Etički izazovi u doba veštačke inteligencije: Pristrasnost pod lupom

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.

Sadržaj

Glava 1

Recenzent — ocena: 5

1.1 O čemu rad govori?

Rad istražuje etičke izazove u primeni veštačke inteligencije, sa fokusom na pristrasnost. Kroz primere sistema za prepoznavanje lica, društvenih mreža, i jezičkih modela, prikazano je kako pristrasnost u podacima dovodi do nepravednih i netačnih zaključaka ovih sistema.

1.2 Krupne primedbe i sugestije

Poruka rada bi bila snažnija ako bi postojala zajednička nit koja na jasan način spaja pomenute primere pristrasnosti veštačke inteligencije. To bi mogla da bude studija koja je takođe obrađivala sve ove teme, ili određeni podaci koji bi mogli da se uporede među ovim primerima. Nismo uspeli da nađemo studiju koja je obrađivala sve ove teme zajedno, ali smo se potrudili da ih na neki način povežemo proširivanjem uvoda.

Značajniji deo rada bi trebalo da bude posvećen predloženim rešenjima ovih problema i mislim da bi bilo efektivnije da je to izdvojeno u jasnije celine. Trenutno se rešenja pominju usputno i mislim da nisu dovoljno istaknuta. Studije i komentari o mogućim rešenjima dodati su u okviru zaključka.

Predložio bih i reorganizaciju teksta gde bi se sekcije sa primerima stavile pod jednu zajedničku sekciju. Takođe, bilo bi korisno imati dodatnu diskusiju o važnosti istraživanja i razumevanja pristrasnosti u veštačkoj inteligenciji koja bi mogla da bude unutar uvodne sekcije, ili kao odvojena sekcija. Trenutno su uvodna i zaključna sekcija nedovoljno razrađene i ne prenose jasno poruku rada. Uvodna i zaključna sekcija su dodatno razrađene. U uvodu je sada navedeno da će se u radu ispitivati različite sfere pristrasnosti veštačke inteligencije, pa mislimo da ima smisla da sve budu u zasebnim poglavljima.

1.3 Sitne primedbe

Nisam primetio štamparske greške. Neke od rečenica su preduge i teške za razumevanje, npr. “Kada veštačka inteligencija obučava sisteme za prepoznavanje

lica na podacima koji ne odražavaju stvarni svet ili sadrže inherentne pristrasnosti, to može rezultovati netačnim identifikacijama, posebno među različitim društvenim grupama, uključujući etničke i rodne kategorije".
Ispravljeno. Rečenica je podeljena na dve: "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 posmatranja različitih društvenih grupa, uključujući etničke i rodne kategorije." Neki termini su prevođeni sa engleskog dok drugi nisu (npr. globalni indeks veba, news feed).
Smatramo da ne postoji adekvatan prevod izraza News Feed na osnovu koga bi bilo jasno o čemu se radi, tako da je taj termin ostavljen u tekstu. Napravljena je ispravka tako da se Global Web Index ne prevodi, već piše: "Prema istraživanju kompanije *Global Web Index*..."

1.4 Provera sadržajnosti i forme seminarskog rada

1. Da li rad dobro odgovara na zadatu temu?
   Rad u velikoj meri odgovara na zadatu temu obrađivanjem predloženih pitanja.

2. Da li je nešto važno propušteno?
   Na osnovu moje upućenosti u tematiku, ne.

3. Da li ima suštinskih grešaka i propusta?
   Ne, rad pokušava da odgovori na pitanja zadata temom.

4. Da li je naslov rada dobro izabran?
   Naslov rada je intrigantan i ima akademski prizvuk. Iz naslova nije jasno da će u radu konkretno biti reči o pristrasnosti veštačke inteligencije. Mislim da je naslov dobro izabran, međutim, možda bi bilo bolje da se umesto *digitalno doba* iskoristio termin *doba veštačke inteligencije*.
   Slažemo se sa sugestijom i promenili smo naslov u "Etički izazovi u doba veštačke inteligencije: Pristrasnost pod lupom".

5. Da li sažetak sadrži prave podatke o radu?
   Da, međutim, sažetak deluje previše neformalno i ne pominje neke od aspekata obrađene tematike niti ukazuje na izvedene zaključke.
   Smatramo da je sažetak u skladu sa zahtevima koje treba da ispunjava. Sažetak sadrži opis problema i navedene su neke od tema koje su obrađene u radu. S obzirom na to da sažetak, naslov i sadržaj treba da stanu na prvu stranu, nije moguće njegovo proširivanje, a delimično neformalan govor služi kao motivacija za čitanje rada.

6. Da li je rad lak-težak za čitanje?
   Rad je lak za čitanje. Korišćeni jezik je u većoj meri neformalan i svi nepoznati pojmovi su u dobroj meri objašnjeni.

7. Da li je za razumevanje teksta potrebno predznanje i u kolikoj meri?
   Za razumevanje je potrebno minimalno tehničko predznanje o načinu funkcionisanja sistema veštačke inteligencije.

8. Da li je u radu navedena odgovarajuća literatura?
   Da, sve tvrdnje i podaci su potkrepljeni referencama iz kojih su preuzeti.

9. Da li su u radu reference korektno navedene?
   Da, autori su konzistentni u načinu na koji navode reference.

10. Da li je struktura rada adekvatna?
    Trenutna struktura je laka za praćenje i deli rad u logične celine. Međutim, mislim da bi bilo prikladnije da su sekcije 2-4 podsekcije u sekciji koja bi se mogla nazvati *Primeri pristrasnosti veštačke inteligencije* ili slično. Uvodna i zaključna sekcija su manjkave i ovakva organizacija bi potencijalno podstakla dodatne prelazne segmente koji bi čitaoca još bolje uveli u tematiku. Odgovoreno u odeljku krupnih sugestija.

11. Da li rad sadrži sve elemente propisane uslovom seminarskog rada (slike, tabele, broj strana...)?
    Rad sadrži propisane elemenate, naime, ima tri slike, četiri tabele, 12 stranica, 27 referenci. Rad ima bar jednu knjigu, bar jedan naučni članak iz časopisa i bar jednu veb adresu kao referencu. Slika 1 i tabele 1 i 2 su autorske.

12. Da li su slike i tabele funkcionalne i adekvatne?
    Da, slike i tabele su većinski preuzete iz referenci koje se pominju u tekstu i prate tekst.

1.5 Ocenite sebe

d) malo upućen

Svestan sam problematike i pročitao sam nekoliko članaka i odgledao nekoliko kratkih predavanja koja su mi bila preporučena na društvenim mrežama. Međutim, nisam aktivno istraživao datu temu i nisam aktivno kritički razmišljao o njoj.

Glava 2

Recenzent — ocena: 5

2.1 O čemu rad govori?

U ovom radu prikazano je kako odabir podataka za treniranje modela veštačke inteligencije može uticati na stvaranje pristrasnost prema određenim podacima prilikom pravljenja predikcija. Konkretno, stavljen je akcenat na analiziranje pojave pristrasnosti kod sistema za prepoznavanje lica, sistema za preporuku sadržaja, detekciju lažnih vesti i kod jezičkih modela koji mogu stvarati određene stereotipe.

2.2 Krupne primedbe i sugestije

U radu su spomenute različite sfere u kojima se javlja problem pristrasnosti prilikom pravljenja predikcija. Kao ideju u okviru detekcije lažnih vesti predložio bih problem *clickbait* naslova. Na Internetu sam naišao na radove koji se bave ovim problemom, tako da mislim da bi rad bio još zanimljiviji ukoliko biste spomenuli i ovaj problem u okviru lažnih vesti. Iako pojam *clickbait* naslova inače pripada temi lažnih vesti, mi se ovde ipak fokusiramo samo na aspekt mogućnosti veštačke inteligencije da te lažne vesti (ne) prepozna zbog "afiniteta" koji je razvila prema određenim vestima. Kako su *clickbait* naslovi samo pokušaj urednika da privuče čitaoce lažnim predstavljanjem onoga što se u tekstu nalazi, smatramo da ovoj temi nije mesto u našem radu.

U okviru sekcije 3 Društvene mreže, čitajući podsekciju 3.2 Filter mehur stvorio sam utisak da spada pod okvire sistema za preporuku sadržaja, pa bih predložio samo da se podsekcija 3.2 uvede kao sekcija 3.1.1. Sugestija je uvažena, te se *Filter mehur* sada nalazi u sekciji 3.1.1.

2.3 Sitne primedbe

U okviru sekcije 3.2 sam primetio dve slovne greške (šlici, Đa). Slovne greške su ispravljene. U okviru sekcije 4.1 nalaze se dve tabele koje prikazuju rezultate modela za prepoznavanje govora mržnje i pogrešnih informacija. Bolji rezultati su obojeni tamnijim nijansama, lošiji svetlijim. Imao bih samo primedbu na odabir boje tabele 3, jer tamnije nijanse crvene boje prirodno daju utisak da je rezultat lošiji, a u vašem slučaju, rezultati obojeni tom bojom su bolji. Boje su promenjene, sada su obe tabele u nijansama zelene boje.

2.4 Provera sadržajnosti i forme seminarskog rada

1. Da li rad dobro odgovara na zadatu temu?
   Da. Rad u potpunosti odgovara na zadatu temu analizirajući različite sfere digitalnog sveta u kojem je prisutna pristrasnost modela veštačke inteligencije.

2. Da li je nešto važno propušteno?
   Ne. Sama suština teme je sadržana u ovom radu.

3. Da li ima suštinskih grešaka i propusta?
   Ne. Sve je korektno.

4. Da li je naslov rada dobro izabran?
   Naslov je malo duži, i na prvi pogled ne bih znao da se radi o pristrasnosti koju stvaraju modeli veštačke inteligencije.
   Naslov je promenjen u "Etički izazovi u doba veštačke inteligencije: Pristrasnost pod lupom" kako bi bilo očiglednije da je u pitanju problem pristrasnosti veštačke inteligencije.

5. Da li sažetak sadrži prave podatke o radu?
   Da. U sažetku se navode ključna pitanja na koja ovaj rad odgovara.

6. Da li je rad lak-težak za čitanje?
   Rad se čini lakim sa čitanje.

7. Da li je za razumevanje teksta potrebno predznanje i u kolikoj meri?
   Za razumevanje ovog rada potrebno je elementarno znanje mašinskog učenja.

8. Da li je u radu navedena odgovarajuća literatura?
   Da.

9. Da li su u radu reference korektno navedene?
   Da.

10. Da li je struktura rada adekvatna?
    Struktura je adekvatna, s primedbom na sekciju 3.2 Filter mehur koju sam već naveo.

11. Da li rad sadrži sve elemente propisane uslovom seminarskog rada (slike, tabele, broj strana...)?
    Rad sadrži 4 tabele, 3 slike, 12 strana i 27 referenci, čime zadovoljava sve propisane uslove.

12. Da li su slike i tabele funkcionalne i adekvatne?
    Da. Sve tabele i slike se čine nezavisnim u odnosu na tekst. Jedina primedba koju sam ranije naveo jeste odabir boje tabele 3. Boje su promenjene, sada su obe tabele u nijansama zelene boje.

2.5 Ocenite sebe

c) srednje upućen
Imam iskustva sa pravljenjem modela veštačke inteligencije i svestan sam problema koji mogu nastati ukoliko se ne odaberu adekvatni podaci za treniranje samog modela.

Glava 3

Dodatne izmene

Nije bilo dodatnih izmena.