Mašininio mokymosi pritaikymas reklamų aptikimui YouTube įrašuose

Karolis Kvedaravičius, Olga Kurasova

Vilniaus universitetas, Duomenų mokslo ir
skaitmeninių technologijų institutas, Akademijos g. 4, Vilnius
karolis.kvedaravicius@mif.stud.vu.lt

Santrauka. Šiame straipsnyje aprašyta kaip tyrimų metu buvo bandoma pritaikyti mašininį mokymąsi reklamų aptikimui *YouTube* vaizdo įrašuose naudojant transkribuotą tekstą. Reklamų aptikimas buvo laikomas teksto klasifikavimo užduotimi ir todėl buvo naudojamas BERT šeimos mašininio mokymosi modelis, kuris pasiekia aukštus rezultatus sprendžiant teksto analizės uždavinius. Tačiau šiam modeliui dėl įvairių priežasčių buvo sunku pasiekti aukštą tikslumo lygį. Bet naudojant antrą straipsnyje pasiūlytą klasifikavimo žingsnį, kuris atsižvelgia į BERT modelio klasifikavimą tam tikram laiko tarpe, rezultatai buvo pagerinti.

Raktiniai žodžiai: Mašininis mokymasis, reklamos, transkribuotas tekstas.

1 Įvadas

Šiais laikais *YouTube* vaizdo įrašų turinyje yra dažnai įterpiamos reklamos, kuriose įrašo kūrėjas perduoda informaciją iš rėmėjų (angl. *sponsor segments*). Šios reklamos ne visada yra aktualios žiūrovams. Jau dabar yra sistemų (viena iš jų yra *SponsorBlock*), kurios, naudodamos vartotojų įkeltus duomenis, žymi ir praleidžia reklamas *YouTube* vaizdo įrašuose. Tačiau ši sistema turi trūkumų, reikia palaukti kol kiti vartotojai sužymės reklamas ir jei įrašas neturi žiūrovų, naudojančių *SponsorBlock,* reklamos niekad nebus aptiktos.

Todėl atrodė vertinga bandyti sukurti automatizuotą sistemą, kuri atliktų reklamų aptikimo procesą automatiškai, pasitelkiant mašininį mokymąsi. Kadangi galima pasiekti *YouTube* įrašų transkribuotą tekstą per *YoutubeTransciptAPI*, šį tekstą galima klasifikuoti naudojant mašininio mokymosi modelius. Šio tyrimo tikslas apmokyti mašininio mokymo modelį reklamų aptikimo uždaviniui naudojant transkribuoto tekstą ir pateikti modelio rezultatus vykdant reklamų aptikimo uždavinį.

DOI: https://doi.org/10.15388/LMITT.2024.10

2 Literatūros apžvalga

Kadangi transkribuotame tekste reklamų paieška yra teksto klasifikavimo uždavinys, buvo nagrinėjama, kokie modernūs modeliai yra tinkami atliekant teksto klasifikavimą. 2017 metais buvo pasiūlytas naujas mašininio mokymo modelis – transformeris [1]. Jis išsiskyrė iš kitų modelių, nes savo architektūroje naudoja dėmesio (angl. attention) mechanizmą, vietoje konvoliucinių sluoksnių.

Transformerių architektūra buvo pritaikyta BERT (angl. *Bidirectional Encoder Representations from transformers*) modeliams. BERT buvo sukurtas 2018 Google mokslininkų [2]. BERT yra iš anksto apmokytas (angl. *Pre-trained*) modelis, kuris apmokytas MLM (angl. *masked language modeling*) ir NSP (angl. *Next Sentence Prediction*) uždaviniams spręsti. BERT modelis gali būti pritaikytas įvairiems natūralios kalbos apdorojimo uždaviniams. Pavyzdžiui, GLUE (angl. *General Language Understanding Evaluation*) pasiekia 80 % įvertinimą.

Taip pat BERT modelis gali būti sėkmingai pritaikytas ne tik tekstų anglų kalba analizei. Yra sukurtas lietuvių ir latvių kalbomis apmokytas BERT modelis, kuris pasiekia aukštesnius rezultatus negu daugiakalbis BERT modelis, taikomas lietuvių ir latvių kalbų uždaviniams [3].

3 Duomenų surinkimas

Kadangi nėra tinkamos viešos duomenų aibės apmokyti BERT modelį reklamų aptikimui *YouTube* įrašuose užduočiai, šio tyrimo eigoje buvo sudaryta nauja duomenų aibė. Duomenų aibė buvo surinkta naudojant *SponsorBlock* atviro kodo duomenų bazę ir *YoutubeTranscriptAPI*. Iš *SponsorBlock* paimamas įrašo URL (kad vėliau būtų galima paimti iš *YoutubeTranscriptAPI* transkribuotą tekstą) *startTime* – reklamos pradžios laiką, *endTime* – reklamos pabaigos laiką, *votes* – vartotojų įvertinimą, *type* – turinio tipą. Kadangi duomenų kiekis didelis ir jų visų negalima patikrinti, pasirenkami tik tie įrašai, kurių tipas *sponsor* ir turi daugiau negu 100 teigiamų įvertinimų.

Antra naudojant *YoutubeTranscriptAPI* ir URL paimamas įrašo transkribuotas tekstas. Grąžinamas tekstas yra suskirstytas į atkarpas (vidutiniškai 40 simbolių ilgio). Taip pat grąžinama, kurioje įrašo sekundėje atkarpa baigiasi ir kiek sekundžių atkarpa tęsiasi.

Tada *SponsorBlock* duomenų bazės atrinktos reklamos ir iš *YoutubeTranscriptAPI* surinktas tekstas naudojami sudaryti mokymosi duomenis su

klasėmis. Kiekvienos transkribuoto teksto atkarpos pradžios ir pabaigos laikai lyginami su iš *SponsorBlock* duomenų bazės surinktais reklamų pradžios ir pabaigos laikais. Jei teksto atkarpos pradžios arba pabaigos laikas įkrenta tarp reklamos pabaigos ir pradžios laiko, transkribuotas tekstas priskiriamas klasei 1 (klasė 1 žymi atkarpas, kurios yra reklamos), kitu atveju priskiriama klasei 0 (klasė 0 žymi atkarpas, kurios nėra reklamos).

Tokiu metodu iš viso buvo surinkta 741815 transkribuoto teksto eilučių. 700697 teksto eilučių priklausė klasei 0, 41118 eilučių priklausė klasei 1. Klasei 1 priklausė tik 5 % visų surinktų eilučių. Surinktų duomenų pavyzdys yra pateiktas 1 lentelėje.

Šis būdas rinkti duomenis iškelia kelias problemas. Kadangi *YouTube* įrašų kūrėjai po įkėlimo turi kelis būdus modifikuoti jau įkeltą įrašą. Vienas iš jų yra iškirpimas tam tikro fragmento įrašo. Dėl to gali būti, kad *SponsorBlock* duomenų bazėje yra pažymėta reklama, kuri naujoje įrašo versijoje neegzistuoja. Taip pat surinkti duomenys yra priklausomi, nuo kokius vaizdo įrašus žiūri *SponsorBlock* vartotojai ir kaip tiksliai yra pažymėtos reklamos šių vartotojų.

1 lentelė. Surinktų duomenų pavyzdys.

Nr.	Tekstas	Klasė
1	luck cause this isn't even the hardest part yet	0
2	just before this video gets going I want	1
3	to give a special thanks and mention to	1
4	nitrous networks our server provider for	1

4 Modelis reklamoms klasifikuoti

Šiame skyriuje aprašomas BERT apmokymas su sudaryta duomenų aibe ir apmokyto BERT modelio rezultatai sprendžiant reklamų aptikimo užduotį. Taip pat aprašyta papildomo klasifikavimo žingsnio veikimas ir rezultatai. Papildomas klasifikavimo žingsnis buvo naudojamas norint pasiekti aukštesnius klasifikavimo rezultatus.

4.1 BERT modelio apmokymas

Reklamų aptikimo iš transkribuoto teksto užduočiai buvo apmokytas *BertForSequenceClassification* modelis, kuris yra BERT modelis, specialiai pritai-

kytas klasifikavimo uždaviniams su papildomais išmetimo ir klasifikavimo sluoksniais. Kadangi BERT modelio apmokymas su pilna sudaryta duomenų aibe užtruktų ilgai, tyrimo metu modelis buvo apmokytas su mažesniu duomenų kiekiu. Apmokymas atliktas su dviem duomenų kiekiais, 100 tūkstančių ir 200 tūkstančių teksto eilučių norint ištirti duomenų kiekio įtaką BERT modelio rezultatams. Taip pat ištestuoti dveji skirtingi klasių balansai, originalus: 95 % - klasė 0, 5 % - klasė 1 ir pakeistas: : 85 % - klasė 0, 15 % - klasė 1, ir to poveikis vertinimo rezultatams. Vertinimo duomenų aibę sudarė 30 tūkstančių teksto eilučių.

Visiems modelio mokymams buvo naudotas tie patys hiperparametrai, kurie pateikti 2 lentelėje.

2 lentelė. Mokymui naudoti hiperparametrai

Partijos dydis	64
Epochų skaičius	4
Mokymosi greitis	$2{,}5e^{-5}$
Maksimalus įvesties ilgis	128

Visų atliktų apmokymų rezultatai pateikti 3–5 lentelėse. Visose lentelėse pateikti geriausių epochų rezultatai. Geriausias bendras tikslumas (angl. *accuracy*) gautas, kai modelis apmokomas su 100 tūkstančių eilučių su originaliu klasių balansu. Tačiau, metrikų reikšmės klasei 1 (reklamos) yra žemos palyginus su klasės 0. Pakeičiant klasių balansą į 85 % - klasė 0, 15 % - klasė 1, metrikos klasei 1 pagerėja, tačiau bendras tikslumas nukrenta. Padvigubinus eilučių kiekį apmokymui rezultatai pagerėja tik labai nežymiai.

BERT modeliui gali būti sunku teisingai klasifikuoti klasę 1, dėl šios klasės retumo palyginus su klase 0. Kita problema galėtų būti transkribuoto teksto eilučių ilgis. Eilučių ilgis vidutiniškai yra 40 simbolių ir galimai modelis gauna nepakankamai informacijos, kad galėtų teisingai klasifikuoti.

3 lentelė. Mokymosi rezultatas su 95000 teksto eilučių priklausant klasei 0, 5000 priklausant klasei 1.

Bendras tikslumas	**Tikslumas**	**Atpažinimas**	**F1-balas**	**Klasė**
0,95	0,96	0,99	0,97	0
	0,57	0,29	0,39	1

4 lentelė. Mokymosi rezultatas su 85000 teksto eilučių priklausant klasei 0, 15000 priklausant klasei 1.

Bendras tikslumas	Tikslumas	Atpažinimas	F1-balas	Klasė
0,92	0,97	0,95	0,96	0
	0,37	0,53	0,44	1

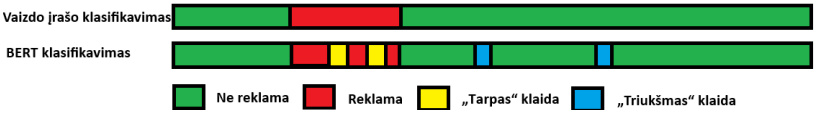
5 lentelė. Mokymosi rezultatas su 170000 teksto eilučių priklausant klasei 0, 30000 priklausant klasei 1.

Bendras tikslumas	Tikslumas	Atpažinimas	F1-balas	Klasė
0,93	0,97	0,96	0,96	0
	0,41	0,45	0,44	1

4.2 Papildomas klasifikavimo žingsnis

Peržiūrėjus kaip modelis klasifikuoja *YouTube* įrašų teksto eilutes pastebėta, kad dažnai pasikartoja dviejų tipų klaidos. *YouTube* įrašuose reklamos dažniausiai susideda iš kelių, viena po kitos einančių eilučių. Tačiau BERT modeliui sunku klasifikuoti visas reklamos eilutes. Reklamose atsiranda „tarpai“, kur neteisingai klasifikuojamos reklamos eilutės kaip ne reklamos. Gali būti, kad „tarpai“ atsiranda, nes teksto eilutės yra trumpos (maždaug 40 simbolių ilgio) ir jose kartais nėra pakankamai informacijos, kad modelis teisingai klasifikuotų.

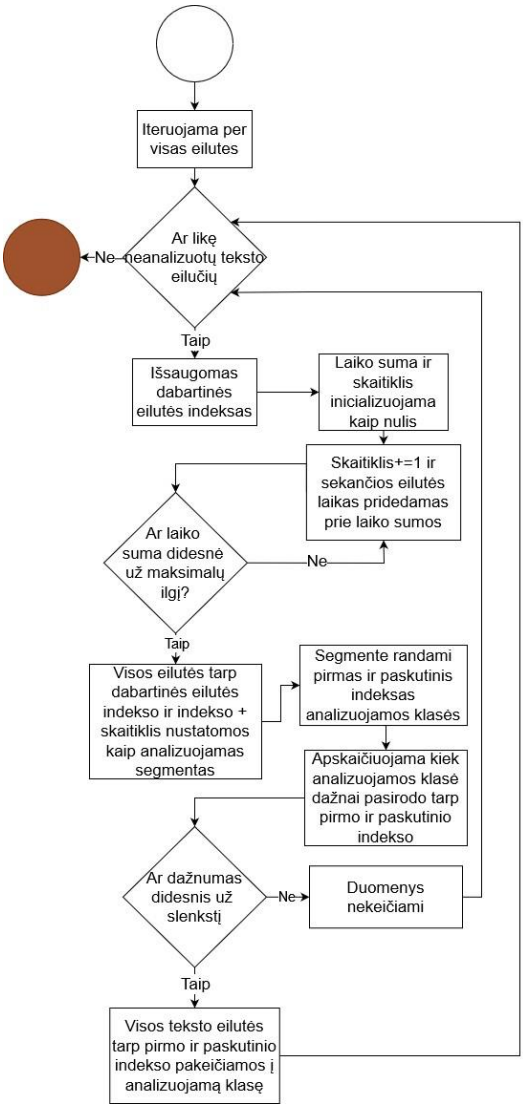
Taip pat, kartais BERT modelis neteisingai klasifikuoja pavienes eilutes kaip reklamas, nors jos nėra reklamos ir yra apsuptos ne reklamos eilučių. Šio tipo klaidos vadinamos triukšmu. 1 pav. pavaizduota kaip šios klaidos atrodytų pažymėtos *YouTube* vaizdo įrašo laiko juostoje.



1 pav. *YouTube* vaizdo įrašo teksto klasifikavimo klaidų pavyzdys

Norint sumažinti šių problemų įtaką rezultatams buvo kuriamas papildomas klasifikavimo algoritmas, kuris darė dvi prielaidas. Pirma, kad *YouTube* įrašuose reklamos yra vientisos, reklamos nebus pertrauktos trumpų sakinių, kurie nėra reklamos dalis. Antra, reklamos yra tam tikro minimalaus

ilgio, jei eilutė, modelio pažymėta kaip reklama, bus apsupta eilučių kurios nėra reklama tai bus laikoma klaida.



2 pav. Papildomas klasifikavimo algoritmas

Pagal idėją algoritmas panašus į vaizdų analizėje dažnai naudojamas morfologines operacijas: eroziją (angl. *erosion*) ir plėtimą (angl. *dilation*). Realizuojant algoritmą naudojama iš *YouTubeTranscriptAPI* gaunama informacija. *YouTubeTranscriptAPI* grąžina ne tik kiekvienos įrašo eilutės tekstą, bet ir eilutės trukmę sekundėmis. Kiekvienos eilutės trukmė naudojama nustatyti, kiek aplinkinių eilučių algoritmas analizuoja, norint nuspręsti ar apmokytas BERT modelis klaidingai klasifikavo eilutę. Algoritmas nusprendžia, kad BERT modelis klaidingai klasifikavo teksto eilutę, priklausomai nuo kaip dažnai pasikartoja tam tikra klasė analizuojamame segmente.

Algoritmo veikimas pavaizduotas 2 pav. Algoritmas turi keturis įvesties duomenis:

- Sąrašas BERT modelio kiekvienos eilutės klasifikavimo.
- Maksimalus analizuojamo segmento ilgis, kuris nustato kiek eilučių analizuojama vienu metu.
- Slenkstis, kuris nustato ar tam tikra klasė pasikartoja analizuojamame segmente pakankamai retai, kad tai būtų laikoma klaida.
- Analizuojama klasė. Jei analizuojama klasė 1 atliekamas „tarpų“ šalinamas. Jei analizuojama klasė 0 atliekamas „triukšmo“ šalinamas.

6 lentelė je pateikti vertinimo rezultatai pridėjus papildomą klasifikavimo žingsnį (BERT modelis apmokytas su 200 tūkstančių eilučių, 85 % - klasė 0, 15 % - klasė 1). Bendras tikslumas pagerėja nuo 0,93 iki 0,95 ir žymiai pagerėja F1-balas klasei 1, nuo 0,44 iki 0,55.

6 lentelė. Rezultatai naudojant papildomą klasifikavimo žingsnį

Bendras tikslumas	Tikslumas	Atpažinimas	F1-balas	Klasė
0,95	0,97	0,98	0,98	0
	0,61	0,5	0,55	1

5 Išvados

Tyrimo metu buvo sudaryta nauja duomenų aibė, kurioje *YouTube* vaizdo įrašų transkribuotas tekstas suskirstytas į reklamas ir ne reklamas. Duomenų aibė sudaryta naudojant viešai pasiekiamus duomenis iš *SponsorBlock* ir *YouTubeTranscriptAPI.* Iš viso surinkta apie 740 tūkstančių eilučių. Su šia duomenų aibe buvo apmokytas BERT modelis reklamų aptikimo uždaviniui. BERT modelis pasiekia gana aukštą bendrą tikslumą klasifikuojant *YouTube*

vaizdo įrašų transkribuotą tekstą į reklamas ir ne reklamas. Tačiau rezultatai reklamos klasei yra gana žemi, F1-balas yra tik 0,45, o ne reklamos klasė F1-balas yra 0,97. Pridėjus antrą klasifikavimo žingsnį, kuris atsižvelgia į BERT modelio klasifikavimą tam tikram laiko tarpe, rezultatai pagerėja, ypač reklamos klasei. Reklamos klasės F1-balas pagerėja nuo 0,45 iki 0,55.

Literatūra

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, „Attention Is All You Need," p. 15, 2017.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, „BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Minneapolis, 2019.

[3] Matej Ulčar, Marko Robnik-Šikonja, „Training dataset and dictionary sizes matter in BERT models: the case of Baltic languages," Lublianos Universitas, Lubliana, 2021.