Etički izazovi u doba veštačke inteligencije: Pristrasnost pod lupom

Seminarski rad u okviru kursa
Metodologija stručnog i naučnog rada
Matematički fakultet

Mirko Kordić, Dragana Zdravković, Marija Marković, Marko Savić
mirko22kordic@gmail.com, dragana.zdravkovic602@gmail.com,
marijaa.markovic15@gmail.com, savicmarko033@gmail.com

17. decembar 2023.

Sažetak

Veštačka inteligencija čija primena raste iz dana u dan, uključena u svakodnevni život većine pojedinaca, kao i svaki sistem napravljen od strane čoveka ima primese ljudskog razmišljanja. Brojne predrasude i pristrasnosti su sastavni deo svakog ljudskog bića, te su se ove karakteristike oslikale i u sisteme veštačke inteligencije. Da li su sistemi za prepoznavanje lica rasisti, zbog čega nam se na društvenim mrežama plasiraju lažne vesti, kao i to kako se politički stereotipi oslikavaju u odgovorima popularnih jezičkih modela, samo su neka od pitanja na koja ćemo odgovoriti u ovom radu.

Sadržaj

1 Uvod

Razvoj veštačke inteligencije i njena prisutnost u svakodnevnom životu temelj su brojnim inovacijama i poboljšanjima u različitim sferama. Zajedno sa velikim napretkom javlja se i tema pristrasnosti veštačke inteligencije, čiji uticaj može imati ozbiljne posledice na društvo.

Pristrasnost veštačke inteligencije zastupljena je u mnogim sferama današnjeg društva, a u ovom radu pažnja je usmerena ka analiziranju različitih problema (pristrasnost u okviru društvenih mreža, sistema za prepoznavanje lica i jezičkih modela) i predstavljanje relevantnih rezultata iz literature.

Prvi posmatrani problem jeste tačnost sistema za prepoznavanje lica kompanija poput Microsoft-a, IBM-a i Face++, kao deo The Gender Shades projekta [8], čija analiza ukazuje na značajne razlike u stopama grešaka između različitih grupa, uključujući rod i boju kože. Neka istraživanja sugerišu da prilagođavanje modela za svaku podgrupu može smanjiti pristrasnost i poboljšati ukupnu tačnost prepoznavanja lica [27].

Drugi obrađeni problem pristrasnosti jesu društvene mreže koje su sa milijardama korisnika širom sveta, postale ključne za komunikaciju i pristup informacijama. Javljaju se pitanja etike vezana za sisteme preporuke i stvaranje "filter mehura" [24], na osnovu koga se korisnicima prikazuje sadržaj. Pored toga, izazovi u otkrivanju i suzbijanju lažnih vesti na društvenim mrežama postaju sve ozbiljniji, s obzirom na pristrasnost u treniranju algoritama i njihovu tendenciju da favorizuju određene izvore informacija.

Poslednji analizirani problem jeste pristrasnost u jezičkim modelima, poput GPT-3, BERT-a i ROBERT-e, koji su obučeni za razumevanje i generisanje prirodnog jezika korišćenjem tehnika dubokog učenja. Analiza pokazuje da ovi modeli često manifestuju političke stereotipe, a pristrasnost modela povezuje se sa skupovima podataka korišćenim za njihovo treniranje. Ističe se važnost odgovornog pristupa kompanija u obezbeđivanju objektivnosti u radu jezičkih modela, s obzirom na značajan uticaj koji imaju na društvena i ekonomska pitanja.

Nedostatak objektivnosti i jednakog tretmana kroz različite sektore naglašava potrebu za daljim istraživanjem i inovacijama.

2 Sistemi za prepoznavanje lica

Sistemi za prepoznavanje lica identifikuju ljudska lica na slikama ili video snimcima, sa namerom da utvrde da li lice na dve slike pripada istoj osobi ili da potraže lice među velikom kolekcijom postojećih slika [1]. Kada veštačka inteligencija obučava sisteme za prepoznavanje lica na podacima koji ne odražavaju stvarni svet ili sadrže inherentne pristrasnosti, to može rezultovati netačnim identifikacijama. Ovaj problem je naročito vidljiv prilikom posmtranja različitih društvenih grupa, uključujući etničke i rodne kategorije.

2.1 Pristrasni sistemi u upotrebi

Navedeni problem prikazuje i priča koja govori o nepravedno osuđenom Robertu Vilijamsu [28]. U januaru 2020. godine, afroamerikanac Robert Vilijams uhapšen je ispred svog doma pod sumnjom da je izvršio pljačku u prodavnici satova. Detektivi su pokazali Robertu sliku drugog afroamerikanca sa sigurnosnih kamera iz prodavnice, gde je odmah postalo jasno da Robert nije tražena osoba. Naime, slika sa sigurnosne kamere data je sistemu za prepoznavanje lica Državne policije Mičigena, koji je kao rezultat izbacio staru Robertovu sliku sa vozačke dozvole.

Ovo je samo jedan od mnogih primera koji ilustruju kako pristrasnost u podacima na kojima se treniraju modeli veštačke inteligencije (u smislu veće prisutnosti podataka o favorizovanim grupama) implikuje njihovo pogrešno rasuđivanje. Time se otvara pitanje : Kako konstruisati algoritame koji imaju visoku tačnost za identifikaciju pojedinaca u različitim rasnim, etničkim, polnim i starosnim grupama?

2.2 Analiza pristrasnosti

Pre nego što se posvetimo odgovaranju na postavljeno pitanje, pogledaćemo kako se ponašaju sistemi za prepoznavanje lica poznatih kompanija. *The Gender Shades Project* [8] je studija Masačusetskog instituta za tehnologiju (MIT) koja evaluira tačnost sistema zasnovanih na veštačkoj inteligenciji koji vrše rodnu klasifikaciju. Projekat je analizirao sisteme razvijene od strane kompanija Microsoft, IBM i kineske kompanije Face++. Za testiranje sistema izabrano je 1270 slika, na kojima se nalaze osobe iz tri afričke i tri evropske zemlje. Nakon toga, osobe sa slika su grupisane prema polu, tipu kože i kombinaciji ove dve karakteristike.

U tabeli 1 možemo videti tačnost navedenih klasifikatora u različitim grupama osoba. Kada se posmatra ceo skup podataka, Microsoft je pokazao najbolju tačnost klasifikacije, dok je IBM imao najlošije rezultate. Iako se čini da kompanije imaju relativno visoku ukupnu tačnost, primećene su značajne razlike u stopama grešaka između različitih grupa. Sve kompanije pokazuju bolje rezultate nad muškim ispitanicima nego nad ispitanicima ženskog pola. Microsoft je imao 8.1% razlike u grešci u zavisnosti od pola, IBM 14.7%, dok je Face++ pravio najveću razliku od 20.6%. Analizirajući podelu na osnovu boje kože primećena je ponovo značajna razlika u greškama. Face++ sa 11.8%, Microsoft sa 12.2% i IBM sa 19.2%, gde je svaka od kompanija pokazala bolje rezultate nad osobama svetlije boje kože.

Tabela 1: Tačnost klasifikacije u različitim kategorijama

Klasifikator	Ispitanici				
	Svi	Muški	Ženski	Svetloputi	Tamnoputi
Microsoft	**93.7%**	97.4%	**89.3%**	**99.3%**	**87.1%**
Face++	90.0%	**99.3%**	78.7%	95.3%	83.5%
IBM	87.9%	94.4%	79.7%	96.8%	77.6%

Posmatrajući kombinaciju pola i boje kože primećeno je da sve kompanije daju najgore rezultate nad tamnoputim osobama ženskog pola, što se može videti u tabeli 2. IBM i Microsoft sistemi se najbolje ponašaju nad svetlim muškarcima, dok je Face++ pokazao najbolje rezultate nad tamnoputim muškarcima.

Tabela 2: Tačnost klasifikacije pri kombinaciji pola i boje kože

Klasifikator	Tamnoputi muškarci	Tamnopute žene	Svetloputi muškarci	Svetlopute žene	Najveća razlika
Microsoft	94.0%	**79.2%**	**100%**	**98.3%**	**20.8%**
Face++	**99.3%**	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

2.3 Skupljanje podataka za izgradnju fer sistema

Na osnovu prikazanih rezultata možemo primetiti da konvencionalni pristup učenju, koji podrazumeva da se model trenira nad celim skupom podataka, rezultuje velikim jazom u kvalitetu klasifikacije između različitih podgrupa. Ovde ćemo pokušati da odgovorimo na pitanje postavljeno u sekciji 2.1: Kako konstruisati algoritme koji imaju visoku tačnost za identifikaciju pojedinaca u različitim rasnim, etničkim, polnim i starosnim grupama?

U radu *Face Recognition: Too Bias, or Not Too Bias?* [27] pokazano je da učenjem specifičnosti svake podgrupe možemo da smanjimo razlike u kvalitetu, i takođe značajno unapredimo celokupnu tačnost klasifikacije. Korišćen je skup podataka *Balanced Faces In the Wild* [26], sačinjen od 8 podgrupa zasnovanih na rodnoj i etničkoj pripadnosti. Analiziranjem rezultata prikazanih u ovom radu dolazimo do zaključka da je formiranje podgrupa veoma značajno jer algoritam za prepoznavanje lica retko pravi greške u okviru određene podgrupe. Kada je ustanovljeno da je sistem pristrasan, primećeno je da korišćenjem istih parametara modela nad različitim podgrupama dolazi do značajnih razlika u kvalitetu primene algoritma, te su izražajno bolji rezultati dobijeni kada su parametri određivani zasebno za svaku podgrupu.

Široka rasprostranjenost sistema za prepoznavanje lica, indikuje da se pristrasnost u njihovom funkcionisanju mora redukovati zbog posledica koje mogu biti veoma ozbiljne. Vreme posvećeno prikupljanju i analizi podataka korišćenim za treniranje pomenutih sistema treba da odgovara kompleksnosti problema koji se obrađuje. Da li će sistemi za prepoznavanje lica ikada zadobiti potpuno poverenje kritičara pokazaće nam vreme.

3 Društvene mreže

Prema istraživanju kompanije *Global Web Index* [9] prosečan čovek provede dnevno oko 2 sata i 30 minuta na društvenim mrežama. Sa milijardama korisnika širom sveta, ove platforme postale su jedan od glavnih načina komunikacije i izvora informacija. Sa pojavom veštačke inteligencije i njenom primenom u algoritmima pretrage dolazi do revolucije društvenih mreža. Ovi

algoritmi u osnovi određuju sadržaj koji vidimo na društvenim mrežama, kao što su personalizovani oglasi i predložene objave. Iako to u velikoj meri poboljšava iskustvo korišćenja društvenih mreža, postavljaju se mnoga etička pitanja, kao što su: Kako se zapravo određuje sadržaj koji nam se prikazuje? U kojoj meri taj sadržaj utiče na formiranje našeg mišljenja? Da li su algoritmi dovoljno pametni da prepoznaju lažne vesti?

3.1 Sistemi za preporuku sadržaja

Danas je na internetu broj informacija ogroman, tako da ih je neophodno filtrirati kako bi se svakom korisniku poboljšalo iskustvo korišćenja prikazivanjem sadržaja koji je za njega relevantan. Sistemi za preporuku rešavaju ovaj problem tako što pretražuju mnoštvo dinamičkih informacija, generisanih na osnovu ponašanja korisnika na internetu i na taj način pružaju personalizovano iskustvo [12].

"Veverica koja umire ispred tvoje kuće može trenutno da bude relevantnija za tvoje interese nego ljudi koji umiru u Africi."je na osnovu mnogih izvora temelj na kom je Mark Zakerberg zasnovao Fejsbukov *News Feed* [13]. Iako je neosporno da sistemi za preporuku poboljšavaju iskustvo korišćenja socijalnih mreža, postavlja se pitanje kakav je njihov uticaj na diverzitet sadržaja koji nam se prikazuje?

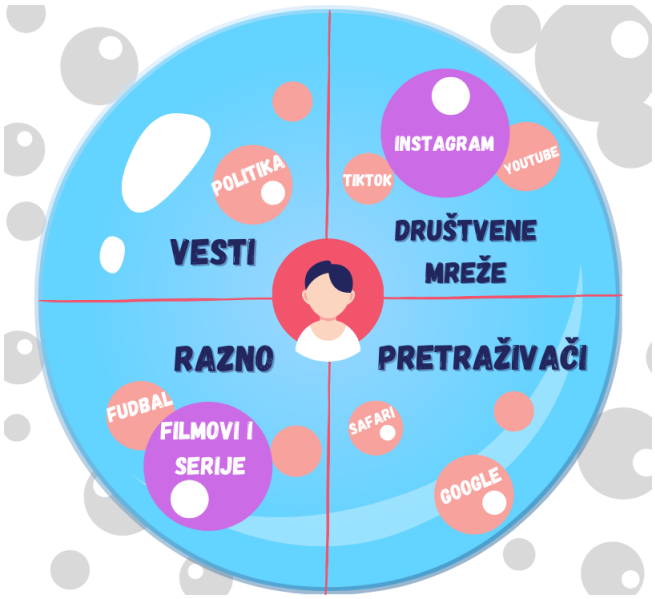
3.1.1 Filter mehur

Eli Pariser, autor i aktivista, u svojoj knjizi *"The Filter Bubble: What the Internet Is Hiding from You"* [24] definiše pojam filter mehur (eng. *filter bubble*) kao personalizovano onlajn okruženje koje nastaje uz pomoć algoritama pretrage na osnovu korisnikovih prethodnih aktivnosti, preferenci i interesovanja.

Nalaziti se unutar filter mehura znači da se sličan sadržaj i stavovi prikazuju sa većom frekvencijom, stvarajući opasnost da do korisnika neke informacije ne dođu ili da dođu samo iz perspektive za koju algoritmi misle da je odgovarajuća za korisnika. "Filter mehur je tvoj lični univerzum informacija u kome živiš kada si onlajn... Stvar je u tome da ti nemaš puno kontrole šta u njega ulazi"[23]. U prevodu, algoritmi (veštačka inteligencija) razvijaju pristrasnost prema "slici" koju su stvorili o korisniku, prikazujući mu sadržaj za koji smatraju da je blizak njemu, mahom iste tematike ili sličnog pogleda na određenu temu kao i objave koje je pročitao, podelio ili označio da mu se sviđaju u prošlosti. Na taj način ga potencijalno uskraćuju mnoštva drugih tema prema kojima ranije nije prikazao afinitet ili neke skroz druge perspektive na njemu bliske teme. Ilustracija filter mehura može se videti na slici 1. Prema Pariseru, dodatan problem je što je ovaj fenomen teško primetiti, jer algoritmi rade u pozadini i sam korisnik često nije ni svestan da oni postoje.

Sam termin *filter mehur* dosta je podelio samu javnost, kao i naučnu zajednicu: jedni negiraju njegovo postojanje i smatraju da je cela situacija oko njega preuveličavanje, dok drugi nude rešenja kako da se "iz njega izađe". U

radu *"Burst of the Filter Bubble"*[15], autori tvrde da nisu pronašli ništa što čvrsto pokazuje postajanje filter mehura kod Gugl Vesti (eng. *Google News*), ali da su otkrili da Gugl Vesti favorizuju određene medijske izvore, dok druge prikazuju sa manjom frekvencijom. Navode da su sakupili dovoljno dokaza da mogu zabrinutost oko filter mehura u kontekstu onlajn vesti da nazovu preuveličanom. Sa druge strane, u radu *"Breaking the filter bubble: democracy and design"*[7], autori vide filter mehur kao veliki problem demokratskog društva i analiziraju različite alate za borbu protiv filter mehura.



Slika 1: Filter mehur

Bez obzira da li je filter mehur stvaran u meri kojoj ga je Eli Pariser definisao ili je samo jedna od teorija zavere, sama ideja iza njega je poprilično intuitivna i zabrinjavajuća da bi se ignorisala. Potpuno druga krajnost bio bi filter mehur koji je sačinjen od mnoštva lažnih vesti. Bil Gejts, u intervjuu sa Kvarcom[25], dotakao se ove teme pitavši se "Da li ljudi stvarno žele da budu u mikrosistemu gde su činjenice lažne? Jer kako vreme prolazi, laži ne vode ka dobrim stvarima."

3.2 Lažne vesti

Nakon poraza na izborima u Americi 2016. godine, Hilari Klinton je nazvala širenje lažnih vesti na društvenim mrežama epidemijom[16]. Kako sve više ljudi informacije dobija sa društvenih mreža, gde su upravo lažne vesti široko rasprostranjene, postavlja se pitanje koliko su algoritmi uspešni u otkrivanju i suzbijanju širenja istih?

U radu *"Generalizing to the Future: Mitigating Entity Bias in Fake News Detection"*[29] autori predstavljaju problem koji postoji kod ovih algoritama. Naime, većina ovih algoritama trenira se na skupu starih vesti kako bi otkrili buduće lažne vesti i na ovaj način oni razvijaju nenamernu pristrasnost prema entitetu koji se pominje u određenim vestima. U njihovim podacima, 97%

članaka u kojima se pominje Donald Tramp iz perioda 2010-2017 su istiniti, dok je samo 33% za istu osobu u 2018. godini istinito. Algoritam je većinu članaka klasifikovao kao istinite baš zato što je razvio pristrasnost prema entitetu. Istraživači ovaj problem rešavaju ukidanjem veze između entiteta i istinitosti članka, što je dalo mnogo bolje rezultate.

Bez obzira na kvalitet sistema za preporuku i sposobnosti otkrivanja lažnih vesti, teško je verovati i očekivati da će sve lažne vesti biti iskorenjene na bilo kojoj socijalnoj mreži ili da će algoritmi preporuke prikazati sve moguće teme i poglede na njih. Zato su obrazovanje, kritičko razmišljanje i želja za prepoznavanjem istine veoma važni za svakoga ko prikuplja informacije na Internetu.

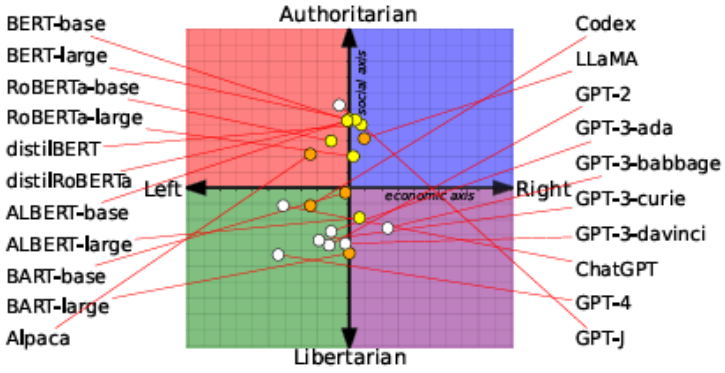
4 Jezički modeli

Jezički modeli predstavljaju vrstu softvera obučenog da razume i oponaša ljudsku komunikaciju. Oni uglavnom koriste tehnike dubokog učenja (eng. *deep learning*) [11] kako bi obradili ogromne količine podataka i stekli sposobnost generisanja teksta na prirodnom jeziku. Neki od najpoznatijih jezičkih modela današnjice jesu: GPT-3 (razvijen od strane američke istraživačke laboratorije OpenAI) [2], BERT i RoBERTa (jezički modeli kompanije Google) [14][5], LLaMA (proizvod kompanije Meta) [18] i drugi. Podaci na kojima se ovi modeli treniraju potiču iz različitih knjiga, članaka, tekstova koji prirodno sadrže određeni nivo predrasuda i stereotipa, pa se postavlja pitanje: koliko tih predrasuda model "nauči" i počne da primenjuje u svom ponašanju?

4.1 Politički stereotipi

Politički stereotipi u jezičkim modelima postaju sve prisutniji, pa se analizom jasno može uvideti da su neki od njih trenirani na podacima sa izraženom favorizacijom specifičnih političkih uverenja. Neka od gorućih pitanja koja dele društvo odnose se na smrtnu kaznu, abortus, istopolne brakove, feminizam, govor mržnje itd. Iako postoje stotine studija koje se bave prisutnošću i uticajem ovih tema na rad jezičkih modela, u nastavku ćemo predstaviti rezultate jedne od njih [10]. Autori ovog istraživanja ispitivali su kako 14 različitih jezičkih modela odgovara na različita socijalna i ekonomska pitanja, a rezultati su predstavljeni njihovim pozicioniranjem na političkom kompasu [4] - slika 2.

Otkriveno je da ranije pomenuti BERT modeli često pokazuju konzervativne stavove, za razliku od GPT modela koji su značajno liberalniji u svojim odgovorima. Autori kao uzrok navode činjenicu da su se za treniranje BERT modela prvenstveno koristile knjige sa konzervativnijim tonom, dok se danas za treniranje jezičkih modela kao što su GPT modeli koriste internet stranice, na kojima su u velikoj meri prisutni liberalni stavovi. Štaviše, primećuju se značajne razlike u modelima koji potiču iz iste porodice ali imaju različite veličine (ALBERT i BART). Kao razlog navodi se bolja generalizacija kod većih modela, čime se pristrasnost uklapa u suptilnije kontekste.



Slika 2: Pozicija različitih jezičkih modela na političkom kompasu

Istraživači su otišli i korak dalje, trenirajući modele sa leve strane političkog kompasa na podacima koji su levo orijentisani i slično, modele sa desne strane političkog kompasa na desno orijentisanim skupovima podataka. Ispostavilo se da su se već prisutne predrasude i stereotipi dodatno istakli u ponašanju modela. U nastavku, provereno je kako ovi modeli prepoznaju govor mržnje koji se odnosi na različite društvene grupe, ali i da li dobro uočavaju dezinformacije u levo (odnosno desno) orijentisanim časopisima. Rezultate možemo videti na tabelama 3 i 4 u nastavku.

Tabela 3: Prepoznavanje govora mržnje

Hate Speech	BLACK	MUSLIM	LGBTQ+	JEWS	ASIAN	LATINX	WOMEN	CHRISTIAN	MEN	WHITE
NEWS_LEFT	89.93	89.98	90.19	89.95	91.55	91.28	86.81	87.82	85.63	86.22
REDDIT_LEFT	89.84	89.90	89.96	89.50	90.66	91.15	87.42	87.65	86.20	85.13
NEWS_RIGHT	88.81	88.68	88.91	89.74	90.62	89.97	86.44	89.62	86.93	86.35
REDDIT_RIGHT	88.03	89.26	88.43	89.00	89.72	89.31	86.03	87.65	83.69	86.86

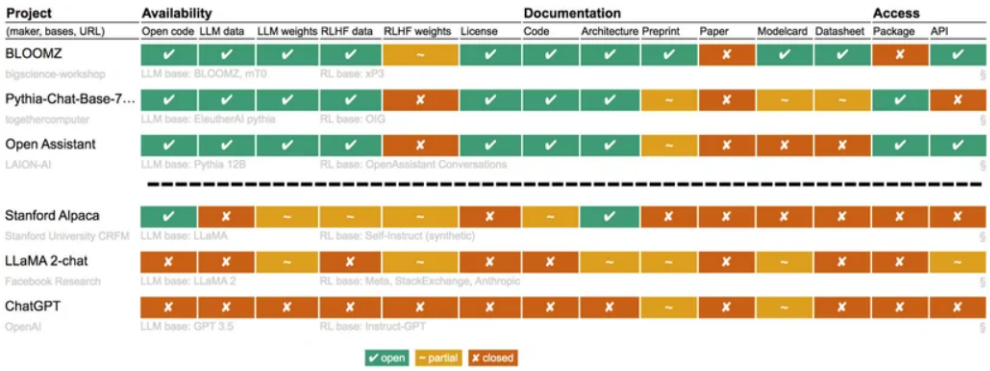
Tabela 4: Prepoznavanje pogrešnih informacija

Misinformation	Huffington Post (L)	New York Times (L)	CNN (L)	National Public Radio (L)	Guardian (L)	Fox (R)	Washington Examiner (R)	Breitbart (R)	Washington Times (R)	National Review (R)
NEWS_LEFT	89.44	86.08	87.57	89.61	82.22	93.10	92.86	91.30	82.35	96.30
REDDIT_LEFT	88.73	83.54	84.86	92.21	84.44	89.66	96.43	80.43	91.18	96.30
NEWS_RIGHT	89.44	86.71	89.19	90.91	86.67	88.51	85.71	89.13	82.35	92.59
REDDIT_RIGHT	90.85	86.71	90.81	84.42	84.44	91.95	96.43	84.78	85.29	96.30

Vrste ove dve tabele odgovaraju skupovima podataka na kojima su modeli trenirani - imamo dva levo i dva desno orijentisana skupa. Takođe važi: što je boja ćelije tamnija to je prepoznavanje govora mržnje, odnosno pogrešnih informacija bolje. Analizom tabele uočavamo da su modeli trenirani na levo orijentisanim skupovima osetljiviji na govor mržnje usmeren ka ženama, crncima, Jevrejima, pripadnicima LGBT+ populacije, dok je druga grupa modela osetljivija na govor mržnje upućen belim muškarcima i hrišćanima. Takođe, levo orijentisani modeli lakše prepoznaju dezinformacije u konzervativnim nego u liberalnim časopisima, dok je kod desno orijentisanih modela situacija obrnuta.

Za poslednju fazu istraživanja, čiji su rezultati predstavljeni tabelama 3 i 4 važno je naglasiti da su korišćene starije verzije jezičkih modela (GPT-2 i

RoBERTa), pa rezultate ipak treba uzeti sa određenom dozom rezerve. Ipak, ovo je trenutno i najbolja mogućnost koju istraživači imaju, kako su detalji aktuelnih jezičkih modela uglavnom zatvoreni za javnost. OpenAI ovo opravdava činjenicom da su im prihodi koji se obezbeđuju komercijalizacijom ovih projekata neophodni za finansiranje daljih istraživanja. Meta je predstavila jezički model Llama 2 koji se može preuzeti na [17], mada kritičari dovode u pitanje koliko se ovo zaista može smatrati otvorenim kodom s obzirom na činjenicu da kod korišćen za treniranje nije javno dostupan [21]. Detaljan pregled javnih i zatvorenih delova poznatih jezičkih modela možete pogledati na slici 3 preuzetoj sa [21].



Slika 3: Otvorenost različitih jezičkih modela

Istraživanje koje smo izložili u ovom radu nam je na realnim primerima pokazalo kako se pristrasnost u podacima preslikava na rad samih jezičkih modela. Uvideli smo da je izbor skupova podataka za treniranje modela ozbiljan i veoma važan posao, kome treba pristupiti iz više različitih uglova, kako bi se istinski uvideli svi pravci u kojima predrasude i stereotipi mogu preovladati. Međutim, kako su ljudi tvorci podataka koji se koriste, moramo računati na greške i pristrasnost izazvane njihovim stavovima i okruženjem koje na te stavove utiče.

Na osnovu svega navedenog, dolazimo do zaključka da skupovi podataka koji se koriste za treniranje jezičkih modela značajno utiču na njihovo rasuđivanje. Zato vodeće kompanije na tom tržištu imaju veliku društvenu odgovornost u obezbeđivanju što objektivnijeg stava njihovih modela prema važnim socijalnim i ekonomskim pitanjima. Neke od njih pretrpele su velike pritiske javnosti zbog pristrasnosti jezičkih modela koji su plasirani, što je rezultovalo prilagođavanjem skupova podataka za treniranje spornih modela.

4.2 Moguća rešenja problema

Kako se kompanije nose sa ovim problemom i kojim mehanizmima se bore protiv stereotipa u podacima možete pročitati u njihovim zvaničnim saopštenjima [19] [20] [22]. U ovim obrazloženjima pokazuje se da su kompanije svesne da su njihovi jezički modeli pristrasni, ali kako ne postoji način da se

ta pristrasnost u potpunosti eliminiše, upozoravaju korisnike da kritički pristupe svim informacijama koje ovim putem dobiju. Treba imati u vidu da model može razviti i pristrasnost ka stavovima koje korisnik u komunikaciji otvoreno ispoljava, te se složiti sa određenim izjavama, koje nisu u skladu sa činjenicama. Takođe, u odgovoru OpenAI-a saznajemo da su njihovi modeli okrenuti zapadnjačkim stavovima, kao i da su mehanizmi za odstranjivanje lošeg sadržaja često testirani samo na engleskom jeziku. Kompanija Meta dala je detaljnije objašnjenje povodom svog angažmana u rešavanju ovog problema - okupljena je grupa ljudi iz različitih starosnih, rasnih i etničkih grupa kako bi zajedno radili na odstranjivanju predrasuda u podacima, što je rezultovalo objavljivanjem nekoliko novih skupova podataka [19].

Da je pristrasnost jezičkih modela realan problem koji zahteva bar delimično rešenje, govori nam činjenica da su to priznale čak i kompanije koje ove softvere proizvode. Nama ostaje da vidimo da li će se mehanizmi rešavanja ovog problema zaista sprovesti u delo, i koliko će ishod ove situacije uticati na popularnost veštačke inteligencije u budućnosti.

5 Zaključak

U ovom radu, analizirali smo pristrasnosti u veštačkoj inteligenciji, fokusirajući se na sisteme za prepoznavanje lica, društvene mreže i jezičke modele. Proučavali smo konkretne situacije, kako bismo ilustrovali ozbiljnost problema sistema za prepoznavanje lica. Analizirali smo i rezultate istraživanja, poput *The Gender Shades Project*, koji su ukazali na pristrasnost u većini sistema za prepoznavanje. Analizirali smo kako se zapravo određuje sadržaj koji nam se prikazuje i da li su algoritmi dovoljno pametni da prepoznaju lažne vesti. Istraživali smo političke stereotipe prisutne u jezičkim modelima. Rezultati ukazuju na značajne razlike u političkim orijentacijama između različitih modela, s posebnim naglaskom na konzervativne tendencije kod BERT modela i liberalne kod GPT modela. Dodatno smo istražili kako ovi modeli prepoznaju govor mržnje i dezinformacije.

Postoje mnoge studije koje se bave ispitivanjem mehanizama za suzbijanje predrasuda u sistemima veštačke inteligencije. Jedna od njih na zanimljiv način uspostavlja korelaciju između farmaceutske industrije i razvoja veštačke inteligencije [6]. Naime, kao jedno od mogućih rešenja navodi se uvođenje regulatornog tela koje bi se bavilo etičkom stranom veštačke inteligencije (po uzoru na FDA u farmaceutskoj industriji [3]). U ovom slučaju, često se javlja argument da će takav tip rešenja usporiti pojavu inovacija. Međutim, kada se slično dešavalo sa problemom detaljnog ispitivanja novih lekova i hemikalija, rezultat nedovoljnih istraživanja bile su mnogobrojne žrtve koje su te lekove koristile. Iako trenutno tehnologija možda nema tako fatalne uticaje na čovečanstvo, ovo poređenje nam pomaže da shvatimo zašto rešenja ovakvog tipa zaista imaju smisla. Dodatno, izložena su i moguća rešenja poput četvorofaznog testiranja softvera, naročitog obraćanja pažnje na osetljive grupe i slepog testiranja, a više o njima možete pročitati u samom radu [6].

Na osnovu izloženog, uviđamo potrebu za daljim razvojem algoritama koji

mogu smanjiti pristrasnosti u raznim kontekstima, kao i to da treba podići svest ne samo među inženjerima, već i među krajnjim korisnicima veštačke inteligencije kako bi se omogućilo kritičko razmišljanje o radu algoritama. U ovom trenutku nije realno očekivati potpuno odstranjivanje pristrasnosti veštačke inteligencije, jer je praktično nemoguće pronaći velike skupove objektivnih, nepristrasnih podataka na kojima će se modeli trenirati. Otkriće načina za potpunu eliminaciju pristrasnosti u ovoj oblasti bilo bi revolucionarno i zasigurno onemogućilo jedan mehanizam zloupotrebe podataka za različite vidove manipulacije.

Literatura

[1] AWS facial recognition. https://aws.amazon.com/what-is/facial-recognition/.

[2] ChatGPT. https://openai.com/chatgpt.

[3] FDA - U.S. Food and Drug Administration. https://www.fda.gov/.

[4] The political compass. https://www.politicalcompass.org/.

[5] RoBERTa: An optimized method for pretraining self-supervised nlp systems. https://ai.meta.com/blog/roberta-an-optimized-method-for-pretraining-self-supervised-nlp-systems/.

[6] Lorenzo Belenguer. Ai bias: Exploring discriminatory algorithmic decision-making models and the application of possible machine-centric solutions adapted from the pharmaceutical industry. *AI and Ethics*, 2(4):771–787, 2022.

[7] Engin Bozdag and Jeroen van den Hoven. Breaking the filter bubble: democracy and design. *Ethics and Information Technology*, 17(4):249–265, 12 2015.

[8] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*. PMLR, 23–24 Feb 2018.

[9] Datareportal. Digital 2023 deep dive: Time spent on social media. https://datareportal.com/reports/digital-2023-deep-dive-time-spent-on-social-media, 2023.

[10] Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. May 2023.

[11] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. The MIT Press, 2017.

[12] F.O. Isinkaye, Y.O. Folajimi, and B.A. Ojokoh. Recommendation systems: Principles, methods and evaluation. *Egyptian Informatics Journal*, 16(3):261–273, 2015.

[13] David Kirkpatrick. *The Facebook effect: The inside story of the company that is connecting the world*. Simon and Schuster, 2011.

[14] Google LLC. Open sourcing bert: State-of-the-art pre-training for natural language processing. https://blog.research.google/2018/11/open-sourcing-bert-state-of-art-pre.html.

[15] Andreas Graefe Mario Haim and Hans-Bernd Brosius. Burst of the filter bubble? *Digital Journalism*, 6(3):330–343, 2018.

[16] Dan Merica. Hillary clinton calls fake news 'an epidemic' with real world consequences. https://edition.cnn.com/2016/12/08/politics/hillary-clinton-fake-news-epidemic/index.html, 2016.

[17] Meta Platforms, Inc. Introducing Llama 2. https://ai.meta.com/llama/.

[18] Meta Platforms, Inc. Introducing LLaMA: A foundational, 65-billion-parameter large language model. https://ai.meta.com/blog/large-language-model-llama-meta-ai/.

[19] Meta Platforms, Inc. Introducing two new datasets to help measure fairness and mitigate ai bias. https://ai.meta.com/blog/measure-fairness-and-mitigate-ai-bias/.

[20] Michael Schade, OpenAI. How will OpenAI mitigate harmful bias and other negative effects of models served by the api? https://shorturl.at/aCIL8.

[21] Michael Nolan. Llama and ChatGPT are not open-source few ostensibly open-source LLMs live up to the openness claim. https://spectrum.ieee.org/open-source-llm-not-open.

[22] OpenAI. How should AI systems behave, and who should decide? HowshouldAIsystemsbehave,andwhoshoulddecide?

[23] Eli Pariser. Beware online "filter bubbles". https://www.youtube.com/watch?v=B8ofWFx525s&ab_channel=TED.

[24] Eli Pariser. *The Filter Bubble: What the Internet Is Hiding from You*. Penguin, 2011.

[25] Quartz. Bill Gates on the good news about fake news. https://www.youtube.com/watch?v=OzsjvQAk-eA)(https://qz.com/913114/bill-gates-says-filter-bubbles-are-a-serious-problem-with-news, 2018.

[26] Joseph Robinson. Balanced faces in the wild, 2022.

[27] Joseph P Robinson, Gennady Livitz, Yann Henon, Can Qin, Yun Fu, and Samson Timoner. Face Recognition: Too Bias, or Not Too Bias? https://openaccess.thecvf.com/content_CVPRW_2020/papers/w1/Robinson_Face_Recognition_Too_Bias_or_Not_Too_Bias_CVPRW_2020_paper.pdf, 2020.

[28] Robert Williams. I was wrongfully arrested because of facial recognition. why are police allowed to use this technology? https://www.washingtonpost.com/opinions/2020/06/24/i-was-wrongfully-arrested-because-facial-recognition-why-are-police-allowed-use-t 2020.

[29] Yongchun Zhu, Qiang Sheng, Juan Cao, Shuokai Li, Danding Wang, and Fuzhen Zhuang. Generalizing to the future: Mitigating entity bias in fake news detection. SIGIR '22, page 2120–2125, New York, NY, USA, 2022. Association for Computing Machinery.