ANIDA VRCIĆ AMAR

Alfa BK univerzitet

avrcicamar@gmail.com

ULFETA MAROVAC

Državni univerzitet u Novom Pazaru, Departman za tehničko-tehnološke nauke

umarovac@np.ac.rs

Emocije u senci algoritama: poređenje ljudskog i mašinskog prepoznavanja emocija u tekstu

Apstrakt

Ovo istraživanje ispituje razlike između ljudske i algoritamske interpretacije emocija u tekstu. Skup podataka sastoji se od 404 objave prikupljene sa srpskog internet foruma. Objave su ručno anotirali studenti psihologije i automatski su analizirane korišćenjem modela GouEmoušns (*GoEmotions*). Ljudski anotatori identifikovali su 144 normalizovane kategorije emocija, koje su mapirane na 28 kategorija modela GouEmoušns korišćenjem soft-label reprezentacije. Rezultati pokazuju umeren nivo saglasnosti između anotatora (F1_micro = 0,58; F1_macro = 0,41) i otkrivaju značajne razlike između ljudske interpretacije emocija i interpretacije generisane modelom. Dok su ljudski anotatori najčešće identifikovali klasične emocije (42,8%), model je ovoj kategoriji dodelio znatno manji udeo (14,5%) i često je klasifikovao tekst kao neutralan (31,9% u odnosu na 1,2% u ljudskim anotacijama). Iako su suprotne predikcije sentimenta bile retke (~6%), model je često ublažavao emocionalni intenzitet klasifikujući emocionalno obojene tekstove kao neutralne. Ovi nalazi ukazuju na sistematske razlike između statističkog prepoznavanja emocija i ljudske kontekstualne interpretacije afektivnog sadržaja.

Ključne reči: prepoznavanje emocija, NLP, GouEmoušns (*GoEmotions*), ljudska anotacija

Uvod

U eri brzog razvoja veštačke inteligencije (*AI*) sve češće se postavlja pitanje u kojoj meri su kognitivne sposobnosti ljudi i mašina uporedive. Napredni veliki jezički modeli (*LLM*), poput Dži-Pi-Ti-4 (*GPT-4*), pokrenuli su raspravu o potencijalnoj „kognitivnoj paritetnosti" između biološke i nebiološke inteligencije (Bubeck et al., 2023). Ipak, emocije – kao jedna od najkompleksnijih dimenzija ljudskog iskustva – ostaju domen u kojem se ljudi i algoritmi i dalje značajno razlikuju. Afektivno

iskustvo obuhvata više od puke kategorizacije osećanja; uključuje telesne senzacije, kognitivnu procenu i kulturološki oblikovanu jezičku ekspresiju (Barrett, 2017). Iako AI sistemi mogu pouzdano prepoznati leksičke indikatore emocija, ostaje otvoreno pitanje da li oni zaista mogu razumeti emocionalni sadržaj na način na koji to ljudi čine, pre svega zbog odsustva iskustvene i telesne dimenzije ljudskog emocionalnog doživljaja (Colombetti, 2014). Tradicionalni psihološki modeli univerzalnih osnovnih emocija polaze od pretpostavke da postoji ograničen skup evolucijski univerzalnih emocija (Ekman, 1992). Nasuprot tome, savremenija empirijska istraživanja pokazuju da ljudi spontano razlikuju znatno veći broj suptilnih emocionalnih stanja, često više od dvadeset kategorija (Cowen & Keltner, 2021). Na osnovu ovih saznanja razvijen je model GouEmoušns, jezički model velikih razmera treniran na više od 58.000 ručno anotiranih komentara, koji omogućava klasifikaciju 27 kategorija emocija (Demszky et al., 2020). Iako model GouEmoušns postiže visoku tačnost u detekciji emocija, njegova sposobnost interpretacije složenijih afektivnih stanja ostaje ograničena. U savremenim NLP pristupima emocije se najčešće predstavljaju kroz latentne i distribucione reprezentacije u prostoru podataka (Buechel & Hahn, 2025). Tekstualni izrazi poput „praznina, „unutrašnja borba" ili „samoponižavanje" nose semantičke i emocionalne slojeve daleko iznad leksičke površine. Razumevanje ovih slojeva zahteva empatijsku simulaciju, iskustvenu usklađenost i kulturološko znanje, što su resursi koje *AI* modeli ne poseduju (Barrett & Bliss-Moreau, 2009). Štaviše, modeli poput GouEmoušnsa ne obuhvataju mnoge moralno ili egzistencijalno složene emocije.

Polazeći od ograničenja savremenih modela za analizu emocija, ovaj rad ispituje razlike između ljudske i mašinske interpretacije emocionalnih izraza u tekstu na srpskom jeziku. Posebna pažnja posvećena je analizi kompleksnih i kontekstualno zavisnih emocija kroz poređenje ljudskih anotacija i predikcija transformatorskih modela.

Rad je organizovan na sledeći način: drugo poglavlje daje teorijski okvir, treće opisuje korpus, proces anotacije i metodologiju, četvrto predstavlja evaluaciju modela, dok peto poglavlje prikazuje rezultate, diskusiju i zaključke.

Teorijski okvir analize emocija u tekstu

Emocije imaju važnu ulogu u oblikovanju ljudskog ponašanja, donošenju odluka i međuljudskoj komunikaciji. U savremenom digitalnom okruženju društvene mreže i forumi generišu velike količine tekstualnih podataka koji sadrže informacije o stavovima i emocionalnim reakcijama korisnika. Obrada prirodnog jezika (*Natural Language Processing – NLP*) omogućava računarima da analiziraju i interpretiraju ljudski jezik, čime se postavlja osnova za prepoznavanje emocija u tekstu (Jurafsky & Martin, 2009). NLP tehnike koriste se i za automatsko označavanje specifičnih termina u domenski orijentisanim tekstovima, kao što su medicinski izveštaji i elektronski zdravstveni kartoni (Avdić, Marovac, & Janković, 2020). Raniji pristupi uglavnom su se zasnivali na tehnikama obrade teksta poput tokenizacije, lematizacije i obrade negacije, dok savremeni pristupi sve više koriste modele zasnovane na dubokom učenju za analizu složenih jezičkih i kontekstualnih obrazaca. Analiza sentimenta određuje opšti emocionalni ton teksta (pozitivan, negativan ili neutralan), dok analiza emocija ide korak dalje i identifikuje specifična emocionalna stanja, poput radosti, tuge, straha ili besa (Liu, 2020: 83). Za modelovanje emocija najčešće se koriste *kategorijalni modeli*, koji definišu skup osnovnih emocija, i *dimenzionalni modeli*, koji emocije opisuju kroz kontinuirane dimenzije poput valence i pobuđenosti (Ekman, 1992). U savremenim istraživanjima često se primenjuju transformatorski modeli, a jedan od najčešće korišćenih okvira za klasifikaciju emocija predstavlja GouEmoušns, skup od 28 kategorija emocija razvijen na velikom korpusu Redit (*Reddit*) komentara (Demszky et al., 2020).

Analiza sentimenta i emocija u srpskom jeziku suočava se sa izazovima karakterističnim za morfološki bogate jezike, kao što su višeznačnost reči, nestandardni pravopis, žargon, obrada negacije i slobodan red reči, dok dodatni problem predstavlja ograničena dostupnost resursa poput anotiranih korpusa i leksičkih baza podataka. Nekoliko resursa razvijeno je za analizu sentimenta u srpskom jeziku. Hibridni pristupi koji kombinuju leksičke metode i mašinsko učenje pokazali su dobre rezultate u klasifikaciji polariteta (Mladenović et al., 2016), dok obrada negacije predstavlja važan faktor za tačnost detekcije sentimenta (Ljajić & Marovac, 2019). Među značajnim leksičkim resursima izdvaja se SentiVordNet 3.0 (*SentiWordNet 3.0*), koji dodeljuje skorove

polariteta sinonimskim skupovima i omogućava modelovanje afektivnih značenja (Baccianella et al., 2010). Takođe je razvijen leksikon Senti_Pol_sr. (*Senti-Pol-sr*), prilagođen srpskim morfološkim varijantama (Ljajić & Marovac, 2019).

Kada su u pitanju korpusi, dostupni su skup recenzija filmova na srpskom jeziku (Batanović et al., 2016) i književni korpusi koji obuhvataju analize starih srpskih romana (Stanković et al., 2022). Ipak, većina postojećih istraživanja fokusira se na klasifikaciju polariteta ili analizu aspekata, dok su studije o prepoznavanju složenih emocija u srpskom jeziku i dalje ograničene (Nikolić et al., 2020; Grljević et al., 2022). Novija istraživanja počinju da razvijaju resurse za modelovanje emocija. Na primer, anotirani korpusi studentskih povratnih informacija korišćeni su za detekciju emocija primenom višejezičnih transformatorskih (transformer) modela (Vidaković et al., 2025), dok su razvijeni i specijalizovani leksikoni emocija za srpski jezik (Šošić et al., 2026).

Međutim, navedena istraživanja prvenstveno su usmerena na razvoj resursa i metoda za automatsku detekciju sentimenta i emocija, dok se poređenje između ljudske i mašinske interpretacije emocija u tekstu ne razmatra direktno.

Skup podataka i metodologija

Skup podataka korišćen u ovom istraživanju prikupljen je sa internet foruma Ana.rs, namenjenog ženskoj populaciji. Analiza obuhvata 485 objava na srpskom jeziku, napisanih kao odgovori na pitanje: „Koji su to osećaji koje nikada nećete zaboraviti?" Nakon čišćenja podataka (uklanjanje duplikata, veoma kratkih i nevažećih odgovora), konačni skup za analizu sadrži 404 objave.

Za potrebe ovog istraživanja skup podataka anotirala su dva studenta psihologije u saradnji sa drugim autorom rada. Anotatorima nije bio unapred zadat fiksni model niti definisan skup kategorija emocija; umesto toga, slobodno su identifikovali emocije koje su prepoznali u svakom tekstu koristeći sopstvene izraze. Kako bi se obezbedilo zajedničko razumevanje zadatka, u početnoj fazi zajednički je anotirano približno 10% korpusa, tokom čega su diskutovani nejasni slučajevi i usklađena tumačenja emocionalnih izraza. Nakon ove kalibracije preostale objave anotirane su nezavisno. Po završetku

anotacije sprovedena je normalizacija oznaka emocija, koja je uključivala ujednačavanje pisma i velikih slova, uklanjanje tipografskih varijacija i objedinjavanje sinonima u standardizovane forme. Nakon normalizacije i objedinjavanja sinonimnih i flektovanih oblika identifikovan je skup od 144 jedinstvene emocije, koje su zatim mapirane na 28 kategorija GoUEmoušns taksonomije. Početnu šemu mapiranja definisao je drugi autor rada (doktor kliničke psihologije), nakon čega je ona usaglašena sa studentima-anotatorima kako bi se obezbedila konzistentnost kategorija. Za svaku objavu dodeljeni su realni skorovi u opsegu [0,1], čime je omogućeno povezivanje jednog teksta sa više emocija. Ovaj pristup predstavlja meko označavanje (*soft-label*) reprezentaciju, koja zadržava informacije o intenzitetu i preklapanju emocija. Saglasnost između anotatora procenjena je nakon normalizacije emocija. Skorovi su binarizovani pragom 0,30, a saglasnost je izračunata pomoću F1-mere u micro i macro režimu. Rezultati ukazuju na umeren nivo saglasnosti (F1-micro = 0,58; F1-macro = 0,41), što je očekivano u istraživanjima anotacije emocija zbog subjektivne prirode njihove interpretacije. U završnoj fazi konstruisane su prosečne ljudske meko označene (*soft-label*) anotacije (avg), dobijene prosečavanjem skorova svih anotatora. Ovaj pristup omogućava objedinjenu reprezentaciju ljudske interpretacije emocija uz očuvanje informacija o njihovom intenzitetu i preklapanju.

Pored ljudskih anotacija, korpus je dodatno analiziran korišćenjem pretreniranog modela GouEmoušns (*bhadresh-savani/bert-base-go-emotion*), zasnovanog na BERT arhitekturi i treniranog na oko 58.000 ručno anotiranih Redit komentara u 27 kategorija emocija i neutralnoj klasi. Pošto je model treniran na engleskom jeziku, objave su prethodno automatski prevedene korišćenjem servisa Gugl prevodilac (*Google Translate*). Model je primenjen na sve objave u analiziranom korpusu kako bi se dobile automatske predikcije emocija. Za svaki tekst model generiše skorove u opsegu [0, 1] za svaku od 27 emocija, kao i za neutralnu klasu, čime se dobija višelabelna reprezentacija emocionalnog sadržaja.

Time je omogućeno direktno poređenje između ljudskih anotacija i modelskih predikcija emocija. Radi lakše interpretacije rezultata, 27 emocija iz GouEmoušns taksonomije grupisane su u četiri šire kategorije:

- Klasične emocije – radost, tuga, bes, strah i iznenađenje.
- Afektivna stanja – složenija i dugotrajnija stanja poput razočaranja, tuge ili ponosa.
- Relacione i socijalne emocije – emocije povezane sa interpersonalnim odnosima, kao što su ljubav, zahvalnost i divljenje.
- Fiziološka i reaktivna stanja – brze emocionalne reakcije poput nervoze ili uzbuđenja.

Ovakvo grupisanje omogućava pregledniju analizu raspodele emocija i lakše poređenje između ljudskih anotacija i predikcija modela. Pošto su sve izvorne oznake emocija mapirane na GouEmoušns taksonomiju, dalja analiza zasniva se na ovoj standardizovanoj klasifikaciji.

Radi poređenja na nivou opšte emocionalne polarizacije, za svaku objavu izračunata je agregirana vrednost sentimenta kao razlika između ukupnih skorova pozitivnih i negativnih emocija. Ako je razlika bila veća od praga 0,40, tekst je klasifikovan kao pozitivan ili negativan, dok su ostali postovi označeni kao neutralni. Tako je svaki tekst svrstan u jednu od tri kategorije: negativan (–1), neutralan (0) ili pozitivan (1).

Na osnovu ljudskih anotacija i predikcija modela GouEmoušns zatim je upoređena raspodela emocionalnih kategorija u korpusu. Analiza obuhvata učestalost četiri konceptualne grupe emocija, kao i razlike između ljudskih i modelskih procena sentimenta, čime se ispituju sličnosti i razlike u interpretaciji emocionalnog sadržaja.

Rezultati

Kako bi se dodatno ispitalo slaganje između ljudskih anotacija i predikcija modela GouEmoušns, izvršeno je poređenje na nivou ukupnog sentimenta. Na osnovu agregiranih skorova pozitivnih i negativnih emocija, svaka objava je klasifikovana u jednu od tri kategorije: negativan, neutralan ili pozitivan sentiment. Tabela 1. prikazuje matricu zabune (konfuzije) između ljudskih anotacija i modelskih predikcija.

Tabela 1. Matrica (zabune) konfuzije između ljudskih anotacija i predikcija modela GouEmoušns

	Model Negative	Model Neutral	Model Positive
Human Negative	**24**	**80**	**20**
Human Neutral	**10**	**50**	**16**
Human Positive	**5**	**122**	**77**

Iako postoje odstupanja između ljudskih anotacija i modelskih predikcija, slučajevi direktnog neslaganja u smislu suprotnog polariteta (negativan ↔ pozitivan) relativno su retki. U posmatranom skupu podataka takvi slučajevi čine približno 6% svih primera, što ukazuje da model retko potpuno pogrešno interpretira emocionalni ton teksta.

Međutim, analiza matrice zabune (konfuzije) pokazuje jedan drugačiji obrazac greške. Najveći broj neslaganja potiče iz činjenice da model veliki broj emocionalno obojenih objava klasifikuje kao neutralne. Na primer, čak 122 objave koje su ljudski anotatori označili kao pozitivne, model klasifikuje kao neutralne, dok je 80 negativnih objava takođe svrstano u neutralnu kategoriju.

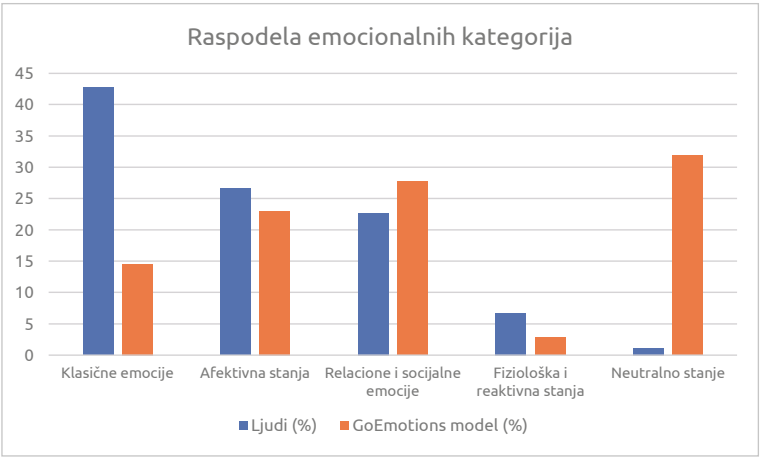
Ovakav obrazac ukazuje na izraženu tendenciju modela ka neutralnoj klasifikaciji. Drugim rečima, model često „ublažava" emocionalni sadržaj teksta i smanjuje intenzitet prepoznatih emocija u odnosu na ljudske anotacije.

Na primer, u objavi: „Meni je nestvaran osećaj prve zaljubljenosti, kad se sav streseš, kad hodaš kao da letiš, skoro dečije, neiskvareno." Model GouEmoušns je dodelio negativan polaritet zato što su emocije poput stresa i blage nervoze dobile relativno visoke skorove u odnosu na pozitivne emocije. Nasuprot tome, u ručnoj anotaciji objava je označena kao pozitivna (ljubav, sreća). Ovakvo neslaganje proizlazi iz činjenice da model interpretira „stres" kao negativnu emociju, iako u ovom kontekstu ona zapravo predstavlja uzbuđenje kao deo pozitivnog iskustva.

Slični primeri ukazuju na razlike u načinu na koji ljudi i modeli interpretiraju emocionalni sadržaj teksta. Dok model obrađuje emocije kao izolovane signale i tehnički ih agregira, ljudi emocije procenjuju u širem kontekstu, naglašavajući dominantnu emociju i njen intenzitet.

Poređenje ljudskih anotacija i predikcija modela GouEmoušns pokazuje jasne razlike u distribuciji emocionalnog sadržaja. Analiza raspodele emocionalnih kategorija na celom skupu podataka otkriva značajne razlike između ljudske interpretacije emocija i automatskih predikcija modela (grafikon 1).

Grafikon 1. Raspodela emocionalnih kategorija u ljudskim anotacijama i predikcijama modela GouEmoušns



U ljudskim anotacijama dominantnu ulogu imaju klasične emocije, koje čine 42,8% svih identifikovanih emocionalnih stanja. Nasuprot tome, model GouEmošns ovoj kategoriji dodeljuje znatno manji udeo (14,5%). Ovaj rezultat ukazuje da ljudski anotatori najčešće prepoznaju osnovne emocije poput radosti, tuge, besa i straha.

Sa druge strane, model znatno češće klasifikuje tekst kao neutralan, pri čemu neutralna kategorija čini 31,9% svih modelskih predikcija, dok je u ljudskim anotacijama gotovo zanemarljiva (1,2%). Ova razlika predstavlja jedan od najuočljivijih obrazaca u poređenju ljudskih i modelskih interpretacija emocionalnog sadržaja.

Razlike su prisutne i u drugim kategorijama, ali su manje. Relacione i socijalne emocije kod modela čine 27,7%, dok su kod ljudskih anotacija zastupljene u 22,7% slučajeva. Afektivna stanja su kod ljudi nešto zastupljenija (26,6%) nego kod modela (23,0%). Najmanji udeo u obe analize imaju fiziološka i reaktivna stanja, koja kod ljudi čine 6,8%, dok ih model prepoznaje u svega 2,8% slučajeva.

Ovakvi rezultati ukazuju na različite obrasce interpretacije emocionalnog sadržaja. Dok ljudski anotatori spontano identifikuju diskretne osnovne emocije, model češće interpretira tekst kroz neutralne ili relacione kategorije, čime se smanjuje vidljivost klasičnih emocija u automatskoj klasifikaciji.

Kako bi se dodatno ispitalo koje emocije prepoznaju model i ljudski anotatori u objavama za koje su se složili u pogledu sentimenta, analizirana je raspodela emocionalnih kategorija u slučajevima slaganja između ljudskih anotacija i predikcija modela GouEmoušns (tabela 2).

Tabela 2. Raspodela emocionalnih kategorija u slučajevima slaganja između ljudskih anotacija i modela (%)

	Pozitivan sentiment		**Negativan sentiment**	
Grupa emocija	**Ljudi (%)**	**Model (%)**	**Ljudi (%)**	**Model (%)**
Klasične emocije	36,46	13,87	51,47	33,66
Afektivna stanja	20,6	15,75	34,54	37,19
Relacione i socijalne emocije	37,44	50,2	4,21	10,7
Fiziološka i reaktivna stanja	4,79	3,55	7,34	3,78
Neutralno	0,71	16,64	2,45	14,67

Rezultati pokazuju da, i pored slaganja u pogledu ukupnog polariteta, ljudi i model različito raspodeljuju emocionalne kategorije. U objavama sa pozitivnim sentimentom model najčešće identifikuje relacione i socijalne emocije (50,2%), dok ljudski anotatori pored visokog udela relacionih emocija (37,44%) značajan deo emocionalnih izraza klasifikuju i kao klasične emocije (36,46%). Takođe je primetno da model znatno češće dodeljuje neutralnu kategoriju (16,64%) u poređenju sa ljudskim anotacijama (0,71%).

Sličan obrazac prisutan je i kod negativnog sentimenta. U ljudskim anotacijama dominiraju klasične emocije (51,47%), dok model češće naglašava afektivna stanja (37,19%). Takođe, model i u negativnim objavama dodeljuje veći udeo neutralnoj kategoriji (14,67%) nego ljudski anotatori (2,45%).

Ovi rezultati ukazuju da model, čak i kada pravilno identifikuje sentiment teksta, drugačije interpretira njegovu emocionalnu strukturu u odnosu na ljudske anotatore.

Kako bi se detaljnije ispitao obrazac neutralne klasifikacije modela, analizirane su objave koje je model GouEmoušns označio kao neutralne, iako su ljudski anotatori prepoznali jasan emocionalni polaritet. Posmatrana su dva slučaja neslaganja: (1) objave koje su ljudi označili kao pozitivne, a model klasifikovao kao neutralne i (2) objave koje su ljudi označili kao negativne, dok ih je model takođe svrstao u neutralnu kategoriju. Tabela 3 prikazuje raspodelu emocionalnih kategorija u ovim slučajevima.

Tabela 3. Raspodela emocionalnih kategorija u postovima koje model GouEmoušns klasifikuje kao neutralne

	Pozitivno–neutralno		**Negativno–neutralno**	
Grupa emocija	**Ljudi (%)**	**Model (%)**	**Ljudi (%)**	**Model (%)**
Klasične emocije	38,52	10,26	48,68	14,74
Afektivna stanja	21,52	21,55	34,92	27,87
Relacione i socijalne emocije	31,58	18,89	7,18	20,28
Fiziološka i reaktivna stanja	7,4	2,48	7,72	2,8
Neutralno	0,97	46,82	1.5	34,32

Rezultati pokazuju konzistentan obrazac razlika između ljudskih anotacija i modelskih predikcija. U objavama koje su ljudi označili kao pozitivne dominiraju klasične emocije (38,5%) i relacione i socijalne emocije (31,6%), dok model gotovo polovinu ovih objava povezuje sa neutralnom klasom (46,8%). Sličan obrazac prisutan je i kod objava koje su ljudi označili kao negativne, gde dominiraju klasične emocije (48,7%) i afektivna stanja (34,9%), dok model ponovo značajan deo emocionalnog sadržaja klasifikuje kao neutralan (34,3%).

Ovi rezultati ukazuju na izraženu tendenciju modela da ublažava emocionalni intenzitet teksta, interpretirajući emocionalno obojene poruke kroz neutralniji okvir u poređenju sa ljudskim anotatorima.

Diskusija

Rezultati ovog istraživanja pružaju uvid u sličnosti i razlike između ljudskog i algoritamskog prepoznavanja emocija u tekstu, koristeći model GouEmoušns kao referentni sistem za automatsku klasifikaciju emocionalnog sadržaja. Analiza matrice zabune (konfuzije) pokazuje da, uprkos određenim odstupanjima, slučajevi potpune neusaglašenosti u smislu suprotnog emocionalnog polariteta (negativan ↔ pozitivan) ostaju relativno retki, čineći oko 6% primera. Ovaj nalaz potvrđuje da modeli uspevaju da prepoznaju osnovni emocionalni ton teksta, što je u skladu sa ranijim radovima o statističkim predikcijama polariteta u velikim jezičkim modelima (Demszky et al., 2020).

Detaljnija analiza pokazuje izraženu tendenciju modela ka neutralnoj klasifikaciji. Konkretno, čak 122 objave koje su ljudski anotatori označili kao pozitivne, model je svrstao u neutralnu kategoriju, dok je 80 negativnih objava takođe dodeljeno neutralnom polju. Ovaj obrazac sugeriše da algoritmi „ublažavaju" emocionalni intenzitet, što je u skladu sa nalazima (Buechel & Hahn, 2018), koji ukazuju na to da statistički zasnovani modeli često ne prepoznaju implicitne emocionalne signale.

Ovakav obrazac može se objasniti načinom na koji modeli identifikuju emocionalne indikatore u tekstu. Većina sistema oslanja se na statističke obrasce pojedinačnih reči ili fraza, zbog čega su implicitne i kontekstualne emocije teže prepoznatljive (Poria, Cambria & Gelbukh, 2016).

Raspodela emocionalnih kategorija pokazuje jasne razlike između ljudskih anotacija i modelskih predikcija. Ljudski anotatori najčešće identifikuju klasične emocije (42,8%), zatim afektivna stanja (26,6%) i relacione i socijalne emocije (22,7%), dok je neutralna kategorija gotovo zanemarljiva (1,2%). Nasuprot tome, model najčešće dodeljuje neutralnu kategoriju (31,9%), zatim relacione i socijalne emocije (27,7%) i afektivna stanja (23,0%), dok su klasične emocije znatno ređe zastupljene (14,5%). Fiziološka i reaktivna stanja imaju najmanju zastupljenost kod ljudi (6,8%) i modela (2,8%).

Ove razlike mogu se objasniti načinom na koji modeli identifikuju emocionalne obrasce u tekstu. Dok se modeli uglavnom oslanjaju na statističke indikatore reči i fraza, ljudski evaluatori uzimaju u obzir širi diskurzivni i psihološki kontekst, što omogućava prepoznavanje složenijih emocionalnih nijansi (Buechel & Hahn, 2018).

Analiza objava sa slaganjem u sentimentu pokazuje da, čak i kada model pravilno prepozna globalni polaritet, razlike u raspodeli emocionalnih kategorija ostaju. Modeli teže da generalizuju emocije prema jezički jasno izraženim i statistički učestalim obrascima, često favorizujući relacione ili neutralne kategorije, dok ljudski anotatori uzimaju u obzir širi kontekst i interpersonalna značenja (Almulla, 2024). Ove razlike ukazuju na ograničenja NLP modela u prepoznavanju složenih i kontekstualnih emocija, koje i dalje predstavljaju izazov za automatizovanu detekciju (García Hernández et al., 2024).

Klinička i praktična važnost ovih nalaza je višestruka. Prema istraživanjima, internet je već dovoljno dugo izvor psihološke podrške pri informisanju i savetovanju u vezi sa mnogim psihološkim stanjima i poremećajima, između ostalog i problemima postpartalne depresije (Fazlagić, 2017), a u novijoj istoriji i kovida-19 (Davidović Rakić et al., 2024). Takav trend se evidentno nastavlja uz sve veći broj korisnika koji se obraća AI sistemima za neformalnu psihološku podršku, često u situacijama kada profesionalna pomoć nije odmah dostupna (Abd-Alrazaq et al., 2020).

AI sistemi mogu analizirati obrasce u tekstu i identifikovati određene emocionalne indikatore, ali njihovo razumevanje psihološkog konteksta i složenih emocionalnih stanja ostaje ograničeno. U kliničkom okruženju to zahteva opreznu primenu, jer digitalni agenti mogu pružiti određeni oblik podrške, ali ne mogu zameniti profesionalnu psihološku procenu i terapiju (Abd-Alrazaq et al., 2020). Modeli poput GouEmoušnsa mogu biti korisni za analizu emocionalnog sadržaja, ali njihova primena u osetljivim kontekstima zahteva pažljivo tumačenje rezultata i ljudski nadzor. Ovi rezultati ukazuju da, uprkos napretku u tačnosti klasifikacije, razumevanje složenih i kontekstualnih emocija i dalje predstavlja izazov za automatizovane sisteme, naglašavajući razliku između algoritamske preciznosti i ljudske empatijske interpretacije emocija.

Zaključak

Rezultati ovog istraživanja ukazuju na jasne razlike između ljudske i algoritamske interpretacije emocija u tekstu. Dok modeli za prepoznavanje emocija mogu uspešno identifikovati određene obrasce u jeziku i postići relativno visoku tačnost klasifikacije, njihova

interpretacija emocionalnog sadržaja često ostaje ograničena na statističke obrasce u podacima. Ljudski anotatori, sa druge strane, češće prepoznaju suptilne i kontekstualne emocionalne nijanse koje nisu eksplicitno izražene kroz leksičke indikatore.

Analiza distribucije emocija pokazala je da modeli imaju tendenciju da emocionalno obojene tekstove češće klasifikuju kao neutralne ili da ublažavaju intenzitet izraženih emocija. Ovakav obrazac sugeriše da automatizovani sistemi mogu imati poteškoća u prepoznavanju kompleksnih ili ambivalentnih emocionalnih stanja koja su u velikoj meri zavisna od konteksta, implicitnog značenja i iskustvene interpretacije.

Dobijeni rezultati ukazuju da savremeni modeli za analizu emocija, iako predstavljaju značajan tehnološki napredak, ne mogu u potpunosti reprodukovati način na koji ljudi razumeju emocionalni sadržaj u jeziku. Razlika između statističke klasifikacije i empatijske interpretacije emocija predstavlja važan izazov za dalji razvoj sistema zasnovanih na veštačkoj inteligenciji.

Ovo istraživanje takođe ukazuje na potrebu za razvojem bogatijih korpusa i modela koji bi bolje uzimali u obzir kontekstualne, kulturne i narativne aspekte emocionalnog izražavanja, posebno u jezicima sa ograničenim resursima. Buduća istraživanja mogla bi uključiti veće korpuse, multimodalne izvore podataka i pristupe koji kombinuju automatske metode sa ljudskom ekspertizom kako bi se unapredilo razumevanje emocionalnog sadržaja u tekstu.

LITERATURA

Abd-Alrazaq, A. A., Rababeh, A., Alajlani, M., Bewick, B. M. & Househ, M. (2020). Effectiveness and safety of using chatbots to improve mental health: Systematic review and meta-analysis. *Journal of Medical Internet Research*, *22*(7), e16021. https://doi.org/10.2196/16021

Almulla, M. A. (2024). On the effect of prior knowledge in text-based emotion recognition. *International Journal of Research and Scientific Innovation (IJRSI), 11*(8), 48–58. https://doi.org/10.51244/IJRSI.2024.1108005

Avdic, A., Marovac, U. & Jankovic, D. (2020). Automated labeling of terms in medical reports in Serbian. *Turkish Journal of Electrical Engineering and Computer Sciences*, *28*(6), 3285–3303. https://doi.org/10.3906/elk-1912-144

Baccianella, S., Esuli, A. & Sebastiani, F. (2010). SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In: N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis & D. Tapias (eds.), *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, (2200–2204). Valletta: European Language Resources Association.

Barrett, L. F. (2017). *How emotions are made: The secret life of the brain*. London: Pan Macmillan.

Barrett, L. F. & Bliss-Moreau, E. (2009). Affect as a psychological primitive. *Advances in Experimental Social Psychology*, *41*, 167–218. https://doi.org/10.1016/S0065-2601(08)00404-8

Batanović, V., Nikolić, B. & Milosavljević, M. (2016). Reliable baselines for sentiment analysis in resource-limited languages: The Serbian movie review dataset. In: N. Calzolari, K. Choukri, K. Declerck, S. Goggi, M. Grobelnik, B. Maegaard & S. Piperidis (eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, (2688–2696). Portorož: European Language Resources Association.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E. & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*. https://doi.org/10.48550/arXiv.2303.12712

Buechel, S. & Hahn, U. (2018). Emotion representation mapping for automatic lexicon construction (mostly) performs on human level. In: *Proceedings of the 27th International Conference on Computational Linguistics*, (2892–2904). Santa Fe, NM: Association for Computational Linguistics.

Buechel, S. & Hahn, U. (2025). Emotion embeddings – Learning stable and homogeneous abstractions from heterogeneous affective datasets. *IEEE Transactions on Affective Computing*. https://doi.org/10.1109/TAFFC.2025.3635686

Colombetti, G. (2014). *The feeling body: Affective science meets the enactive mind*. Cambridge, MA: MIT Press.

Cowen, A. S. & Keltner, D. (2021). Semantic space theory: A computational approach to emotion. *Trends in Cognitive Sciences*, *25*(2), 124–136. https://doi.org/10.1016/j.tics.2020.11.008

Davidović Rakić, J., Vrcić Amar, A., Belić, M. & Maksimović, A. (2024). Sociodemographic characteristics, hope and information in the media as predictors of the fear of COVID-19. *Romanian Journal of Communication and Public Relations*, *26*(1), 25–43. https://doi.org/10.21018/rjcpr.2024.1.458

Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A. S., Nemade, G. & Mohit, B. (2020). GoEmotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*. https://doi.org/10.48550/arXiv.2005.00547

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, *6*(3–4), 169–200. https://doi.org/10.1080/02699939208411068

Fazlagić, A. (2017). Da li žene znaju šta je postpartalna depresija?. *Materia Medica*, *33*(1), 1475–1480.

García Hernández, R. A., Luna García, H., Celaya Padilla, J. M., García Hernández, A., Reveles Gómez, L. C., Flores Chaires, L. A. & Villalba Condori, K. O. (2024). A systematic literature review of modalities, trends, and limitations in emotion recognition, affective computing, and sentiment analysis. *Applied Sciences*, *14*(16), 7165. https://doi.org/10.3390/app14167165

Grljević, O., Bošnjak, Z. & Kovačević, A. (2022). Opinion mining in higher education: A corpus-based approach. *Enterprise Information Systems*, *16*(5), 1773542. https://doi.org/10.1080/17517575.2020.1773542

Jurafsky, D. & Martin, J. H. (2009). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (2nd ed.). Upper Saddle River, NJ: Pearson.

Liu, B. (2020). *Sentiment analysis: Mining opinions, sentiments, and emotions* (2nd ed.). Cambridge: Cambridge University Press.

Ljajić, A. & Marovac, U. (2019). Improving sentiment analysis for Twitter data by handling negation rules in the Serbian language. *Computer Science and Information Systems*, *16*(1), 289–311. https://doi.org/10.2298/CSIS181022051L

Mladenović, M., Mitrović, J., Krstev, C. & Vitas, D. (2016). Hybrid sentiment analysis framework for a morphologically rich language. *Journal of Intelligent Information Systems*, *46*(3), 599–620. https://doi.org/10.1007/s10844-015-0389-8

Nikolić, N., Grljević, O. & Kovačević, A. (2020). Aspect-based sentiment analysis of reviews in the domain of higher education. *The Electronic Library, 38*(1), 44–64. https://doi.org/10.1108/EL-08-2019-0189

Poria, S., Cambria, E. & Gelbukh, A. (2016). Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In: J. Su, X. Carreras & K. Duh (eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (2539–2544). Austin, TX: Association for Computational Linguistics. https://doi.org/10.18653/v1/D16-1263

Stanković, R., Košprdić, M., Ikonić Nešić, M. & Radović, T. (2022). Sentiment analysis of Serbian old novels. In: D. Milošević, C. Banea & R. Mihalcea (eds.), *Proceedings of the 2nd Workshop on Sentiment Analysis and Linguistic Linked Data (SALLD-2)*, (31–38). Marseille: European Language Resources Association.

Šošić, M., Graovac, J. & Stanković, R. (2026). Building an emotion lexicon for Serbian using curated language resources. *Language Resources and Evaluation*, *60*(1), Article 9. https://doi.org/10.1007/s10579-025-09894-5

Vidaković, D., Luburić, N., Kovačević, A. & Slivka, J. (2025). Enhancing software and learning with Serbian student feedback corpora. *Language Resources and Evaluation*, *59*(4), 4193–4221. https://doi.org/10.1007/s10579-025-09855-y

Anida Vrcić Amar
Ulfeta Marovac

Emotions in the Shadow of Algorithms: Comparing Human and Machine Emotion Recognition in Text

Abstract

This study examines the differences between human and algorithmic interpretation of emotions in text. The dataset consists of 404 posts collected from a Serbian online forum. The posts were manually annotated by psychology students and automatically analyzed using the GoEmotions model. Human annotators identified 144 normalized emotion categories, which were mapped to the 28 GoEmotions categories using a soft-label representation. The results show a moderate level of agreement between annotators (F1_micro = 0.58; F1_macro = 0.41) and reveal significant differences between human and model-generated emotion interpretations. While human annotators most frequently identified classical emotions (42.8%), the model assigned a considerably smaller share to this category (14.5%) and often classified the text as neutral (31.9% compared to 1.2% in human annotations). Although opposite sentiment predictions were rare (~6%), the model frequently attenuated emotional intensity by classifying emotionally expressive texts as neutral. These findings indicate systematic differences between statistical emotion detection and human contextual interpretation of affective content.
Keywords: emotion recognition, NLP, GoEmotions, human annotation