Primernost mer FPR in FNR kot kriterija pravičnosti v strojnem učenju*

Assessing the Suitability of FPR and FNR as Fairness Metrics in Machine Learning

Ana Farič†
Kognitivna znanost
Univerza v Ljubljani, Pedagoška fakulteta
Ljubljana, Slovenija
ana.faric@fri.uni-lj.si

Ivan Bratko
Univerza v Ljubljani, Fakulteta za računalništvo in informatiko
Ljubljana, Slovenija
ivan.bratko@fri.uni-lj.si

Povzetek

Sistemi strojnega učenja so v družbeno kritičnih domenah pogosto obtoženi pristranskosti, kar se navadno utemeljuje z asimetrijo napačne pozitivne (FPR) in napačne negativne mere (FNR) med demografskimi skupinami. To je bil tudi glavni argument o pristranskosti sistema COMPAS v znamenitem članku v ProPublici. Vendar v tem prispevku pokažemo, da so razlike v omenjenih merah neizogibna matematična posledica različnih osnovnih stopenj pojava in porazdelitve pomembnih atributov v učnih podatkih. Z analizo realne podatkovne zbirke ProPublica empirično kažemo, da lahko zgolj s spremembo demografske strukture asimetrijo napak izničimo ali celo popolnoma obrnemo. Z analizo vseh možnih uravnoteženih porazdelitev dodatno kažemo, da kriterij izenačenih možnosti, izražen z merami FPR in FNR, pogosto verjetno ni realen za praktično uporabo. Zaključujemo, da sta meri nezadosten in potencialno zavajajoč kazalnik, če ju ne interpretiramo v kontekstu osnovnih stopenj in porazdelitve podatkov.

Ključne besede

Strojno učenje, COMPAS, pristranskost, napačna negativna mera, napačna pozitivna mera

Abstract

Machine learning systems in socially critical domains are frequently accused of bias, which is typically justified by the asymmetry of the false positive (FPR) and false negative rate (FNR) across demographic groups. This was also the main argument against the COMPAS system in the famous paper in ProPublica. Our paper shows that disparities in these measures are an inevitable mathematical consequence of differing base rates and the distribution of important attribute values within the training data. Through the perturbation of the real ProPublica dataset, we empirically demonstrate that altering the demographic structure alone can eliminate or completely reverse the error asymmetry. Through the analysis of all possible balanced distributions, we further show that the fairness criterion expressed by the FPR and FNR measures is likely unrealistic for practical use. We conclude that the two measures are an insufficient and potentially misleading indicator if they are not interpreted in the context of base rates and the distribution of data.

Keywords

Machine learning, COMPAS, machine bias, false negative rate, false positive rate

1 Uvod

Ko sistemi strojnega učenja vstopajo v družbeno kritične domene, kot je pravosodje, postaja pristranskost pogost raziskovalni problem, ki ga zaznamujeta manjko univerzalne definicije [10] in nasprotujoči kriteriji pravičnosti [4][17]. V literaturi jasne definicije pogosto ali niso podane ali pa so pomanjkljivo opredeljene [14]. V splošnem lahko pristranskost razumemo kot razliko v obravnavi posameznikov ali skupin glede na neko lastnost, kot je spol ali rasa [20].

Prediktivno pravosodje (angl. *predictive justice*) vključuje modele za napovedovanje in podporo sodnim odločitvam [21], med katerimi so ključni instrumenti za oceno tveganja povratništva (angl. *risk assessment instruments*). Njihova naloga je identifikacija posameznikov, ki bodo najverjetneje ponovili kaznivo dejanje [15]. Pogosto temeljijo na veliki količini podatkov in na podlagi vrednosti različnih atributov osumljenca izračunajo verjetnost oz. oceno tveganja povratništva. Takšne napovedi se nato uporabljajo pri različnih kazenskih postopkih, od odločanja o varščini do izreka kazni [2].

COMPAS (*Correctional Offender Management Profiling for Alternative Sanctions*) je eden najbolj znanih instrumentov za oceno tveganja. V številnih zveznih državah ZDA, vključno z New Yorkom, Wisconsinom, Kalifornijo in Florido, se od leta 2000 uporablja za ocenjevanje tveganja povratništva [8][9]. Deluje na podlagi algoritma, ki analizira 137 atributov in poda oceno od 1 do 10, pri čemer višja ocena pomeni večjo verjetnost, da bo posameznik v roku dveh let ponovil kaznivo dejanje [5][18].

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
Information Society 2026, 5–9 October 2026, Ljubljana, Slovenia
© 2026 Copyright held by the owner/author(s).
http://doi.org/10.70314/is.2026.cogni.137

Leta 2016 je odmeven članek v ProPublici [1] sprožil razpravo o COMPAS-ovi rasni pristranskosti. Njihova analiza je temeljila na dveh merah uspešnosti modela:

Napačna pozitivna mera (angl. *False Positive Rate* – FPR) pomeni delež dejansko negativnih primerov, ki jim je model napačno napovedal pozitiven razred. Pozitiven razred je definiran kot povratništvo. Izračuna se tako, da število napačnih pozitivnih napovedi (angl. *False Positives* – FP) delimo s številom vseh resnično negativnih primerov, tj. vsoto FP in pravilnih negativnih napovedi (angl. *True Negatives* – TN):

$$FPR = \frac{FP}{FP + TN}$$

Napačna negativna mera (angl. *False Negative* Rate – FNR) pomeni delež dejansko pozitivnih primerov, ki jim je model napačno napovedal negativni razred. Izračuna se tako, da število napačnih negativnih napovedi (angl. *False Negatives* – FN) delimo s številom vseh resnično pozitivnih primerov, tj. vsoto FN in pravilnih pozitivnih napovedi (angl. *True Positives* – TP) [3]:

$$FNR = \frac{FN}{FN + TP}$$

Angwin idr. [1] so ugotovili, da je FPR za temnopolte posameznike skoraj dvakrat višji (44,9 %) kot za belce (23,5 %), medtem ko je FNR bistveno višji za belce (47,7 %) kot temnopolte (28 %). Uporabili so kriterij izenačenih možnosti (angl. *equalized odds*), ki zahteva, da imajo skupine enake vrednosti mer FPR in FNR. Od tedaj se različni vrednosti FNR in FPR med skupinama pogosto avtomatsko sprejme kot jasen dokaz pristranskosti.

Javne razprave [13] in veliko študij [6][8][9][16][19], ki se ukvarjajo s pristranskostjo, podpirajo zaključek Angwin idr. [1], da je COMPAS pristranski. Poleg tega mu upravičeno očitajo še netransparentnost (t. i. problem črne škatle, na katerega opozarja Rudin [22]) in slabo točnost, ki po ugotovitvah Dressela in Farida [8] ni boljša od ocen neekspertov.

Zanimivo je, da predhodne raziskave [4][7][8][12][17], ne glede na njihov zaključek o COMPAS-ovi pristranskosti, konsistentno reproducirajo zloglasno asimetrijo v vzorcu napak med skupinama, pogosto pa manjka globlji razmislek o tem, zakaj do asimetrije sploh pride. V nadaljevanju se posvečamo mehanizmom, ki ta vzorec generirajo, in raziskujemo, ali sta omenjeni meri primerna kazalnika pravičnosti modelov.

2 Sintetični scenarij

Pomemben del strojnega učenja je iskanje vzorcev v učnih podatkih in optimizacija točnosti. Zamislimo si ekstremno preprost scenarij. Imamo množico primerov, izmed katerih je 60 % pozitivnih in 40 % negativnih, model pa nima na voljo nobenih atributov. Model bo naredil edino smiselno potezo (alternativa je naključna klasifikacija) in vsem novim primerom napovedal večinski (pozitiven) razred. Neizogibna posledica bo ogromno število napačno pozitivnih primerov (FPR). Po kriteriju izenačenih možnosti je takšen model pristranski, čeprav je jasno, da gre v tem primeru zgolj za matematično optimizacijo napovedne točnosti.

Da bi izolirali mehanizme, ki prispevajo k asimetriji v merah FNR in FPR pri bolj kompleksnih modelih, smo ustvarili sintetičen primer. Znano je, da že zgolj razlika v osnovni stopnji pojava (npr. različna stopnja kriminalitete) poruši simetrijo mer FPR in FNR [4][7][17]. Vendar pa naslednji sintetični scenarij razkriva še en pomemben in pogosto premalo raziskan mehanizem porazdelitve atributov.

Ustvarimo podatkovno zbirko (N = 2400), kjer imamo dve enako veliki skupini (A: N = 1200 (rdeča barva) in B: N = 1200 (zelena barva)). Na voljo imamo dva atributa, ki se v pravni literaturi kažeta kot najbolj povedna za napovedovanje povratništva [22]. Gre za število predhodnih kaznivih dejanj in starost. Število predhodnih dejanj je tu obravnavano binarno: največ eno (0–1), dve ali več (2+). Starost obravnavamo kot ternarni atribut: starostna kategorija (mlajši, srednji, starejši). Pri tem sledimo pravilu, da imajo mlajši posamezniki z več predhodnimi dejanji večjo verjetnost za pozitivno napoved (+). Skupini imata enako stopnjo povratništva (600 povratnikov in 600 nepovratnikov; 50 %) in distribucijo predhodnih kaznivih dejanj (300 posameznikov iz vsake skupine ima 0–1, 300 pa več kot 2 kaznivi dejanji). Skupini se razlikujeta zgolj po starostni strukturi (skupina A je mlajša od skupine B in ima 200 primerov več v mlajši kategoriji). Model za vsako od šestih kombinacij atributov napove pozitiven (+) oziroma negativen (–) razred (Tabela 1).

Če za takšen model izračunamo mere uspešnosti, opazimo asimetrijo v FPR in FNR merah (Tabela 1). Ker je mladost pomemben dejavnik tveganja, model vsem primerom, ki spadajo v mlajšo starostno kategorijo, napove pozitiven razred. Ker je osnovna stopnja dejanskih povratnikov enaka, večje število pozitivnih napovedi neizogibno privede do večjega FPR za skupino A. Na drugi strani model starejši skupini B v primerjavi s skupino A pogosteje napove nepovratništvo, s čimer pogosteje spregleda dejanske povratnike, kar pomeni višji FNR.

V študijah s področja pristranskosti se takšna neizogibna asimetrija pogosto interpretira kot pristranskost modela, opisani sintetični primer pa kaže, da je vzorec napak neizogibna posledica porazdelitve učnih podatkov.

Tabela 1: Mere uspešnosti, poročane na splošno za vse primere in posebej za obe skupini

	splošno	A	B
točnost [%]	62,5	63,3	61,7
FPR [%]	37,5	45	30
FNR [%]	37,5	28,3	46,7

3 Eksperimenti s podatkovno zbirko ProPublica

Da bi ugotovitev preverili v praksi, smo na podatkovni zbirki *compas-scores-two-years*, ki so jo v svoji izvorni študiji uporabili Angwin idr. [1], izvedli serijo eksperimentov. Podatkovna zbirka vsebuje 7214 primerov iz okrožja Broward v Floridi in 53 atributov. Ker se ukvarjamo s skupinama temnopoltih in belcev, so iz analize izključeni pripadniki drugih

Tabela 2: Distribucija primerov (skupine A in B) glede na dva atributa (število predhodnih kaznivih dejanj in starostna kategorija)

		starostna skupina									
		mlajši			srednji			starejši			Σ
število predhodnih kaznivih dejanj	0-1	130 +	90 +	+	150 +	130 +	-	20 +	80 +	-	600
		70 -	10 -		150 -	170 -		80 -	120 -		600
	2+	130 +	80 +	+	170 +	150 +	+	0 +	70 +	-	600
		70 -	20 -		130 -	150 -		100 -	130 -		600
	Σ	400	200		600	600		200	400		

etničnih skupin (ostane 5278 primerov). Za učenje so bili uporabljeni atributi *age* (starost v letih), *age_cat* (kategorični atribut) in *priors_count* (število predhodnih kaznivih dejanj). Ciljna spremenljivka je *two_years_recid* (binarna: povratništvo oziroma nepovratništvo).

V teh podatkih so temnopolti obtoženci v povprečju mlajši (32 let) od belcev (38 let) in imajo več predhodnih dejanj (4,2 v primerjavi z 2,3). Zanimalo nas je, kako na FNR in FPR vpliva spreminjanje te demografske strukture.

Eksperimentalni postopek je zasnovan na metodologiji preteklih raziskav [8][11], kjer so bili za napovedovanje povratništva uporabljeni enostavnejši modeli (logistična regresija). Za filtriranje podatkov in učenje smo uporabili odprtokodno orodje *Orange Data Mining* (https://orangedatamining.com). Uporabljen je bil postopek naključnega vzorčenja, kjer je bilo 80 % podatkov namenjenih učenju, 20 % pa testiranju. Skozi 100 ponovitev je bilo izračunano povprečje točnosti, FNR in FPR za vse primere ter posebej za temnopolte in belce (Tabela 3).

Eksperiment 1: Naključno smo odstranili polovico vseh temnopoltih posameznikov. Rezultati so ostali praktično enaki izvorni študiji ProPublice, kar kaže, da v tem primeru sama velikost skupine ni poglavitni vzrok za asimetrijo napak.

Eksperiment 2: Iz podatkov smo naključno odstranili zgolj polovico temnopoltih povratnikov, s čimer smo umetno znižali osnovno stopnjo povratništva temnopoltih. Kljub temu se vzorec napak ni spremenil. Čeprav smo zmanjšali število dejanskih povratnikov, so preostali temnopolti v vzorcu ohranili tvegane atribute (nižja starost, več predhodnih kaznivih dejanj). Ob manjšem številu dejanskih povratnikov se je število napačno pozitivnih primerov (FP) celo nekoliko povečalo.

Eksperiment 3: Obdržali smo vse belce, pri temnopoltih pa smo odstranili izrazito tvegane posameznike in obdržali starejše od 20 let, ki imajo manj kot 3 predhodna dejanja. S tem smo FNR med skupinama praktično izenačili, razliko v FPR pa drastično zmanjšali.

Eksperiment 4: Demografsko sliko smo popolnoma obrnili. Ustvarili smo situacijo, kjer belci predstavljajo skupino z visokim tveganjem za povratništvo (obdržali smo samo mlajše od 40 let), temnopolti pa skupino z nizkim tveganjem (obdržali smo starejše od 25 let z manj kot 3 predhodnimi dejanji). Rezultat je popoln obrat trenda, kjer se model glede na kriterij FPR in FNR sedaj zdi pristranski do belcev.

Tabela 1: Mere uspešnosti za štiri eksperimente

Eksperiment 1			
	splošno [%]	temnopolti [%]	belci [%]
točnost	66,8 ± 1,2	66,4 ± 1,5	67,1 ± 1,4
FPR	32,8 ± 1,8	46,5 ± 2,1	24,7 ± 1,6
FNR	33,8 ± 1,7	21,9 ± 1,5	45,8 ± 2,2

Eksperiment 2			
	splošno [%]	temnopolti [%]	belci [%]
točnost	65,6 ± 1,3	63,7 ± 1,6	67,5 ± 1,3
FPR	34,2 ± 1,9	43,4 ± 2	23,2 ± 1,5
FNR	35 ± 1,8	23 ± 1,4	46,9 ± 2,3

Eksperiment 3			
	splošno [%]	temnopolti [%]	belci [%]
točnost	61,8 ± 1,4	59,1 ± 1,8	63,8 ± 1,5
FPR	41,7 ± 2	46,4 ± 2,2	38 ± 1,8
FNR	32,7 ± 1,6	31,7 ± 1,7	33,4 ± 1,9

Eksperiment 4			
	splošno [%]	temnopolti [%]	belci [%]
točnost	64,8 ± 1,5	67,9 ± 1,9	62,4 ± 1,6
FPR	28,5 ± 1,6	14,9 ± 1,2	41,6 ± 2,1
FNR	46 ± 2,2	73,8 ± 3,1	32,6 ± 1,8

Eksperimenti kažejo, da je zloglasna asimetrija v napakah zgolj refleksija asimetrične strukture v izvornih podatkih.

4 Analiza primernosti kriterijev FPR in FNR

Ugotovitvi, da je asimetrija v merah FNR in FPR povezana s porazdelitvijo podatkov, sledi vprašanje, če je sploh realistično pričakovati, da bosta meri v praksi kdaj izenačeni. To smo raziskali z eksperimentom, pri katerem smo sistematično generirali vse primerne porazdelitve primerov za dve skupini.

Vse generirane porazdelitve so bile uravnotežene tako, da je bilo v vsaki skupini enako število primerov in enako število pozitivnih in negativnih izidov. Ustvarili smo domeno z enim binarnim atributom (*age* z možnima vrednostima *younger* in *older*), pri čemer je veljalo pravilo, da je, če je oseba mlajša, napoved pozitvna. To ustreza običajnemu klasifikacijskemu pravilu v domeni COMPAS-a.

Rezultati kažejo, da je idealna nepristranskost (kjer sta FPR in FNR med skupinama povsem enaka) praktično neobstoječa. Tudi v primeru, ko smo kriterij pravičnosti zrahljali in določili »dovolj nepristransko« porazdelitev (kjer vsota vseh

napak v obeh skupinah ne preseže 20 %). Pri majhnem vzorcu (N = 10 primerov, 5 v vsaki skupini) obstaja 76 možnih uravnoteženih porazdelitev. Izmed njih so zrahljanemu pogoju zadostne nepristranskosti ustrezale zgolj 4 porazdelitve (5,3 %). Z večanjem števila primerov ta delež hitro pada (Tabela 4).

V realnih podatkovnih zbirkah (kot je zbirka ProPublice), kjer imamo opravka z na tisoče primeri, je verjetnost, da bi naravna porazdelitev ustrezala kriterijem izenačenih mer, praktično enaka nič. To nakazuje, da kriterij pravičnosti, izražen z merami FNR in FPR, verjetno ni realen za praktično uporabo.

Tabela 4: Prikaz števila vseh uravnoteženih porazdelitev in deleža »dovolj pravičnih« porazdelitev glede na skupno število primerov (N)

Skupno št. primerov (N)	Št. vseh porazdelitev	Št. »dovolj pravičnih« in odstotek od vseh
10	76	4 (5,3 %)
20	1.479	29 (1,96 %)
30	8.604	78 (0,91 %)
40	32.777	169 (0,52 %)
50	91.888	338 (0,37 %)
60	219.200	611 (0,28 %)

5 Zaključek

Pristranskost v strojnem učenju ostaja močno polemična tema, predvsem zaradi pomanjkanja univerzalnih definicij in uporabe medsebojno nezdružljivih kriterijev pravičnosti. Naš prispevek združuje sintetične primere, empirično spreminjanje realnih podatkov ProPublice in matematični argument, ki skupaj nakazujejo, da asimetrija v merah FPR in FNR sama po sebi ni zadosten indikator pristranskosti modelov strojnega učenja, in da jo je treba interpretirati v kontekstu osnovnih stopenj in porazdelitve podatkov.

Razlike, ki so sprožile razpravo o COMPAS-u, so v resnici neizogibna matematična posledica različnih osnovnih stopenj in asimetrične porazdelitve pomembnih atributov (kot sta starost in kriminalna preteklost) v izvorni populaciji. Naši rezultati nakazujejo, da je slepo zanašanje na kriterij izenačenih možnosti za ocenjevanje pravičnosti modelov v praksi nerealistično in potencialno zavajajoče. Prihodnje študije se morajo zato odmakniti od naivnega iskanja izenačenosti teh specifičnih mer ter se posvetiti transparentnosti uporabljenih atributov in globljemu razumevanju mehanizmov strojnega učenja ob soočanju z realnimi, naravno neuravnoteženimi podatki.

Literatura

[1] Angwin, J., Kirchner, L., Larson, J. in Mattu, S. 2016. Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica*.

[2] Bansak, K. 2019. Can nonexperts really emulate statistical learning methods? A comment on »The accuracy, fairness, and limits of predicting recidivism«. *Political Analysis, 27*(3), 370–380. DOI: https://doi.org/10.1017/pan.2018.55.

[3] Bradley, A. P. 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition, 30*(7), 1145–1159. DOI: https://doi.org/10.1016/S0031-3203(96)00142-2.

[4] Chouldechova, A. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data, 5*(2), 153–163. DOI: https://doi.org/10.1089/big.2016.0047.

[5] Corbett-Davies, S., Pierson, E., Feller, A. in Goel, S. 2016. A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear. *The Washington Post*.

[6] Crespo, V., Rando, J., Eva, B., Keswani, V. in Sinnott-Armstrong, W. 2023. Reevaluating COMPAS: Base Rate Tracking and Racial Bias. 37th *NeurIPS Conference*.

[7] Dieterich, W., Mendoza, S. in Brennan, T. 2016. COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity.

[8] Dressel, J. in Farid, H. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science Advances, 4*(1). DOI: https://doi.org/10.1126/sciadv.aao5580.

[9] Engel, C., Linhardt, L. in Schubrt, M. 2024. Code is law: how COMPAS affects the way the judiciary handles the risk of recidivism. *Artificial Intelligence and Law, 22*, 873–874. DOI: https://doi.org/10.1007/s105060-024-09400-2.

[10] Farič, A. in Bratko, I. 2024. Machine Bias: A Survey of Issues. *Informatica, 48*(2). DOI: https://doi.org/10.31449/inf.v48i2.5971.

[11] Farič, A. in Bratko, I. 2025. Machine bias: new experiments with COMPAS data. *Conference Information Society*. DOI: https://doi.org

[12] Flores, A. W., Bechtel, K. in Lowenkamp, C. T. 2016. False Positives, False Negatives, and False Analyses: A Rejoinder to »Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks. *Federal Probation, 80*(2).

[13] Hamilton, M. 2019. The sexist algorithm. *Behavioral Sciences & the Law, 37*(2). DOI: https://10.1002/bsl.2406.

[14] Hellström, T., Dignum, V. in Bensch, S. 2020. Bias in Machine Learning – What is it Good for? *arXiv preprint:2004.00686*. DOI: https://doi.org/10.48550/arXiv.2004.00686.

[15] Johnson, J. L., Lowenkamp, C. T., VanBenschoten, S. W. in Robinson, C. R. 2011. The Construction and Validation of the Federal Post Conviction Risk Assessment (PCRA). *Federal Probation, 75*(5).

[16] Khademi, A. in Honavar, V. 2019. Algorithmic Bias in Recidivism Prediction: A Causal Perspective. *arXiv preprint:1911.10640*. DOI: https://doi.org/10.48550/arXiv.1911.10640.

[17] Kleinberg, J., Mullainathan, S. in Raghavan, M. 2021. Inherent Trade-Offs in the Fair Determination of Risk Scores. *arXiv preprint:1609.05807*. DOI: https://doi.org/10.48550/arXiv.1609.05807.

[18] Lippert-Rasmussen, K. 2025. Algorithmic and Non-Algorithmic Fairness: Should We Revise Our View of the Latter Given Our View of the Former? *Law and Philosophy, 44*(2), 155–179. DOI: https://doi-org.nukweb.nuk.uni-lj.si/10.1007/s10982-024-09505-4.

[19] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. in Galstyan, A. 2019. A Survey on Bias and Fairness in Machine Learning. *arXiv preprint:1908.09635*. DOI: https://doi.org/10.48550/arXiv.1908.09635.

[20] Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M. E., …, Staab, S. 2020. Bias in data-driven artificial intelligence systems – An introductory survey. *WIREs Data Mining and Knowledge Discovery, 10*(3). DOI: https://doi.org/10.1002/wodm.1356.

[21] Porebski, A. 2023. Machine Learning and law. *Research Handbook on Law and Technology*. Cheltenham, UK: Edward Elgar Publishing. DOI: https://doi.org/10.4337/9781803921327.00037.

[22] Rudin, C. 2018. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *arXiv preprint:1811.10154*. DOI: https://doi.org/10.48550/arXiv.1811.10154.