Neapykantos kalbos atpažinimas lietuviškuose komentaruose panaudojant dirbtinį intelektą

Eglė Kankevičiūtė[1], Milita Songailaitė[2], Justina Mandravickaitė[3]

Vytauto Didžiojo universitetas, Vileikos g. 8, Kaunas
[1] *egle.kankeviciute@vdu.lt,*
[2] *milita.songailaite@vdu.lt,*
[3] *justina.mandravickaite@vdu.lt*

Santrauka. Šiame darbe pateikiame neapykantos kalbos aptikimo modelių palyginimą lietuvių kalbai. Neapykantos kalbai aptikti naudojome tris giliojo mokymosi modelius: daugiakalbį BERT, *LitLat* BERT ir *Electra*. Visi trys modeliai buvo adaptuoti lietuviškų komentarų klasifikavimui į tris klases: neapykantos, įžeidžią ir neutralią kalbą. Norint adaptuoti modelius atpažinti neapykantos kalbą, buvo parengtas anotuotas duomenų rinkinys, kuriame yra 25 219 lietuviški komentarai. Apmokyti modeliai buvo įvertinti naudojant tikslumo, atkūrimo, preciziškumo ir F1 statistikos metrikas. Geriausiai pasirodė *LitLat* BERT, kurio F1 statistikos reikšmė buvo 0,72. Antroje vietoje liko daugiakalbis BERT, kurio F1 statistika buvo 0,63, o trečioje vietoje liko *Electra*, kurio F1 statistika pasiekė 0,55.

Raktiniai žodžiai: dirbtinis intelektas, teksto klasifikavimas, neapykantos kalba, transformerių neuroniniai tinklai

1 Įvadas

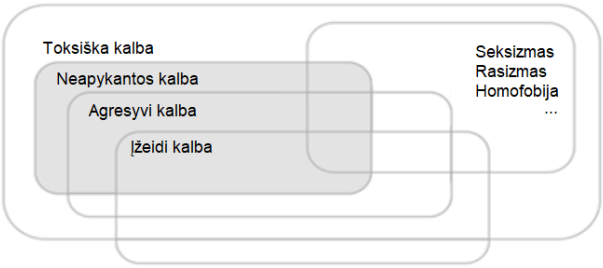
Šio darbo tikslas yra adaptuoti ir palyginti tris transformerių neuroninių tinklų modelius – daugiakalbį BERT, *LitLat* BERT ir *Electra* – neapykantos kalbos klasifikavimo uždaviniui lietuvių kalba spręsti. Tikslui pasiekti išsikelti šie uždaviniai: išanalizuoti literatūrą neapykantos kalbos atpažinimo tema; paruošti lietuviškų komentarų anotuotą tekstyną; adaptuoti pasirinktus modelius lietuviškų komentarų klasifikavimui ir galiausiai palyginti gautus modelių rezultatus.

DOI: https://doi.org/10.15388/LMITT.2023.3

2 Literatūros apžvalga neapykantos kalbos atpažinimo tema

Neapykantos kalba yra kompleksiškas reiškinys, kurio aptikimas nėra lengvas. Mokslininkai prisideda prie neapykantos kalbos identifikavimo kurdami tam skirtus sprendimus, rengdami anotuotus tekstynus, išskirdami reikšmingus požymius ir išbandydami klasifikavimo algoritmus. Neapykantos kalbos aptikimo įvairiomis kalbomis tyrimai bei parengti lyginamieji tekstynai toliau skatina neapykantos kalbos automatinio aptikimo plėtrą, nes tai gali padėti kovoti su smurto ir neapykantos internete eskalavimu ar netikrų naujienų sklaida [1].

Atpažįstant neapykantos kalbą, iššūkių gali kelti neapykantos kalbos apibrėžčių įvairovė. Neapykantos kalbai priklauso ir tokios giminingos sąvokos, kaip piktnaudžiavimas, agresyvumas, rasizmas ir kt. Išsamų šių susijusių sąvokų atvaizdavimą žr. 1 pav.



1 pav. Neapykantos kalbos ir susijusių sąvokų ryšiai [1]

Kita vertus, net ir tarp sąvokų įvairovės esama nuoseklumo. Pavyzdžiui, Fortuna ir Nunes [2] išanalizavo turimus neapykantos kalbos apibrėžimus ir nustatė šiuos panašumus:

- Neapykantos kalba turi taikinį.
- Neapykantos kalba skatina smurtą arba neapykantą.
- Neapykantos kalba puola arba žemina.
- Neapykantos kalba gali turėti tam tikrų humoro rūšių, pavyzdžiui, sarkazmo.

Apibrėžimų įvairovė leidžia daryti prielaidą, kad neapykantos kalbos suvokimas skiriasi. Šie skirtingi apibrėžimai daro įtaką esamiems duomenų rinkiniams, nes jie anotuojami remiantis jais. Tad panašūs atvejai, atsižvel-

giant į šiuos sąvokų skirtumus, gali būti priskirti skirtingoms anotavimo kategorijoms. Panašiai ir [3] nustatė, kad dauguma viešai prieinamų duomenų rinkinių yra nesuderinami dėl skirtingų neapykantos kalbos apibrėžimų, priskiriamų panašioms sąvokoms. Be to, neapykantos kalbos duomenų rinkiniai kartais turi labai panašius žymėjimus (anotavimus, pažymint teksto priskyrimą tam tikrai klasei), todėl neretai tyrimuose kai kurios atskiros duomenų klasės sujungiamos į vieną klasę (paprastai taip siekiama sumažinti klasių disbalansą) [4]. Tokia praktika gali daryti neigiamą poveikį tyrimams, nes klases atskirti būtina. [5]Pavyzdžiui, buvo pasiūlyta įžeidžią kalbą skirti nuo neapykantos kalbos, nes tai nėra tas pats, todėl šių klasių nereikėtų sujungti [5].

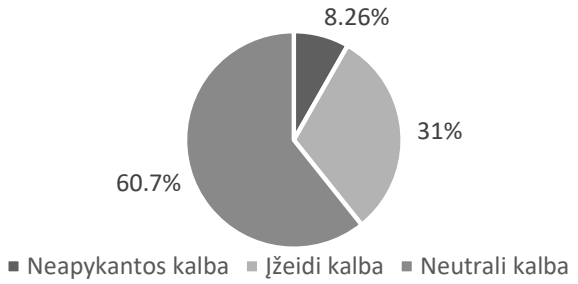
Neapykantos kalbos ir susijusio piktnaudžiaujamo elgesio aptikimo metodai tampa vis populiaresni, jų tikslumas taip pat auga [6], [7], ypač šio uždavinio sprendimui ėmus taikyti gilųjį mokymą. Kita vertus dabartiniai naujausi sprendimai vis dar turi tikslumo trūkumų, todėl jų praktinis taikymas realiuoju laiku yra ribotas [8]. Be to, neapykantos kalbos aptikimas vis dar yra itin sudėtingas uždavinys, ypač kai neapykantos apraiškos yra netiesioginės [9].

3 Duomenys, metodai ir pasiruošimas eksperimentams

3.1 Anotuotas neapykantos kalbos tekstynas

Neapykantos kalbos atpažinimo lietuvių kalba sprendimo sukūrimui buvo surinkta apie 60 000 komentarų iš įvairių naujienų portalų (15min.lt, alkas.lt, delfi.lt), taip pat tekstynas buvo papildytas 226 776 komentarais iš naujienų portalo lrytas.lt ir tūkstančiu rankiniu būdu surinktų, tik neapykantos kalbos komentarų iš įvairių socialinių tinklų puslapių ir naujienų portalų. Pastarieji komentarai buvo renkami pagal konkrečias temas. Iš viso buvo suanotuota 25 219 komentarų. Anotavime dalyvavo keturi anotatoriai, o komentarų anotavimo schemą sudarė 3 klasės (žymos): neutrali kalba, įžeidi kalba ir neapykantos kalba.

Surinktame duomenų rinkinyje duomenys pasiskirstė netolygiai. Daugiausia komentaruose rasta *non-hate* (neutralios kalbos), kuri sudaro 60,7 proc. visos duomenų aibės, mažesnę dalį užima *offensive* (įžeidžios kalbos) kategorija, kurios yra 31 proc., ir mažiausią dalį užima *hate* (neapykantos kalbos) kategorija, kuri sudaro tik 8,26 proc. visų anotuotų duomenų (žr. 2 pav.).



2 pav. Kategorijų pasiskirstymas suanotuotame tekstyne

3.2 Neapykantos kalbos atpažinimo metodologija

3.2.1 BERT architektūra paremti modeliai

Transformeris BERT (*Bidirectional Encoder Representations from Transformers*) yra dėmesio mechanizmu [10] paremtas giliojo mokymosi modelis, dažniausiai taikomas įvairiems kalbos technologijų uždaviniams spręsti [11]. Šis modelis veikia žinių perkėlimo (angl. *transfer learning*) principu [12], kuomet neuroninis tinklas yra apmokomas sugeneruoti žodžių įterpinius (angl. *word embeddings*), kurie vėliau naudojami kaip įvesties funkcijos modeliams, sprendžiantiems įvairius kalbos technologijų uždavinius.

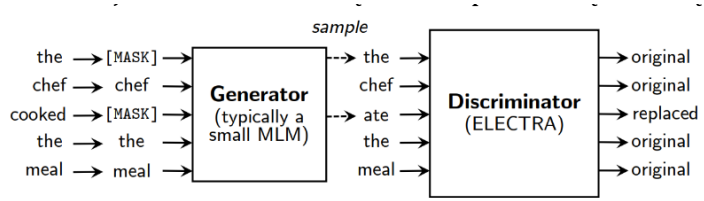
Vienas didžiausių BERT modelio pranašumų prieš kitus neuroninių tinklų modelius yra konteksto tarp žodžių, esančių tekste, supratimas. Kontekstą modelis išmoksta pasinaudodamas transformaciniams modeliams būdingu dėmesio mechanizmu, kurį sudaro užkodavimo (angl. *encoder*) ir atkodavimo (angl. *decoder*) mechanizmai [13]. Kuriant neapykantos kalbos atpažinimo sprendimą, naudoti du BERT modeliai: *Multilingual* BERT ir *LitLat* BERT.

Multilingual arba kitaip – **daugiakalbis BERT** naudoja BERT modelio architektūrą ir yra apmokytas Google komandos, panaudojant 104 skirtingas kalbas, įskaitant ir lietuvių [14]. Apmokymui buvo naudojami tekstai iš skirtingų kalbų Vikipedijos portalų, kurie nebuvo anotuojami ar kitaip apdorojami [15].

***LitLat* BERT** yra trikalbis modelis, naudojantis XLM-RoBERTa-base tvirtai optimizuotą (angl. *robustly optimized*) architektūrą [16]) ir apmokytas su lietuvių, latvių ir anglų kalbų duomenimis. Remiantis moksline literatūra, *LitLat* BERT koncentruojasi tik į tris kalbas, tad veikia geriau nei daugiakalbis BERT [17].

3.2.2 Electra modelis

Electra yra transformacinis modelis (kitaip – transformeris), naudojantis išankstinio apmokymo metodą, pagal kurį apmokomi du transformerių modeliai: generatoriaus (angl. *generator*) ir diskriminatoriaus (angl. *discriminator*). Generatoriaus tikslas yra pakeisti sekos leksemas, todėl jis mokomas kaip užmaskuotos kalbos modelis. Tuo tarpu diskriminatorius bando nustatyti, kurias leksemas pakeitė generatorius [18] (žr. 3 pav.). Generatoriumi gali būti bet koks kalbos modelis, kuris sukuria išvesties pasiskirstymą pagal leksemas, tačiau įprastai yra naudojamas MLM (angl. *Masked Language Model*), kuris apmokomas kartu su diskriminatoriumi. Po pirminio apmokymo generatorius yra išmetamas ir atliekant tolesnes užduotis tikslinamas tik diskriminatorius (*Electra* modelis) [19]. Toks pasiūlytas mokymo būdas yra žymiai efektyvesnis nei BERT modeliuose naudojamas maskuoto mokymo metodas. Būtent dėl šios priežasties *Electra* modelis apmokymui reikalauja mažiau duomenų bei kompiuterinių resursų.



3 pav. Pakeistos leksemos aptikimo pavyzdys [19]

Pagrindinis šio modelio trūkumas, lyginant su prieš tai apžvelgtais *daugiakalbiu* BERT ir *LitLat* BERT modeliais yra tai, kad lietuvių kalbai *Electra* modelis nėra iš anksto apmokytas, tad tam reikia panaudoti kuo daugiau surinktų lietuviškų tekstų ir taip iš anksto apmokyti šį modelį. Nepaisant to, šio modelio apmokymui išnaudojama mažiau kompiuterio resursų nei mokant BERT modelius.

3.2.3 Modelių apmokymas klasifikavimui

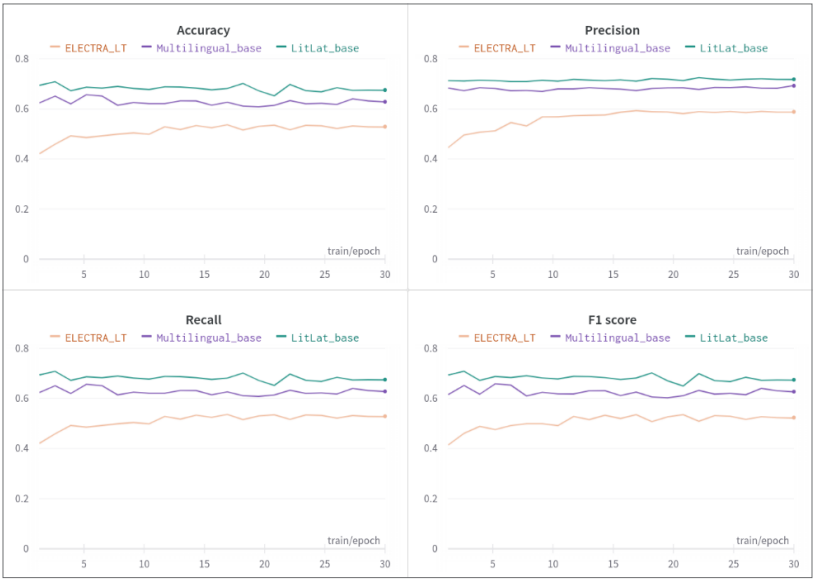
Visi trys įterpinių modeliai buvo papildomai apmokyti su anotuotu lietuviškų komentarų duomenų rinkiniu. Šį rinkinį sudarė 25 219 komentarai, suanotuoti į jau anksčiau minėtas tris klases:

1. neapykantos kalba (2082 komentarų);
2. neutrali kalba (15 316 komentarų);
3. įžeidi kalba (7821 komentarų).

Šie duomenys buvo paskirstyti į mokymo, validavimo ir testavimo duomenų aibes santykiu 0.6:0.2:0.2. Neapykantos ir įžeidžios kalbų klasių komentarai kiekviename iš duomenų rinkinių buvo pakartojami (dubliuojami) tam, kad sulygintume komentarų skaičių kiekvienoje klasėje. Kadangi sugeneruojami įterpiniai buvo 512 arba 128 skaičių ilgio vektoriai, tai visi komentarai, kurie buvo ilgesni nei 512 simbolių, buvo atmesti.

4 Rezultatai

Kiekvienas klasifikavimo modelis buvo mokomas po tris kartus, nustačius skirtingus atsitiktinių skaičių generavimo parametrus (angl. *seed*). Modelius apmokius buvo išsirenkamas didžiausius tikslumo įverčius turintis modelis ir būtent su šiuo modeliu atliekami tolimesni testavimai ir palyginimai. Modeliai buvo vertinami pagal keturias metrikas: tikslumą (angl. *accuracy*), preciziškumą (angl. *precision*), atkūrimą (angl. *recall*) ir F1 statistiką (angl. *F1-score*). Visų trijų modelių metrikų reikšmės apmokymo metu pateiktos 4 paveikslėlyje. Matome, kad geriausiai veikiantis modelis, naudojant validavimo imtį buvo



4 pav. Klasifikavimo vertinimo kreivės naudojant tikslumo (viršuje kairėje), preciziškumo (viršuje dešinėje), atkūrimo (apačioje kairėje) ir F1-įverčio (apačioje dešinėje) metrikas.

LitLat BERT, o prasčiausiai – *Electra*. Kalbant apie *Electra* modelį, galime daryti prielaidą, kad tokius rezultatus gavome todėl, kad *Electra* transformeris buvo apmokytas tik su 70 mln. lietuviškų žodžių, kai tuo metu *LitLat* BERT modelis – su 1,21 mlrd. 101 kalbos žodžiais. Taigi net jei *Electra* transformerio struktūra leidžia apmokyti modelį naudojant mažesnius kiekius duomenų, turimo duomenų kiekio vis tiek neužteko tiksliam modelio apmokymui.

Taip pat iš pateikto grafiko matosi, kad rezultatai didėjant epochų skaičiui beveik nekinta, tai parodo, kad modeliai greitai išmoksta tikslinių klasių savybes (pavyzdžiui, neapykantos kalba išsiskiria tam tikrais žodžiais, kaip „gėjus“, „žydas“, „rusas“ ar kt.), to pasekoje modelis greitai apsimoko klasifikuoti komentarus. Šią problemą išspręstų didesnis duomenų kiekis, kuriuose būtų daugiau skirtingų komentarų, skirtingoms klasėms (pavyzdžiui, neutralioj klasėj turėtų būti daugiau teigiamų komentarų apie gėjus ar kitų tautybių žmones, o neapykantos kalbos klasėje turėtų būti pateikiama bendrai daugiau skirtingų komentarų).

Toliau modeliai buvo testuojami panaudojant ir testinę duomenų aibę. Žemiau pateiktoje lentelėje galima matyti, kaip modelis geba atskirti kiekvieną klasę, remiantis prieš tai minėtomis metrikomis. Kiekvieno modelio ir kiekvienos klasės, įvertinimo rezultatai, panaudojant testinę duomenų aibę atvaizduojami žemiau pateiktoje lentelėje (žr. 1 lentelė). Matome, kad *LitLat* BERT modelis neapykantos kalbos klasę atskyrė geriausiai ir F1 statistika siekia 78 proc., o tikslumas – 72 proc.

1 lentelė. *LitLat* BERT, daugiakalbio BERT ir *Electra* modelių klasifikavimo įvertinimai, naudojant testinius duomenis.

Modelis	Klasė	Preciziškumas (%)	Atkūrimas (%)	F1 statistika (%)	Tikslumas (%)
LitLat BERT	neapykantos	80	76	78	72
	neutrali	69	74	72	
	įžeidi	69	67	68	
Daugiakalbis BERT	neapykantos	87	52	65	63
	neutrali	54	81	65	
	įžeidi	62	58	60	
Electra	neapykantos	83	40	54	55
	neutrali	49	65	56	
	įžeidi	51	61	56	

Atlikus modelių klasifikavimo tyrimą, galime matyti, kad geriausiai neapykantos kalbą bei kitas klases aptinkantis modelis yra *LitLat* BERT, todėl šį modelį galima panaudoti neapykantos kalbos aptikimo lietuviškuose žinių bei socialinių tinklų komentaruose prototipui kurti. Tačiau norint pagerinti modelio efektyvumą ir patikimumą, reikalingas didesnis kiekis ir įvairesnis komentarų turinys.

5 Išvados

1. Išanalizavus surinktą ir suanotuotą 25 219 komentarų turintį tekstyną, buvo pastebėta, kad neapykantos kalba pasireiškė rečiausiai ir užima 8,26 proc. viso duomenų rinkinio. Dažniausiai pasireiškė neutrali kalba (60,7 proc.).
2. Pritaikius modelius neapykantos kalbos atpažinimui lietuvių kalbai ir remiantis precizijos, atkuriamumo, tikslumo bei F1-įverčio metrikomis buvo išsiaiškinta, kad geriausiai veikia *LitLat* BERT modelis, kurio tikslumas siekia 72 proc., todėl šis modelis buvo pasirinktas neapykantos kalbos atpažinimo prototipo realizavimui.

Literatūra

[1] F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, and V. Patti, "Resources and benchmark corpora for hate speech detection: a systematic review," Lang Resour Eval, vol. 55, no. 2, pp. 477–523, 2021, doi: 10.1007/s10579-020-09502-8.

[2] P. Fortuna and S. Nunes, "A survey on automatic detection of hate speech in text," ACM Comput Surv, vol. 51, no. 4, 2018, doi: 10.1145/3232676.

[3] P. Fortuna, J. Soler, and L. Wanner, "Toxic, Hateful, Offensive or Abusive? What Are We Really Classifying? An Empirical Analysis of Hate Speech Datasets," in Proceedings of the Twelfth Language Resources and Evaluation Conference, Marseille, France: European Language Resources Association, May 2020, pp. 6786–6794. [Online]. Available: https://aclanthology.org/2020.lrec-1.838

[4] K. Madukwe, X. Gao, and B. Xue, "In Data We Trust: A Critical Analysis of Hate Speech Detection Datasets," in Proceedings of the Fourth Workshop on Online Abuse and Harms, Online: Association for Computational Linguistics, Nov. 2020, pp. 150–161. doi: 10.18653/v1/2020.alw-1.18.

[5] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated Hate Speech Detection and the Problem of Offensive Language," Proceedings of the International AAAI Conference on Web and Social Media, vol. 11, no. 1, pp. 512–515, May 2017, doi: 10.1609/icwsm.v11i1.14955.

[6] S. Mishra and S. Mishra, "3Idiots at HASOC 2019: Fine-tuning Transformer Neural Networks for Hate Speech Identification in Indo-European Languages.," in FIRE (Working Notes), 2019, pp. 208–213.

[7] A. Schmidt and M. Wiegand, "A Survey on Hate Speech Detection using Natural Language Processing," in Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 1–10. doi: 10.18653/v1/W17-1101.
[8] E. Mosca, M. Wich, and G. Groh, "Understanding and Interpreting the Impact of User Context in Hate Speech Detection," in Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media, Online: Association for Computational Linguistics, Jun. 2021, pp. 91–102. doi: 10.18653/v1/2021.socialnlp-1.8.
[9] Z. Waseem, T. Davidson, D. Warmsley, and I. Weber, "Understanding Abuse: A Typology of Abusive Language Detection Subtasks." 2017.
[10] A. Vaswani et al., "Attention is all you need," Adv Neural Inf Process Syst, vol. 30, 2017.
[11] J. Devlin, M.-W. Chang, K. Lee, K. T. Google, and A. I. Language, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." [Online]. Available: https://github.com/tensorflow/tensor2tensor
[12] S. Ruder, M. E. Peters, S. Swayamdipta, and T. Wolf, "Transfer Learning in Natural Language Processing," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials, Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 15–18. doi: 10.18653/v1/N19-5004.
[13] S. Lei, W. Yi, C. Ying, and W. Ruibin, "Review of attention mechanism in natural language processing," Data Analysis and Knowledge Discovery, vol. 4, no. 5, pp. 1–14, 2020.
[14] T. Pires, E. Schlinger, and D. Garrette, "How multilingual is multilingual BERT?," ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, pp. 4996–5001, 2020, doi: 10.18653/v1/p19-1493.
[15] Hugging Face, "bert-base-multilingual-cased." https://huggingface.co/bert-base-multilingual-cased (accessed Dec. 20, 2022).
[16] Y. Zhao and X. Tao, "ZYJ123@DravidianLangTech-EACL2021: Offensive Language Identification based on XLM-RoBERTa with DPCNN," Proceedings of the 1st Workshop on Speech and Language Technologies for Dravidian Languages, DravidianLangTech 2021 at 16th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2021, pp. 216–221, 2021.
[17] EMBEDDIA, "litlat-bert." https://huggingface.co/EMBEDDIA/litlat-bert (accessed Dec. 20, 2022).
[18] Hugging Face, "ELECTRA." https://huggingface.co/docs/transformers/model_doc/electra (accessed May 27, 2022).
[19] K. Clark, M.-T. Luong, G. Brain, Q. V Le Google Brain, and C. D. Manning, "ELECTRA: PRE-TRAINING TEXT ENCODERS AS DISCRIMINATORS RATHER THAN GENERATORS." [Online]. Available: https://github.com/google-research/