimą, teksto generavimą ir kt. Vienas esminių šių modelių komponentų – dėmesio mechanizmas – leidžia jiems efektyviau užfiksuoti žodžių kontekstą nei ankstesnės architektūros, tokios kaip rekurentiniai neuroniniai tinklai (RNN) [5]. Spartus transformeriais pagrįstų kalbos modelių vystymasis atvedė iki tokių plačiai žinomų ir naudojamų modelių rūšių sukūrimo kaip GPT ar BERT. GPT ir kiti panašūs dideli kalbos modeliai (angl. *Large language models*) yra vienos krypties (autoregresiniai), t. y. jie generuoja tekstą nuosekliai, žodis po žodžio. Todėl, nors ir gali būti sėkmingai pritaikyti įvairioms užduotims, pagrindinis jų gebėjimas – teksto generavimas [6]. Tuo tarpu BERT yra dvikryptis modelis, kuris tekstą gali apdoroti abejomis kryptimis, t. y. iš kairės į dešinę ir iš dešinės į kairę [7]. Tai leidžia šiam ir kitiems panašiems modeliams geriau užfiksuoti kontekstą ir žodžių tarpusavio ryšius tekste, o tai daro jį tinkamesnį užduotims, kurioms reikalingas gilus konteksto supratimas. Todėl BERT šeimos modeliai yra pažangiausi sprendžiant kalbos supratimo, tame tarpe ir propagandos atpažinimo, uždavinius [8]. Įvairūs iš anksto apmokyti kalbos modeliai yra aktyviai tyrinėjami propagandos aptikimo srityje. Dideli kalbos modeliai

GPT-3 ir GPT-4 pasiekia palyginamus rezultatus [9], tačiau BERT šeimos modelis RoBERTa išlieka pažangiausias adaptavus jį propagandos atpažinimo uždaviniams spręsti [8].

Yra sukurta daugybė transformeriais pagrįstų kalbos modelių, apmokytų anglų kalba. Tuo tarpu mažiau išteklių turinčios kalbos, įskaitant lietuvių, susiduria su problemomis. Šiuo metu nėra sukurto nei vieno plačiau žinomo modelio, iš anksto apmokyto išskirtinai lietuvių kalba. Dėl šios priežasties, naudojant pažangius iš anksto apmokytus kalbos modelius spręsti lietuvių kalbos apdorojimo užduotis, dažnai yra pasitelkiami daugiakalbiai modeliai. Trys plačiausiai naudojami tokio tipo užduotims spręsti modeliai yra *LitLat* [10], *mBERT* [11, 12] ir *XLM-R* [13, 14]. *LitLat* yra apmokytas lietuvių, latvių ir anglų kalbomis; *mBERT* ir *XLM-R* – atitinkamai 106 ir 100 kalbų, įskaitant ir lietuvių. 1 lentelėje vaizduojamas bendras šių modelių palyginimas.

1 lentelė. Daugiakalbių modelių bendras palyginimas.

Modelis	Architektūra	Kalbų kiekis	Žodyno dydis
mBERT	BERT	104	119547
XLM-R	XLM-RoBERTa	100	250002
LitLat	XLM-RoBERTa	3	84201

3 Eksperimento rezultatai

Darbe nagrinėjamų modelių propagandos atpažinimo galimybėms išbandyti atliktas eksperimentas: kiekvieno jų adaptavimas ir ištestavimas sprendžiant binarinį klasifikavimo uždavinį. Adaptuoto modelio tikslas – klasifikavimas, ar duotas tekstas yra propagandinis, ar ne. Šiam uždaviniui naudotas ekspertų suanotuotas duomenų rinkinys (N = 750), nusakantis, kuriai klasei tekstas priklauso. Pagal klases lygiai subalansuotas duomenų rinkinys buvo suskirstytas į mokymo (85 %) ir testavimo (15 %) poaibius.

Kadangi BERT ir RoBERTa architektūros kaip įvestį gali priimti tik iki 512 teksto vienetų, susidedančių iš žodžių dalių, skyrybos ženklų ir specialių teksto vienetų, ilgio sekas, modeliai buvo apmokyti klasifikuoti 512 teksto vienetų ilgio gabalus. Šie gabalai sudaryti originalų tekstą skaidant slenkančio lango principu: imama 50 persidengiančių teksto vienetų iš senesnio gabalo ir 462 teksto vienetai iš originalaus gabalo. Vieną originalų tekstą vidutiniškai sudaro apie 2,73 dalys. Bendrai visiems modeliams adaptuoti naudoti parametrai pavaizduoti 2 lentelėje.

2 lentelė. Bendri modelių apmokymo hiperparametrai. Hiperparametrai buvo parinkti nepagrindžiant pasirinktų jų verčių tyrimais.

Hiperparametras	**Reikšmė**
Optimizatorius	AdamW
Mokymo žingsnis	2e-5
Paketo dydis	8
Epochų kiekis	2

Modelio testavimas vykdytas kiekvieną testavimo duomenų rinkinio tekstą suskaidžius taip pat į 512 teksto vienetų dydžio gabalus slenkančio lango principu. Kiekvienam šiam gabalui vykdyta atskira klasifikacija. Galutinė prognozė rėmėsi daugumos principu, t. y. galutinę prognozę nulemia tai, kokios klasės gabalų tekste modelis prognozavo daugiau. Rezultatai pavaizduoti 3 lentelėje. Geriausius rezultatus, 88,5 % F1 statistikos įvertį, pasiekia *LitLat* modelis. Tuo tarpu *mBERT* ir *XLM-R* modelių rezultatai tarpusavyje yra beveik identiški ir nuo geriausio modelio atsilieka nedaug – apytiksliai 2,7 %.

3 lentelė. Klasifikavimo rezultatai.

Modelis	**Tikslumas**	**Preciziškumas**	**Atkūrimas**	**F1**
mBERT	0,858	0,867	0,858	0,858
XLM-R	0,858	0,860	0,858	0,858
LitLat	**0,885**	**0,885**	**0,885**	**0,885**

4 Išvados

Šiuo metu pasaulyje automatinis propagandos aptikimas yra aktyviai tyrinėjama natūralios kalbos apdorojimo sritis. Šis susidomėjimas kyla tiek dėl jos aktualumo socialine, tiek dėl netrivialumo mašininio mokymosi prasme. Tuo tarpu panašūs uždaviniai lietuvių kalba nėra pakankamai ištirti. Šiame darbe buvo išbandyti 3 skirtingi transformeriais paremti, iš anksto apmokyti daugiakalbiai modeliai dvejetainiam propagandos klasifikavimo uždaviniui spręsti. Pagal visus pagrindinius statistikos įverčius *LitLat* modelis pasiekė geriausius rezultatus, tačiau skirtumas nuo kitų, *mBERT* ir *XLM-R*, modelių nėra ryškus. Atsižvelgiant į šiuos rezultatus visi išbandyti modeliai gali būti laikomi tinkamais tolimesniems tyrimams.

Ateityje planuojame atlikti technikų klasifikavimą ir analizuoti jų ryšį su viso teksto klasifikavimo rezultatais, kuriuos gavome šiame straipsnyje. Toks metodas leis mums geriau suprasti, kaip atskiri teksto segmentai prisideda prie bendro teksto vertinimo, bei padės išsiaiškinti, kuriuos temos dominuoja lietuviškuose, su propaganda susijusiuose, tekstuose.

Padėka

Dėkojame projektui „Propagandos ir dezinformacijos tyrimai: automatinis atpažinimas mašininio mokymo metodais, poveikis ir visuomenės atsparumas“ (Lietuvos Respublikos Vyriausybės prioritetinių mokslinių tyrimų programa (įgyvendinama per Lietuvos mokslo tarybą) „Visuomenės atsparumo stiprinimas ir krizių valdymas šiuolaikinių geopolitinių įvykių kontekste“, dotacijos numeris S-VIS-23-8) už suteiktus duomenis ir pagalbą atliekant analizę. Taip pat, esame dėkingi Vilniaus universiteto Informacinių technologijų paslaugų centrui (VU ITPC) už suteiktus didelio našumo skaičiavimo išteklius.

Literatūra

[1] Da San Martino, G., Shaar, S., Zhang, Y., Sh, Y., Barrón-Cedeno, A., & Nakov, P. (2020). Prta: A system to support the analysis of propaganda techniques in the news. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 287-293.

[2] Lietuvos nacionalinis radijas ir televizija (LRT), URL: https://www.lrt.lt/naujienos/lrt-tyrimai/5/1700792/lrt-tyrimas-lietuvos-penktoji-kolona-rusijos-propaganda-platina-seimos-gynejai-sektos-ir-knygu-apie-stalina-leidejai (žiūrėta: 2024-02-22).

[3] Martino, G., Barrón-Cedeno, A., Wachsmuth, H., Petrov, R., & Nakov, P. (2020). SemEval-2020 task 11: Detection of propaganda techniques in news articles. arXiv preprint arXiv:2009.02696.

[4] Piskorski, J., Stefanovitch, N., Da San Martino, G., & Nakov, P. (2023, July). Semeval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multi-lingual setup. In Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023) (pp. 2343-2361).

[5] Ashish, V. (2017). Attention is all you need. Advances in neural information processing systems, 30, I.

[6] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, Ch., McCandlish, S., Radford, A., Sutskever, I. & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.

[7] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
[8] Abdullah, M., Altiti, O., & Obiedat, R. (2022, June). Detecting propaganda techniques in english news articles using pre-trained transformers. In 2022 13th International Conference on Information and Communication Systems (ICICS) (pp. 301-308). IEEE.
[9] Sprenkamp, K., Jones, D. G., & Zavolokina, L. (2023). Large Language Models for Propaganda Detection. arXiv preprint arXiv:2310.06422.
[10] LitLat BERT. URL: https://huggingface.co/EMBEDDIA/litlat-bert
[11] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
[12] Multilingual BERT. URL: https://huggingface.co/google-bert/bert-base-multilingual-cased
[13] Conneau, A., & Lample, G. (2019). Cross-lingual language model pretraining. Advances in neural information processing systems, 32.
[14] XLM-Roberta. URL: https://huggingface.co/transformers/model_doc/xlmroberta.html