Uporaba umjetne inteligencije za verifikaciju informacija

Lisak, Lucija

SVEUČILIŠTE U ZAGREBU
FILOZOFSKI FAKULTET
ODSJEK ZA INFORMACIJSKE I KOMUNIKACIJSKE ZNANOSTI

Ak. god. 2023./2024.

Lucija Lisak

Uporaba umjetne inteligencije za verifikaciju informacija

Završni rad

Mentor: dr. sc. Nives Mikelić Preradović

Zagreb, kolovoz 2024.

Izjava o akademskoj čestitosti

Izjavljujem i svojim potpisom potvrđujem da je ovaj rad rezultat mog vlastitog rada koji se temelji na istraživanjima te objavljenoj i citiranoj literaturi. Izjavljujem da nijedan dio rada nije napisan na nedozvoljen način, odnosno da je prepisan iz necitiranog rada, te da nijedan dio rada ne krši bilo čija autorska prava. Također izjavljujem da nijedan dio rada nije korišten za bilo koji drugi rad u bilo kojoj drugoj visokoškolskoj, znanstvenoj ili obrazovnoj ustanovi.

_____

(potpis)

Sadržaj

1. Uvod

U digitalnom okruženju u kakvom danas živimo, protok i kolanje informacija su, zahvaljujući društvenim mrežama, iznimno brzi i učinkoviti. Mijenjaju se načini i brzina širenja i primanja informacija te korisnici postaju preopterećeni raznovrsnim informacijama među kojima je teško razlikovati koje su istinite, a koje lažne. Brzo širenje informacija internetom i društvenim mrežama rezultira velikim brojem neistinitih i nepouzdanih informacija, poznatih kao dezinformacije. Posljedice širenja dezinformacija su dalekosežne, jer mogu, na primjer, utjecati na javno mišljenje i stavove ljudi u društvu. Kao potencijalno rješenje za verifikaciju informacija i zaustavljanje širenja dezinformacija javljaju se umjetna inteligencija i sustavi za automatiziranu provjeru informacija, čime bi se doprinijelo jačanju povjerenja javnosti u medije. Mogli bismo opisati umjetnu inteligenciju (AI) kao proces automatizacije zadataka koji obično zahtijevaju ljudsku inteligenciju, odnosno repliciranje različitih aspekata ljudskog mišljenja i ponašanja. U ovom kontekstu, važna je ona strana umjetne inteligencije koja omogućava računalima da zamijene ljudsku inteligenciju u obavljanju zadataka poput učenja, razmišljanja, rješavanja problema i donošenja odluka. Algoritmi i matematički modeli omogućavaju AI sustavima prikupljanje i analizu podataka, učeći iz njih kako bi primijenili novo znanje i unaprijedili postojeće.

Ovaj rad istražuje kako se umjetna inteligencija već koristi u verifikaciji informacija i detekciji dezinformacija, uključujući automatizirano provjeravanje informacija (AFC), Human-in-the-Loop (HITL) pristup, kao i detekciju deepfake slika, videozapisa i audiozapisa. Uvođenje umjetne inteligencije u proces provjere informacija donosi značajne prednosti, ali i izazove s kojima se treba suočiti. Cilj je potvrditi autentičnost informacija na internetu, identificirati lažne vijesti, dezinformacije i manipulativne sadržaje, te konačno osvijestiti korisnike i pružiti im pouzdane informacije.

2. Informacija i dezinformacija

Pojam „informacija“ (ili „obavijest“) u Hrvatskoj enciklopediji definiran je kao „skup podataka s pripisanim značenjem, osnovni element komunikacije koji, primljen u određenoj situaciji, povećava čovjekovo znanje“ („Hrvatska Enciklopedija, mrežno izdanje“, bez dat.).

Danas je protok informacija u velikoj mjeri ubrzan zahvaljujući internetu, društvenim mrežama i općenito digitalizaciji. Promjene u načinu i brzini širenja i zaprimanja informacija dovele su do gomilanja različitih informacija među kojima može biti izazov razaznati koje su istinite, a koje su lažne. Iznimno brzo širenje informacija internetom i društvenim mrežama rezultira velikim udjelom neistinitih i nepouzdanih informacija koje se nazivaju dezinformacijama. Te lako dostupne pogrešne informacije velikom se brzinom kreću digitalnim okruženjem te na taj način postaju dostupne svima (Santos, 2023). Pojam lažnih vijesti (u ovom slučaju pojam vijest poistovjećujemo s pojmom informacija) obilježava fenomen značajan za ovu temu, a prvi se put pojavio već u izvješću iz 2016. godine vezanom uz prijetnju uplitanja Rusije u izbore za Europski parlament i potencijalne terorističke propagande. Europska komisija izvješćem iz 2018. godine skrenula je pozornost na pojavu lažnih vijesti u formi satire, propagande i obmane. Takvo širenje lažnih vijesti definiralo se i sintagmom „informacijski poremećaji“ (Grmuša i Prelog, 2020). Spomenuta Europska komisija dezinformacije definira kao „lažne, netočne ili zavaravajuće informacije osmišljene, predstavljene i promovirane s namjerom da izazovu javnu štetu ili ostvare profit“ (Santos, 2023).

Kulić (kao što citiraju Grmuša i Prelog) lažne vijesti klasificira u tri kategorije – *false news*, *fake news* i *news satira*. Ono što međusobno razlikuje navedene kategorije jest istinitost informacija i namjera iza stvaranja i dijeljenja istih. *False news* i *fake news* su kategorije koje karakteriziraju u potpunosti neistinite informacije, no za razliku od *false news* koje mogu biti slučajne, *fake news* stvorene su s namjerom da obmane čitatelje, odnosno primatelje informacija. Prema Kulić, s druge strane, za *news satiru* nije nužno da bude lažna. Lažna vijest zapravo je oksimoron jer je istinitost temeljni čimbenik vijesti, a prema Beck (kao što citiraju Grmuša i Prelog) lažna vijest postaje vijest „samo ako je recipijent prihvati kao pravu“. Lažne vijesti mogu se nazivati i lažnim informiranjem gdje se radi o novom obliku digitalne manipulacije koja može rezultirati sukobom i propagandom (Grmuša i Prelog, 2020). U ovom kontekstu važno je istaknuti politiku i zdravstvo kao područja u kojima dezinformacije ostavljaju dubok trag. Za primjer možemo uzeti kolanje netočnih informacija

u vrijeme pandemije COVID-19 (Santos, 2023). Što se politike tiče, tehnološki napredak u velikoj se mjeri koristi za manipulaciju informacijama i širenje dezinformacija. Kertysova (2018) izdvaja četiri vrste prijetnji – profiliranje korisnika i segmentacija; hiperpersonalizirano ciljanje; „deep fakes"; ljudi pozicionirani „out-of-the-loop" u AI sustavima. Razvoj strojnog učenja omogućava detaljniju identifikaciju osobina, potreba i stavova korisnika što dovodi do personalizacije sadržaja čime se najučinkovitije može djelovati na korisnike. Hiperpersonalizirano ciljanje u kombinaciji s alatima za generiranje prirodnog jezika može se koristiti za automatsko generiranje sadržaja za korisnike. Negativna strana toga je ugrožavanje privatnosti i zaštite osobnih podataka. Deepfake tehnologija svojim stalnim napredovanjem ugrožava poimanje istine i autentičnosti. Njome se stvaraju uvjerljivi lažni prikazi ljudi, što predstavlja ozbiljnu prijetnju njihovom identitetu i privatnosti. Pozicioniranjem ljudi „out-of-the-loop" zanemaruje se činjenica da je ograničavanje automatske provedbe odluka temeljenih na problemima koje je identificirala umjetna inteligencija ključno za održavanje ljudskog nadzora.

Posljedice širenja dezinformacija dalekosežne su, one, primjerice, mogu utjecati na javno mišljenje i stavove ljudi u društvu. Kao što je već spomenuto, zbog velikog kolanja informacija putem interneta dolazi do problema po pitanju razlikovanja istine i fikcije („AI-Driven Recruitment Trends #13 | AI's Crucial Role in Fact-Checking and Misinformation", 2023). Drugi problem koji se javlja jest i potkopavanje kredibiliteta novinara kao izvora vijesti u društvu. Santos (2023) pojašnjava kako su napredak u tehnologiji i umjetna inteligencija istovremeno utjecali i na širenje dezinformacija, ali i na automatizirano stvaranje istih. Grmuša i Prelog (2020) kao faktore koji utječu na sustav vrijednosti navodi načine distribucije i jačanje korisničkog sadržaja te promjene do kojih je došlo u medijskoj strukturi. Oni smatraju da bi umjetna inteligencija i sustavi za automatiziranu provjeru informacija mogli biti rješenje kada je riječ o verifikaciji informacija i zaustavljanju protoka dezinformacija, a samim time pridonijeli bi povjerenju javnosti u medije.

3. Uloga umjetne inteligencije u verifikaciji informacija i detekciji dezinformacija

Russell i Norvig (kao što citira Santos) umjetnu inteligenciju (AI) definiraju kao „proces automatizacije zadataka koji obično zahtijevaju ljudsku inteligenciju. Drugim riječima, uključuje repliciranje različitih aspekata ljudskog razmišljanja i ponašanja“. Autor članka „Validation and Verification of Artificial Intelligence“ (2024) opisuje umjetnu inteligenciju u kontekstu računala koja mogu zamijeniti ljudsku inteligenciju u izvršavanju zadataka kao što su: učenje, promišljanje, rješavanje problema i donošenje odluka. Algoritmi i matematički modeli omogućuju AI sustavima da prikupljaju i analiziraju podatke te uče od njih kako bi prikupljeno, novo znanje kasnije primijenili, a staro unaprijedili. Beckett („Roundtable: How artificial intelligence is used to debunk fake news“, 2020) definira umjetnu inteligenciju kao tehnologiju koju čine strojno učenje (ML), algoritmi i podaci. Zadaci kojima se AI trenutno bavi uglavnom su repetitivni, no uspješno izvršava i zadatke koje ljudi ne mogu, bilo zbog vlastitih sposobnosti ili prevelikog opsega posla. Kertysova (2018) zaključuje kako nema općeprihvaćene i jedinstvene definicije umjetne inteligencije, već da je ono što se pojavljuje u većini objašnjenja mogućnost sustava da uspješno izvršava zadatke na razini koja zahtijeva ljudsku inteligenciju.

Kao važno polje unutar umjetne inteligencije potrebno je navesti strojno učenje (engl. *machine learning*, ML). Zahvaljujući strojnom učenju, prepoznavanje uzoraka u velikim skupovima podataka automatizirano je pomoću algoritama. Na taj način ono omogućava da strojevi prikupljaju podatke, nauče ih i kasnije izvršavaju zadatke na temelju tog stečenog znanja (Santos, 2023). Kertysova (2018.) strojno učenje definira kao „upotrebu algoritama i velikih baza podataka za uvježbavanje računalnih sustava da prepoznaju uzorke koji prethodno nisu definirani te mogućnost tih sustava da nauče iz podataka i razaznaju važne informacije bez da su prethodno programirani za isključivo to.“ U svojem radu Kertysova (2018) se bavi tehnikama strojnog učenja čiji je razvoj usmjeren prema umjetnoj inteligenciji. Riječ je o programima za audio-vizualne analize koji su algoritmom programirani da prepoznaju i kontroliraju dvojben sadržaj na internetu.

Ünver (2023) u radu ukratko sažima povijest razvoja umjetne inteligencije. Kao sami početak ističe 1950. godinu kada je britanski matematičar Alan Turing napisao „Computing Machinery and Intelligence“, rad u kojem je postavio pitanje „Can machines think?“ te je u

sklopu toga razvio i test za umjetnu svijest (engl. *artificial consciousness*, AFC). U to doba polje umjetne inteligencije uglavnom je bilo teoretsko s obzirom na to da računalstvo nije bilo dovoljno napredno da bi se Turingove ideje provele u djela. Godine 1956. formiran je pojam umjetna inteligencija, a s vremenom je porastao i interes za to polje. Herbert Simon i Allen Newell 1956. godine projektirali su prvi inteligentni stroj. Riječ je o programu Logic Theorist koji je rješavao matematičke probleme. Uskoro je napredak u polju umjetne inteligencije privukao i interes američke vlade koja je zatim počela sufinancirati istraživačke programe. 1960-ih američko-njemački informatičar Joseph Weizenbaum napravio je prvi „chatbot" nazvan ELIZA. Chatbot je mogao imitirati razgovore između ljudi te se stoga smatra prvim primjerom neurolingvističkog programiranja (NLP). Nadalje, 1997. godine IBM-ov Deep Blue pobijedio je Garryja Kasparova, svjetskog prvaka u šahu te je iste godine Microsoft's Windows implementirao softver za prepoznavanje glasa. Polje umjetne inteligencije ubrzano se razvijalo, fokus je stavljan na duboko strojno učenje (engl. *deep machine learning*, DML) i neuronske mreže. Pojam strojno učenje (ML) definiran je 1959. godine u sklopu programa razvijenog za samoučenje kako bi igrao igru dame. 2020. godine razvijen je model na kojem se temelje ChatGPT i GPT-3, a u narednim je godinama OpenAI objavio i manje generativne AI modele (Ünver, 2023).

Utjecaj umjetne inteligencije na današnji svakodnevni život iznimno je velik i neosporan te da se taj fenomen odrazio na i infiltrirao u svaki aspekt života. Strojno učenje svojim je napretkom dovelo do revolucionarnih otkrića i razvoja tehnologije, industrije, obrazovanja, medicine i sl. („Validation and Verification of Artificial Intelligence", 2024).

S obzirom na to da se distribucija dezinformacija očituje kao sve veći problem za koji je potrebno naći rješenje, umjetna inteligencija se javlja kao potencijalan odgovor. Činjenica je da umjetna inteligencija može biti korištena u stvaranju i širenju dezinformacija, međutim može igrati važnu ulogu i u verifikaciji informacija i detekciji dezinformacija. Multimodalna automatska detekcija, strojno učenje i razni klasifikacijski algoritmi od velike su koristi u takvim zadacima. Dok multimodalna automatska detekcija prikuplja tekstualne i vizualne naznake prilikom identificiranja dezinformacije, strojno učenje i klasifikacijski algoritmi pojedine podatke organiziraju u kategorije na temelju njihovih karakteristika (Santos, 2023). Važno je napomenuti da lažan sadržaj može biti djelo različitih izvora u digitalnom informacijskom okruženju. Takav sadržaj može biti računalno generiran (engl. *computer-generated*) (sintetski) i korisnički generiran (engl. *user-generated*) što bi značilo da je prikupljen na internetu i ručno modificiran. Posao analitičara prilikom razotkrivanja i

kontroliranja dezinformacija sastoji se od prikupljanja i pregledavanja velikih količina tvrdnji pomoću alata i softvera. Drugim riječima, potrebno je identificirati uzorke, klasificirati tekstualne i audio-vizualne podatke, naznačiti sličnosti među uzorcima i sl. Uloga umjetne inteligencije u ovom dijelu posla jest značajna jer modeli osmišljeni za odrađivanje tih poslova mogu uvelike ubrzati proces i biti koristan alat analitičarima (Juršėnas, Karlauskas, Ledinauskas, Maskeliūnas, Ruseckas i Rondomanskas, 2022).

Autor članka „AI-Driven Recruitment Trends #13 | AI's Crucial Role in Fact-Checking and Misinformation“ (2023) izdvaja pet načina na koji se AI pokazao korisnim, posebno po pitanju strojnog učenja (ML) i obrade prirodnog jezika (engl. *natural language processing*, NLP):

1. Automatizirano provjeravanje informacija (engl. *automated fact-checking*) – pomoću algoritama za analizu transkripcija političkih govora i debata, AI platforme poput ClaimBuster detektiraju tvrdnje koje je potrebno verificirati.
2. Detekcija tvrdnji (engl. *claim detection*) – Full Fact's Live Monitor i slični alati prate emisije koje se emitiraju uživo i transkripcije te u stvarno vremenu provjeravaju tvrdnje koje su izrekli političari i javne ličnosti.
3. Verifikacija slika (engl. *image verification*) – na temelju analize promjena na razini piksela, AI algoritmi mogu detektirati uređivane slike te na taj način pridonijeti identifikaciji digitalno izmijenjenog vizualnog sadržaja.
4. Identifikacija lažnih vijesti (engl. *identifying fake news*) – GPT-3 i slični AI modeli detektiraju izmišljene i manipulirane članke na način da detektiraju uzorke uobičajene za lažne vijesti.
5. Praćenje i analiza društvenih mreža (engl. *social media monitoring*) – InVID i slične platforme pomoću analize kadrova i obrnutog pretraživanja na društvenim mrežama detektiraju lažne videozapise.

Juršėnas i ostali (2022) skreću pozornost na to da je važno da AI sustav obavijesti i objasni administratorima platforme zašto je određeni sadržaj prosudio kao potencijalni izvor dezinformacije. Na taj se način može procijeniti vjerojatnost pristranosti algoritma u određenoj odluci, a samim time se osnažuje povjerenje administratora u AI alate. K tome, cijeli proces pomaže administratorima da obrazlože korisnicima zašto je pojedini račun klasificiran kao onaj koji izvršava neautentične radnje (Juršėnas i ostali, 2022). Umjetna inteligencija u velikoj se mjeri već koristi u novinarstvu za izvršavanje repetitivnih zadataka. Ti poslovi podrazumijevaju prikupljanje podataka i detekciju relevantnih uzoraka, sve što bi

novinarima oduzelo iznimno puno vremena i truda (Santos, 2023). Štoviše, uključenost AI u produkciju vijesti nije ograničena samo na prikupljanje podataka već ima ulogu i u oblikovanju komentara, provjeri informacija te verifikaciji medijskog sadržaja. Autor izdvaja alate za prepoznavanje lažnih izvora vijesti kao iznimno bitne s obzirom da je njihova funkcija da detektiraju dezinformaciju prije negoli je objave („Human vs AI Fact-Checkers: A Comparative“, 2024).

Kintsurashvili („Roundtable: How artificial intelligence is used to debunk fake news“, 2020) čvrsto stoji iza tvrdnje da umjetna inteligencija ne može zamijeniti ljudski rad, bez obzira na dosadašnji napredak. Kao što je već rečeno, prednosti AI očituju se u uštođenom vremenu i detekciji problema. Kao primjer daje The Myth Detector koji se bavi provjerom informacija te omogućuje pristup velikoj količini podataka koji su trenutno u opticaju na internetu. S obzirom na to da velike količine informacija kolaju internetom i dostupne su iznimno velikom broju korisnika, važno je reagirati u stvarnom vremenu. Kertysova (2018) konstatira kako će u budućnosti napredak po pitanju preciznosti i izvedbe umjetne inteligencije rezultirati strojevima koji uspijevaju u zadacima koje ljudi ne mogu izvršiti. To bi značilo da, primjerice, pristranost ne bi predstavljala problem kada je riječ o donošenju odluka.

Elemary („Roundtable: How artificial intelligence is used to debunk fake news“, 2020) podsjeća kako je u pozadini umjetne inteligencije uvijek ljudska inteligencija. U svojem radu Kertysova (2018) skreće pozornost na činjenicu da bez obzira na automatiziranu detekciju i uklanjanje lažnog sadržaja, umjetna inteligencija također sa sobom donosi ograničenja, izazove i nehotične posljedice.

4. Automatizirano provjeravanje informacija (Fact-Checking)

4.1. Što je Fact-Checking?

„Fact-checking se odnosi na proces verifikacije informacija kako bi točnost izjava baziranih na činjenicama bila potvrđena“ (Santos, 2023). „Fact-checking kao zadatak ima procijeniti jesu li pisane ili izrečene tvrdnje istinite. To je osnovni zadatak u novinarstvu i obično ga organizacije posvećene tome, poput PolitiFact, izvršavaju ručno“ (Guo, Schlichtkrull i Vlachos, 2022). Cilj sustava za provjeru informacija jest borba protiv dezinformacija, što se očituje u verifikaciji izjava te stavljanju neistinitih informacija u kontekst. Inicijative koje se bave provjerom i verifikacijom podataka i informacija pronađenih na internetu, na društvenim mrežama ili izjavama koje imaju učinak na društvo mogu biti neovisne ili specificirani odjeli (Gutiérrez-Caneda i Vázquez-Herrero, 2024). Skromniji cilj projekata usmjerenih provjeri informacija jest da izdvajanjem i ukazivanjem na dezinformacije koje kruže društvenim mrežama vrate vjeru u novinarstvo, s naglaskom na transparentnost i kredibilitet (Santos, 2023).

1995. godine u Sjedinjenim Američkim Državama stvorena je prva inicijativa za provjeru informacija, što je potaknulo val novih organizacija diljem svijeta. Prema istraživanju provedenom 2019. godine, tada je bilo aktivno 135 web-stranica koje su se bavile provjerom informacija. Porast broja takvih organizacija sve je veći, što se može vidjeti u istraživanju provedenom 2023. godine prema kojem je tada bilo aktivno 417 web-stranica za provjeru informacija (Gutiérrez-Caneda i Vázquez-Herrero, 2024). Santos (2023) ističe 2003. godinu kao onu kada se javio interes za provjeru informaciju u medijima uoči pokretanja Factcheck.org. Tada je fokus prvenstveno bio na izjavama javnih ličnosti. Gutiérrez-Caneda i Vázquez-Herrero (2024) pojašnjavaju kako je u poslovnom smislu provjeravanje činjenica usmjereno prvenstveno prema novinarstvu što se i očituje u količini pretraživanja, uspoređivanja, obrade podataka te vizualizaciji informacija.

Autor članka „Human vs AI Fact-Checkers: A Comparative“ (2024) podsjeća kako je provjera informacija od iznimno velike važnosti prilikom ne samo detekcije dezinformacija već i objave istinitih informacija čime se izbjegava nastajanje zablude. Novinari i analitičari su ti koji se tradicionalno bave poslovima kao što je ručna provjera informacija, međutim ukoliko uzmemo u obzir količinu informacija koje putuju internetom dolazimo do jasnog zaključka da je to proces koji zahtijeva mnogo vremena i napora.

4.2. Uloga umjetne inteligencije u procesu provjere informacija

Razvojem tehnologije napreduje i umjetna inteligencija te njena primjena u raznim područjima ljudskog života. U kontekstu informacijske pismenosti, umjetna inteligencija svoju je primjenu našla u provjeri informacija („Human vs AI Fact-Checkers: A Comparative“, 2024). Zbog brzine kojom se danas medijima šire informacije i dezinformacije, analitičari se prilikom provjere okreću automatizaciji. Pokrenuta su istraživanja i projekti automatizacije provjeravanja informacija na temelju prirodne obrade jezika, strojnog učenja, prikaza znanja (engl. *knowledge representation*, KR) i baza podataka. Cilj je automatski verificirati tvrdnje (Guo i ostali, 2022).

Kao što je već rečeno, provjeru informacija izvorno su ručno provodili „fact-checkers“, no vrijeme i količina informacija u opticaju pokazali su da je takav način rada nedjelotvoran i neefikasan. Interes za provjeru informacija potpomognutu umjetnom inteligencijom javio se u vrijeme predizborne kampanje američkog kandidata za mjesto predsjednika, Donalda Trumpa. Projekti automatizacije s godinama su sve više financirani kako bi njihov razvoj pridonio: identifikaciji, verifikaciji i ispravnosti sadržaja koji se pojavljuje u medijima (Kertysova, 2018). Međutim, činjenica da iza brzog porasta lažnog sadržaja stoji i umjetna inteligencija dovodi stručnjake u dvojbu. S jedne strane, sadržaj koji je stvoren umjetnom inteligencijom nastaje brže, što podrazumijeva i iznimno velike količine podataka koje je potrebno provjeriti. K tome, pojedini analitičari smatraju da softveri prirodne obrade jezika nisu uvijek pouzdani, već ponekad stvaraju dodatan lažan sadržaj i obmane. S druge strane, kada se uzmu u obzir sve mogućnosti koje umjetna inteligencija nudi, ona se javlja kao koristan alat u borbi protiv dezinformacija i lažnog sadržaja (Gutiérrez-Caneda i Vázquez-Herrero, 2024). Brojni projekti ručne provjere informacija pokrenuti su kako bi se zaustavio val dezinformacija, među kojima se kao bitniji izdvajaju PolitiFact, FactCheck.org, FullFact i Snopes (Nakov i ostali, 2021).

Umjetna inteligencija pruža niz usluga u ovome polju, na način da omogućuje analitičarima da analiziraju veći obujam podataka, automatiziraju mehaničke zadatke i komuniciraju s korisnicima o dezinformacijama koje su otkrivene. Sve to uvelike olakšava posao analitičara i omogućuje da se obradi što veći dio sadržaja u što kraćem vremenu (Gutiérrez-Caneda i Vázquez-Herrero, 2024). Međutim, problem na koji Gutiérrez-Caneda i Vázquez-Herrero (2024) ukazuju su zakon te etika i deontologija koji ograničavaju djelovanje umjetne inteligencije na postupke kojima izbjegavaju rizike glede autorskih prava, privatnosti korisnika te pristranosti algoritma. Brandtzaeg (kao što citiraju Gutiérrez-Caneda i Vázquez-

Herrero): „Uporaba umjetne inteligencije u provjeri činjenica zahtijeva prepoznavanje njezinih ograničenja i objašnjenje kriterija odabira argumenata za donošenje zaključka, a ne samo fokusiranje na krajnji rezultat“.

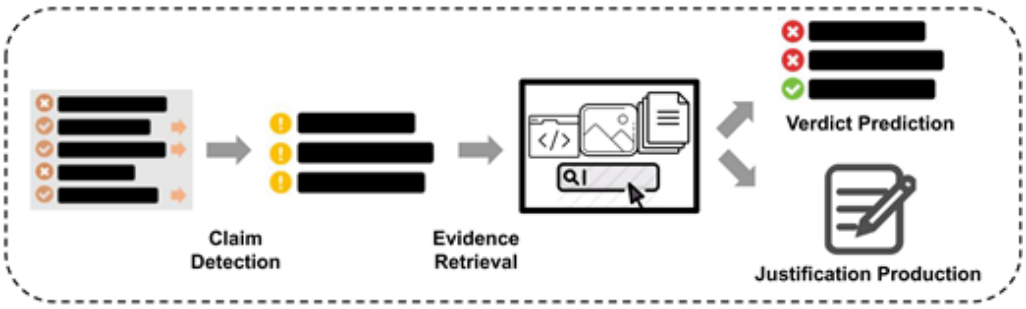
Gutiérrez-Caneda i Vázquez-Herrero (2024) pojašnjavaju kako postoji mnogo termina za uporabu umjetne inteligencije u novinarstvu, kao što su “artificial journalism”, “algorithmic journalism”, “computational journalism”, “robot journalism” i “automated journalism”. Što se tiče povijesti uporabe umjetne inteligencije u novinarstvu, ona započinje 2010. godine, razvojem alata pod nazivom Quakebot. Radilo se o alatu Los Angeles Timesa koji je stvarao vijesti vezane uz potrese u toj regiji na način da je prikupljao informacije iz javnih izvora i automatski pisao i objavljivao prikupljene informacije na web-stranici novina. To je pokrenulo širu uporabu umjetne inteligencije u različitim medijima, ne samo u proizvodnji vijesti, već i u diseminaciji, personalizaciji i verifikaciji sadržaja (Gutiérrez-Caneda i Vázquez-Herrero, 2024). Dosad su uglavnom neprofitne i neovisne organizacije bile te koje su razvijale i implementirale automatiziranu provjeru informacija (Kertysova, 2018). Graves (2018) ističe da je fokus inicijativa automatizirane provjere informacija na jednom ili više od ukupno tri cilja koja se međusobno isprepliću. Ti ciljevi su uočavanje neistinitih ili upitnih tvrdnji na internetu ili drugom obliku medija, verifikacija tvrdnji koje su upitne ili pomoć novinarima u provjeri informacija te da ispravne tvrdnje budu izložene korisnicima koji su prethodno bili izloženi neistinitim informacijama.

4.3. Elementi provjere informacija

Ono prema čemu se sustavi za provjeru informacija, za početak, usmjeravaju su sustavi koji nadziru različite vrste diskursa. Riječ može biti o (online ili u tradicionalnim medijima) govorima, komentarima, debatama ili vijestima. Zadaci koji AFC (automatizirano provjeravanje informacija, engl. *automated fact-checking*) obično podrazumijevaju su rad s prijepisima i drugim materijalima prikupljenima na internetu, praćenje feedova titlova uživo te uporaba automatskog prijepisa. Tek kada je taj dio obavljen, na red dolazi identificiranje i provjera tvrdnji (Graves, 2018).

U svojem radu Guo i ostali (2022) proces provjere informacija dijele u tri faze – detekcija tvrdnje (engl. *claim detection*), prikupljanje dokaza (engl. *evidence retrieval*) i verifikacija tvrdnje (engl. *claim verification*) koja je podijeljena na predviđanje presude (engl. *verdict prediction*) i produkciju opravdanja (engl. *justification production*). Predviđanje presude

podrazumijeva označavanje istinitosti tvrdnji, a produkcija opravdanja kao zadatak ima objasniti zašto se tvrdnja smatra ili se ne smatra vjerodostojnom. Slika 1 prikaz je NLP okvira imenovanih faza. Ukratko, faza detekcije tvrdnje podrazumijeva identifikaciju tvrdnji koje je potrebno verificirati; faza prikupljanja dokaza usmjerena je na pronalazak materijala, odnosno dokaza, koji ili potvrđuju ili opovrgavaju tvrdnju; faza verifikacije tvrdnje finalno određuje vjerodostojnost tvrdnje na temelju prikupljenih dokaza. Guo i ostali (2022) pojašnjavaju kako dok je prva faza zasebna, posljednje dvije faze, prikupljanje dokaza i verifikacija tvrdnje, često imaju zajednički zadatak pod nazivom činjenična provjera (engl. *factual verification*).



Slika 1. *Proces provjere informacija* (2022) Preuzeto 10.08.2024. sa https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00454/109469/A-Survey-on-Automated-Fact-Checking

4.3.1. Detekcija tvrdnje

„Fact-checkers“ koji se bave provjerom informacija prvenstveno moraju izdvojiti tvrdnje koje je potrebno provjeriti. U sklopu toga važno je procijeniti potencijalnu štetu koju dezinformacije mogu nanijeti društvu, kao što su demokratski procesi, ugrožavanje zdravlja ili pogoršanje hitnih situacija. Pažnja se treba skrenuti i na trud i radnu snagu potrebnu za provjeru tvrdnje, kao i na važnost nepristranosti ljudi koji te informacije obrađuju (Nakov i ostali, 2021). Nakov i ostali (2021) ističu kako određivanje „checkworthness“ može predstavljati veliki izazov, posebice ako se radi o „breaking news“ iz razloga što su često takve priče i istinite i popularne. Štoviše, velik broj tvrdnji koje su vrijedne provjere, ostaje po strani i neprovjereno zbog nedostatka resursa. Hassan i ostali (kao što citiraju Guo i ostali) definiraju tvrdnje vrijedne provjere kao „one za koje bi široj javnosti bilo od interesa znati istinu.“ Proces odabira može se bazirati na binarnim odlukama za svaku potencijalnu tvrdnju, na rangiranju tvrdnji na temelju njihove važnosti ili na temelju detekcije glasina koje određuju relevantnost tvrdnje. Nepotvrđene priče ili izjave koje „kruže“, odnosno glasine,

imaju velik utjecaj na protok informacija na društvenim mrežama. Njihova detekcija podrazumijeva procjenu subjektivnosti jezika i širenja informacije, a taj se proces najčešće odvija na način da sustav prikuplja niz objava s društvenih mreža, a zatim klasifikator binarnim odabirom određuje koje su objave glasine. Faktori koji mogu utjecati na odabir klasifikatora su metapodaci poput broja lajkova i dijeljenja (Guo i ostali, 2022).

4.3.2. Prikupljanje dokaza

Za drugu fazu provjere informacija važno je pronaći i informacije koje okružuju zadanu tvrdnju, odnosno dokaze. Tu može biti riječ o tekstovima, tablicama, slikama, bazama podataka i metapodacima koji na neki način mogu ukazivati na vjerodostojnost tvrdnje. Pokazalo se kako oslanjanje isključivo na površinske uzorke tvrdnji, bez uzimanja šireg konteksta u obzir, u određenim slučajevima rezultira time da se dezinformacija propusti. Razlog tome može biti i velika doza kreativnosti u osmišljavanju informacije, koja ne mora nužno biti produkt ljudske kreativnosti, već se može raditi i o informaciji generiranoj računalom. Štoviše, tekstovi generirani na taj način u pojedinim su se slučajevima pokazali od većeg kredibiliteta negoli tekstovi koje su osmislili ljudi. Iz tih je razloga od velike važnosti priložiti dokaze kako bi se pred korisnicima tvrdnja opravdala kao istinita ili lažna (Guo i ostali, 2022). Autori ukazuju na detekciju stava (engl. *stance detection*) kao na instatizaciju prikupljanja dokaza. Ona podrazumijeva manje potencijalnih dokaza te predviđa svoj stav prema tvrdnji. Međutim, i u ovoj se fazi javlja problem po pitanju vjerodostojnosti informacija koje mogu poslužiti kao dokazi. Pod pretpostavkom da se radi o izvorima s kredibilitetom, većina metoda provjere činjenica kreće se enciklopedijama ili rezultatima dobivenim pretraživanjem tražilica. U ovom kontekstu Guo i ostali (2022) definiraju dokaz kao „informaciju koja može biti dohvaćena iz ovog izvora“, a vjerodostojnost kao „koherentnost dokaza“. Kada je riječ o praktičnoj primjeni, do dokaza se dolazi radom novinara, automatiziranim procesom ili kombinacijom oboje. Za primjer daju FullFact koji kao dokaze prilaže tablice i pravne dokumente vladinih organizacija (Guo i ostali, 2022). Razlog zašto često dolazi do kombinacije rada novinara i automatiziranog procesa prikupljanja dokaza jest, ponovno, ušteda vremena. Primjerice, ako se radi o iznimno velikim dokumentima, dokumentima na stranom jeziku (kojeg novinar ne razumije) ili audio-vizualnim materijalima, automatska transkripcija, sažimanje, prijevod i pretraživanje koje automatizacija pruža, čine prikupljanje dokaza mogućim za novinare (Nakov i ostali, 2021).

4.3.3. Verifikacija tvrdnje

Posljednja faza provjere informacija jest verifikacija tvrdnje kod koje se kao najčešći i najjednostavniji pristup koristi binarna klasifikacija, odnosno označavanje tvrdnje kao točne ili netočne. U slučajevima kada je prethodna faza zadovoljena, odnosno prikupljeni dokaz je korišten prilikom verifikacije, poželjno je umjesto točno/netočno navesti potkrijepljeno/opovrgnuto (dokazom) (Guo i ostali, 2022). Međutim, Nakov i ostali (2021) ističu kako neke tvrdnje ne moraju nužno biti ili točne ili netočne. U pojedinim slučajevima može se dogoditi da je određena tvrdnja točna, no svejedno može čitatelja/gledatelja navesti na krivi trag ako je dana izvan konteksta. Ovdje je važna uloga „fact-checkera“ jer je njihov zadatak ne samo verificirati tvrdnju, već omogućiti korisnicima da na pravi način razumiju sadržaj. Ukoliko se radi o kompliciranom i potencijalno dvosmislenom sadržaju, binarna klasifikacija može se pokazati kao neadekvatna. Guo i ostali (2022) pojašnjavaju kako binarna klasifikacija, odnosno označavanje neke tvrdnje kao točne ili netočne često nije dovoljno uvjerljivo za korisnike. Štoviše, može dovesti do daljnjeg uvjerenja da je originalna, netočna tvrdnja točna.

4.4. AFC danas

Najveći uspjeh automatizirane provjere informacija dosad se pokazao prilikom izdvajanja činjeničnih tvrdnji iz: tekstova, govora, članka te objava na internetu. Kao grane umjetne inteligencije koje najviše pridonose provjeri činjenica, potrebno je izdvojiti strojno učenje i obradu prirodnog jezika čijom je kombinacijom moguće identificirati i rangirati tvrdnje koje zahtijevaju provjeru. Štoviše, sve je više fact-checking ogranaka širom cijelog svijeta koji u svoj rad uključuju i softvere čiji je zadatak pomoći analitičarima da uoče tvrdnje koje je potrebno provjeriti (Graves, 2018). Među izazovima, na koje je u početnim stadijima razvoj AFC-a naišao, izdvaja se nedostatak podataka na kojima su modeli uopće mogli biti obučeni i učiti. Važno je bilo dakako omogućiti modelima da rade sa što većim količinama podataka, no iznimno je bitno da su ti podaci i kvalitetni kako bi krajnji rezultat bili učinkoviti algoritmi (Santos, 2023).

Kako bi došle do zaključaka, profesionalne fact-checking organizacije moraju uzeti u obzir kontekst i prikupljene dokaze te primijeniti kritičko mišljenje. Nerijetko su tvrdnje koje se analiziraju kompleksne do razine gdje ne mogu lako biti klasificirane kao točne ili netočne,

štoviše, takve su tvrdnje često uspoređivane s rezultatima drugih „fact-checkera“ ili se o njima konzultira s mjerodavnim izvorima. Također, moguć je i izvod vjerodostojnosti iz sekundarnih pokazatelja. Zasad se kao najučinkovitiji pristup pokazalo uspoređivanje tvrdnji s bazom podataka prethodno provjerenih tvrdnji. Na taj način analitičari mogu raditi na složenijim prosudbama koristeći automatizaciju u svrhu poboljšanja vlastite učinkovitosti (Graves, 2018). Evidentno je da potreba za čim većim uključenjem umjetne inteligencije u proces provjere informacija raste što se odražava i na veću uporabu automatiziranih alata poput strojnog učenja koje se koristi kod modela dubokog učenja namijenjenih detekciji lažnih vijesti. Druge metode identifikacije dezinformacija mogu biti komponente sustava za provjeru informacija ili mogu služiti i kao samostalna rješenja. Osim toga, AI sustavi preporuke mogu imati ključnu ulogu kada se radi o nadziranju društvenih mreža u svrhu smanjenja distribucije dezinformacija (Santos, 2023).

Razvoj i uporaba umjetne inteligencije pokazali su se kao obećavajući, a na temelju trenutnog stanja u budućnosti je moguće očekivati napredak po pitanju algoritama, rješavanja kompleksnih zadataka, smanjenju pristranosti i detekciji dezinformacija. Važno je da umjetna inteligencija sudjeluje i u verifikaciji informacija na društvenim mrežama, na način da u stvarnom vremenu obavijesti korisnike o pojedinoj dezinformaciji te ih potakne na kritičko razmišljanje prilikom korištenja društvenih mreža („AI-Driven Recruitment Trends #13 | AI's Crucial Role in Fact-Checking and Misinformation“, 2023). Iako se naziru brojne mogućnosti koje umjetna inteligencija nudi u području provjere informacija, vjerojatnost da bi takvi sustavi u bliskoj budućnosti mogli zamijeniti ljude koji se time bave jest iznimno mala. Prednost koju ljudi posjeduju jest dublje razumijevanje konteksta, tona i nijansi te namjere, što su iznimno bitni faktori prilikom provjere tvrdnji („Human vs AI Fact-Checkers: A Comparative“, 2024). K tome, važno je osvijestiti da „ground truth“ koju su prilikom vježbanja utvrdili algoritmi nije univerzalna niti trajna (Graves, 2018). Umjesto da fokus bude na zamjeni ljudskih ruku umjetnom inteligencijom, pažnju je potrebno skrenuti na potencijal koji ima njihov zajednički rad, odnosno iskorištavanje njihovih prednosti. Činjenica je da bi kombinacija ljudskog uma i umjetne inteligencije bila učinkovita u provjeri informacija. Primjerice, dok analitičari mogu verificirati informacije uzimajući u obzir kontekst, ton i namjeru, umjetna inteligencija prepoznaje moguće dezinformacije. Cilj ove suradnje jest veća točnost i pouzdanost procesa provjere informacija („Human vs AI Fact-Checkers: A Comparative“, 2024).

4.4.1. Prednosti i nedostaci

Članak „Human vs AI Fact-Checkers: A Comparative“ (2024) govori o ograničenjima, prednostima i nedostacima implementacije umjetne inteligencije u provjeri informacija. Kao primjer daje LongShot Fact Checker koji služi kao AI alat koji analizira najnovije informacije prikupljene na internetu, detektira tvrdnje i verificira ih te korisniku pruža prijedloge ispravnih tvrdnji koje mogu zamijeniti pronađene neistinite tvrdnje. S implementacijom umjetne inteligencije u područje provjere informacija, dolaze značajne prednosti kao i nedostaci. U članku su izdvojene sljedeće prednosti i nedostaci:

Prednosti automatizirane provjere informacija su:

1. Brzina i efikasnost – umjetna inteligencija omogućava brži i točniji pregled većih količina informacija, što utječe i na bržu verifikaciju činjenica. Važnost ove karakteristike posebno je naglašena kada se radi o „breaking news“ u sklopu kojih se mnogo informacija brzo stvara i širi.
2. Dosljednost i nepristranost – ono što AI sustave razlikuje od ljudi jest to što nisu pristrani i nemaju emocije koje bi potencijalno utjecale na njihovo donošenje odluka, odnosno pogled na pojedine tvrdnje koje su u procesu provjere. Zahvaljujući tome, sustavi provjeravaju tvrdnje objektivnim pristupom i bez ikakvih vanjskih utjecaja.
3. Rukovanje velikim količinama podataka – ljudski rad pod utjecajem je fizičkog stanja i energije osobe, dok kod umjetne inteligencije takva ograničenja ne stvaraju problem. Automatizirani sustavi provjere informacija mogu obraditi veće količine podataka djelotvornije i učinkovitije što olakšava upravljanje procesom provjere.

Nedostaci automatizirane provjere informacija su:

1. Nedostatak razumijevanja konteksta – programiranje sustava koji razumiju kontekst prilikom analiziranja zasebnih informacija velik je izazov. Problem predstavljaju nijanse u jeziku te dvosmislenost pojedinih riječi i izraza ovisno o kontekstu, što može dovesti sustav do pogrešne analize i verifikacije.
2. Potencijalna pristranost – iako su AI sustavi programirani kao nepristrani, može se dogoditi da podaci korišteni u svrhu obuke AI programa budu pristrani ili da algoritam nije osmišljen na pravi način. Pristranost može uzrokovati pogrešnu analizu i verifikaciju.
3. Ograničena sposobnost analize tona i namjere – AI sustavi osmišljeni su da identificiraju činjenične greške, no kada se radi o tvrdnjama koje čovjek shvaća kao

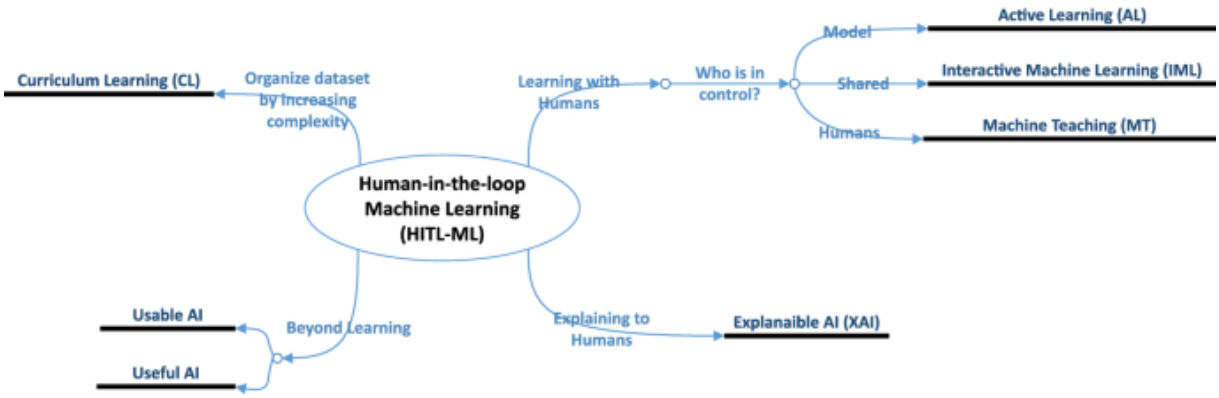
sarkastične ili ironične, sustav ih ne prepoznaje kao takve. U ovakvim slučajevima sustav nerijetko pogrešno analizira i verificira informacije.

5. Human-in-the-Loop (HITL)

Koncept „Human-in-the-Loop“ (HITL) dobiva na važnosti u pogledu istraživanja integracije ljudskog znanja i iskustva sa strojnim učenjem (ML). Činjenica jest da strojno učenje usprkos razvoju i mogućnostima ne može u potpunosti zamijeniti ljudski um. Fokus HITL je na tome da se stvore točni modeli predviđanja uključenjem i ljudskog doprinosa i automatiziranih sustava. Ljudi ponajprije mogu doprinijeti radu strojnog učenja na način da ga opskrbe podacima na kojima može vježbati te asistencijom u rješavanju kompleksnih zadataka koji za same strojeve predstavljaju izazov (Wu, Xiao, Sun, Zhang, Ma i He, 2022). Tradicionalno, razvoj modela strojnog učenja donekle je krut proces – dizajnirani su, testirani i zatim uvedeni bez naknadnih izmjena. Ovakav pristup može rezultirati modelima koji ne skaliraju dobro, postaju zastarjeli ili gube učinkovitost ovisno o promjenama u kontekstu implementacije. Štoviše, trenutne tehnike strojnog učenja još nisu osposobljene za napredno logičko razmišljanje i razumijevanje uzročno-posljedičnih veza. Novija istraživanja kao cilj imaju uspostavljanje novih oblika interakcije između ljudi i ML sustava, što se skupno naziva Human-in-the-loop machine learning (HITL-ML). Nakana ovog pristupa jest poboljšanje točnosti i brzine kojom strojno učenje procesira informacije te unaprjeđenje ljudske učinkovitosti i djelotvornosti u korištenju takvih sustava (Mosqueira-Rey, Hernández-Pereira, Alonso-Ríos, Bobes-Bascarán i Fernández-Leal, 2023).

Holmberg (2020) ističe tri različita pristupa HITL-ML-u, ovisno o tome tko je u kontroli procesa učenja:

1. Aktivno učenje (Active Learning) – sustav ostaje u kontroli procesa učenja, a ljude koristi kao savjetnike za označavanje neoznačenih podataka (Holmberg, 2020).
2. Interaktivno strojno učenje (Interactive Machine Learning) – naglasak je na interakciji korisnika i sustava učenja. Ljudi opskrbljuju sustav informacijama na fokusiraniji i konzistentniji način u usporedbi s tradicionalnim strojnim učenjem (Mosqueira-Rey i ostali, 2023).
3. Strojno podučavanje (Machine Teaching) – ljudi koji su analitičari u određenom području imaju kontrolu u procesu učenja ograničavanjem znanja koje žele prenijeti u model strojnog učenja (Holmberg, 2020).



Slika 2. *Pristupi HITL-ML-u* (2023) Preuzeto 10.08.2024. sa https://link.springer.com/article/10.1007/s10462-022-10246-w

Mosqueira-Rey i ostali (2023) ističu da ljudi imaju važnu ulogu u procesu učenja i u situacijama kada oni nisu ti u kontroli. Kao primjer daje činjenicu da je ljudsko učenje utjecalo na oblikovanje različitih algoritama u kontekstu razvoja strojnog učenja. Kao važnu karakteristiku ljudskog učenja, potrebno je istaknuti kurikulum, odnosno učenje smislenim redoslijedom. Takav se tip učenja otisnuo i na strojno učenje te je time stvorena poddisciplina pod nazivom učenje kurikuluma (engl. *curriculum learning*, CL). CL podrazumijeva nametanje neke strukture skupu za obuku u svrhu ubrzanja i poboljšanja učenja. Mosqueira-Rey i ostali (2023) podsjećaju na važnost toga da algoritam objasni svoje odluke i rezultate ljudima. Drugim riječima, sposobnost algoritma da riješi problem nije dovoljna ukoliko on ne uspijeva pružiti objašnjenje zašto je odabrao neko rješenje. Ovdje se radi o objašnjivoj umjetnoj inteligenciji (XAI), polju koje kao zadatak ima učiniti rezultate umjetne inteligencije razumljive ljudima.

5.1. Primjena HITL u verifikaciji informacija

Uključivanje analitičara kao „human-in-the-loop“ (HITL) unutar standardnih okvira strojnog učenja, kao što su analiza mišljenja, analiza mišljenja na temelju aspekata i detekcija stava, može učiniti ove metode održivima za djelomično automatizirani sustav detekcije dezinformacija (Daniel, 2021). Kada se u obzir uzmu ograničenja i umjetne inteligencije i ljudi, javlja se ideja o kombinaciji ljudi i umjetne inteligencije kao rješenja u borbi protiv dezinformacija. Upravo je to poanta HITL sustava. Oni spajaju prednosti koje pruža svaka od strana – ljudi filtriraju ono na čemu će se raditi dok strojevi izvršavaju zadatke koji se tiču baza podataka i algoritama strojnog učenja. Potencijal koji umjetna inteligencija pruža u

velikim obradama i sposobnost ljudske inteligencije u zadacima gdje umjetna inteligencija nailazi na probleme moguće je iskoristiti u sklopu HITL sustava (Santos, 2023).

Daniel (2021) u svojem radu opisuje po kojem principu funkcionira sustav koji je hibrid čovjeka i umjetne inteligencije. Za početak, čovjek identificira određenu temu ili tvrdnju za koju sumnja da sadrži dezinformacije, nakon čega počinje prikupljati što je više moguće informacija o tome te ukoliko je moguće, identificira osobu koja te informacije širi. Algoritam strojnog učenja zatim upotrebljava tu tvrdnju, odnosno temu u sklopu velike količine teksta prikupljene na internetu te koristi korisnički definirane odnose odobravanja i neodobravanja kako bi kategorizirao tekst u dvije skupine – relevantne dezinformacije i ostali sadržaj (koji podrazumijeva ispravne informacije te nevezane dezinformacije). Kako bi zatvorio petlju, čovjek naposlijetku pregleda rezultate pretraživanja da bi odredio odgovaraju li njegovim potrebama. Ako ne odgovaraju, on može doraditi kriterije pretraživanja i ponoviti proces (Daniel, 2021).

Izazovi koji stoje pred HITL sustavima i dalje su vrijedni spomena, a uključuju prepreke u postizanju optimalne funkcionalnosti i generiranju valjanih rezultata. Primjerice, rezultati koji su dobiveni mogu biti produkt nedosljednosti kvalitete podataka uzrokovane varijacijama među sudionicima koji su bili uključeni u obuku AI sustava (Santos, 2023). Kao što i sam naziv daje naznačiti, interakcije ljudi i umjetne inteligencije u HITL sustavima izrazito su kompleksne. Iako se AI sustavi generalno smatraju objektivnima, njihovo prosuđivanje ovisi o različitim faktorima, poput individualnih očekivanja, interaktivnosti sustava i tome pruža li AI informacije o svojim presudama i preporukama (DeVerna, Yan, Yang i Menczer, 2024). Unatoč ograničenjima, važno je da se od razvoja i rada na ovakvim sustavima ne odustane već da se oni razviju u učinkovite i kvalitetne kontrolne mehanizme za „nabavu iz mnoštva“ (engl. *crowdsourcing*) kako bi se postigle velike količine kvalitetnih oznaka. Problem o kojem se dosta istražuje i raspravlja jest potencijalna pristranost sustava koju ljudi mogu unijeti i/ili pojačati. S obzirom na to da se predrasude i stereotipi anotatora mogu odraziti na generirane oznake, ta pristranost može utjecati na način rada modela koji su obučeni na tim označenim podacima (Santos, 2023).

5.2. Out-of-the-Loop

Kertysova (2018) je u svojem radu opisala i koncept suprotan HITL, pod nazivom „humans out of the loop“ AI sustava. Ovaj pristup podrazumijeva minimalnu uključenost ljudi u proces

donošenja odluke AI sustava. Takva se metoda obično upotrebljava u sustavima koji su obučeni na velikim skupovima podataka te koriste napredne algoritme za samostalno donošenje odluka. Drugim riječima, njihov je rad u potpunosti bez direktnog ljudskog unosa.

6. Detekcija deepfake slika, videozapisa i audiozapisa

Svakim danom važnost razvoja područja umjetne inteligencije, strojnog učenja i dubokog učenja raste sve više. Razvijaju se nove metode i alati koji mogu oblikovati, mijenjati i analizirati multimedije. Iako su u početku ove tehnologije korištene u legitimne svrhe poput zabave i obrazovanja, danas se one nažalost sve više odmiču od toga i bivaju korištene u ilegalne svrhe (Rana, Nobi, Murali i Sung, 2022). Primjerice, realistični lažni videozapisi, slike ili audiozapisi, koji se često nazivaju „deepfakes“, stvoreni su s namjerom širenja dezinformacija u svrhu poticanja političkih nemira, uznemirivanja ili ucjenjivanja. U početku korišteni u filmskoj industriji, deepfakeovi sada su primarno usmjereni prema zabavi na internetu, obmani potrošača te političkim i međunarodnim sferama (Kertysova, 2018). Potencijal koji deepfake ima da dovede u pitanje koncepte istine i autentičnosti jest značajan jer mogu stvoriti uvjerljive lažne prikaze ljudi te na taj način predstavljaju ozbiljnu prijetnju identitetu i privatnosti. Na temelju istraživanja koja su provedena moguće je izvući zaključak da se sposobnost deepfakeova da repliciraju identitete proteže i izvan vizualnih prikaza (Oladoyinbo, Olabanji, Olaniyi, Adebiyi, Okunleye i Ismaila Alao, 2024). Kao što je već konstatirano, deepfake videozapisi koji su generirani algoritmima dubokog učenja sve su aktualnija tema vrijedna proučavanja. Deepfake tehnologija može rezultirati iznimno realističnim manipulacijama lica na videozapisu ili slici. Iako algoritam sam po sebi nije ni dobronamjeran ni zlonamjeran, pokazalo se kako se ova tehnologija nerijetko koristi u loše svrhe (Yu, Xia, Fei i Lu, 2021).

Sam termin „deepfake“ potječe od termina „Deep Learning“ (duboko učenje) i „Fake“ (lažan), a opisuje realistične videozapise ili slike koje su generirane pomoću dubokog učenja. Termin je uveden u upotrebu 2017. godine kada je korisnik na Reddit-u pomoću metoda strojnog učenja zamijenio lice jedne osobe licem druge osobe u pornografskom videozapisu te na taj način stvorio realistične lažne videozapise. Tehnologija koja je iza deepfakea podrazumijeva korištenje dviju neuronskih mreža – generativne mreže koja proizvodi lažne slike i diskriminirajuće mreže koja procjenjuje njihovu autentičnost, zajednički poznatih kao Generative Adversarial Networks (GAN). Početkom 2018. godine, u siječnju, razne web-stranice počele su nuditi usluge stvaranja deepfakeova s potporom privatnih sponzora. Mjesec dana kasnije, velik je broj web-stranica, među kojima su bili i Gfycat, Pornhub i Twitter, zabranili takve usluge (Rana i ostali, 2022). U lipnju 2019. godine, stvorena je i aplikacija Deepnude koja je pokrenula opću paniku zbog mogućnosti „razodijevanja“ ljudi. Videozapisi stvoreni u ovoj aplikaciji nanijeli su štetu po pitanju osobne privatnosti te utjecali na političke

kampanje i javno mišljenje (Yu i ostali, 2021). S obzirom na to da prijetnju koju deepfake predstavlja za javnost, istraživanja usmjerena na distribuciju i detekciju deepfakea sve su brojnija. Rossler i ostali stoje iza alata za otkrivanje skupa video-podataka namijenjenih obuci medijske forenzike deepfakea pod nazivom FaceForensic. Nakon toga, istraživači Sveučilišta u Stanfordu objavili su „Deep video portraits“, metodu koja omogućuje realistične animacije portretnih videozapisa. Istraživači UC Berkeley zaslužni su za razvoj novog pristupa za prijenos pokreta tijela jedne osobe na drugu osobu u videozapisu (Rana i ostali, 2022).

Kertysova (2018) skreće pažnju na činjenicu da će netko uskoro moći stvoriti veoma uvjerljiv lažni videozapis nekoga sve dok ima pristup i može locirati visokokvalitetne slike ili zvuk za tu osobu. Ono što dodatno zabrinjava jest to da će razvoj dodatno zamagliti granicu između autentične multimedije i medija generiranog umjetnom inteligencijom, čineći strojeve sve manje vještima u razlikovanju te dvije vrste, ali i ljude (Kertysova, 2018). Deepfake pornografija samo je jedna od negativnih posljedica koje proizlaze iz deepfakea, a postoji i mnogo drugih nezakonitih ili zlonamjernih aktivnosti u koje se netko može upustiti kao što je širenje lažnih vijesti, poticanje državnog udara i sl., a sve spadaju u neetičke prakse. Ovo je jedan od glavnih povoda stvaranju polja detekcije deepfakea (Rana i ostali, 2022).

6.1. Detekcija sintetički stvorenih ili manipuliranih slika

Juršėnas i ostali (2022) pišu o detekciji sintetički stvorenih ili manipuliranih slika, videozapisa i audiozapisa. Što se tiče detekcije manipuliranih slika, autori ističu činjenicu da generativni modeli postaju sve raznovrsniji te da se kao jedan od najvećih izazova pokazuje suočavanje s lažnim slikama i videozapisima. FaceForensics++ jest skup podataka stvoren s namjerom da riješi probleme po pitanju otkrivanja manipuliranih slika. Cijeli skup podataka čine slike iz tisuću različitih videozapisa. Zatim iste te slike bivaju manipulirane popularnim tehnikama krivotvorenja kao što su zamjena lica ili rekonstrukcija lica. Skup podataka također uzima u obzir i kvalitetu kompresije H.264 koja se obično koristi u objavama na društvenim medijima. AI sustavi obučeni su da detektiraju i manipulirane slike vrlo niske kvalitete (uslijed primjerice velike kompresije), za razliku od ljudi koje takve slike mogu zavarati. FaceForensics++ sadrži unaprijed obučene modele koji omogućavaju transferno učenje, što je od iznimne važnosti s obzirom na to da se stalno pojavljuju nove tehnologije manipulacije koje bi potencijalno mogle poraziti postojeće detektore (Juršėnas i ostali, 2022).

Stariji deepfakeovi, poput onih s thispersondoesnotexist.com, često sadrže vidljive greške (izobličenja pozadine ili nejednake naušnice i zubi (ljudi brzo uočavaju takve probleme ručnom inspekcijom slike)) koje ih čine prepoznatljivima. Jedan od najpreciznijih detektora je Expectation Maximization (EM), algoritam koji detektira konvolucione tragove koji ostaju na slikama, slično kao i otisci prstiju na fizičkim fotografijama. On može prepoznati lažne slike stvorene različitim modelima, kao što su: AttGAN, GDWCT, StarGAN, StyleGAN i StyleGAN2, te to radi s 99.81% točnosti (Juršėnas i ostali, 2022).

6.2. Detekcija sintetički stvorenog ili manipuliranog videozapisa

Deepfake videozapisi sve su veća prijetnja osobnoj privatnosti, ali i sigurnosti te se iz tog razloga radi na stvaranju što više različitih metoda detekcije. Za razliku od prijašnjih projekata gdje je fokus bio na analizi problematike sinteze lica, novije metode fokusiraju se na ključne aspekte sličnosti lica. Nedavno su uvedena mrežna rješenja za zaustavljanje generiranja deepfakeova, poput uobičajenih zadataka klasifikacije. Drugi pristup koristi vremensku dosljednost za detekciju pronalaženjem aktivnosti koja je ostala otvorena između kadrova u deepfakeovima. K tome, često se moguće osloniti na vizualne artefakte koji se pojavljuju u procesu miješanja i daju naznačiti da je nešto lažno. Neke novije metode koje su se pokazale obećavajućima usmjerene su na karakteristike niske razine, poput otisaka prstiju kamere i biometrijska mjerenja (Yu i ostali, 2021). Prilikom testiranja na nekim vrstama falsificiranih videozapisa, poput onih kreiranih metodom Face2Face, stope otkrivanja uglavnom su bile ispod onoga što statističari nazivaju pragom značajnosti – što pokazuje da su pogrešni odabiri tek malo bolji od nasumičnih nagađanja. Iako ove vrste videozapisa ne moraju nužno izazivati sumnje među gledateljima, moguće je da mnoge metode strojnog učenja mogu otkriti i najsuptilnije naznake (Juršėnas i ostali, 2022).

Nedavni napredak po pitanju klasifikacije slika primijenjen je kako bi se unaprijedila detekcija deepfake videozapisa. U ovom pristupu, slike lica iz videozapisa koriste se za obuku detekcijske mreže, koja zatim predviđa autentičnost svih kadrova. Konačni rezultat određen je strategijama izračunavanja prosjeka ili glasanja, zbog čega točnost detekcije uvelike ovisi o neuronskim mrežama bez potrebe za posebnim razlikovnim obilježjima (Yu i ostali, 2021). Prema metodi iza koje stoje Yang, Li i Lyu, moguće je otkriti deepfake videozapise na temelju identifikacije nedosljednosti u predviđenim 3D-orijentacijama lica, s obzirom na to da deepfake modeli nerijetko ne uspijevaju zadržati ispravne položaje dijelova

lica kada transformiraju jedno lice u drugo. K tome, važan aspekt je vrijeme. Neusklađeni pokreti usta ili neprirodni obrasci treptanja također mogu doprinijeti identifikaciji manipuliranih videozapisa, čak i kada pojedinačni kadrovi izgledaju realistično (Juršėnas i ostali, 2022).

6.3. Detekcija sintetički stvorenog audiozapisa

U zadnje vrijeme, kloniranje glasa uvelike je napredovalo pomoću metoda poput konverzije glasa (promjene glasa jedne osobe da bi zvučao kao glas druge) i text-to-speech (TTS) generiranja (stvaranje govora koji oponaša zadani glasi iz tekstualnog unosa). Ove metode, koje se uvelike oslanjaju na modele dubinskog učenja, mogu se svrstati pod zajednički naziv „deepfake voice technology“. U primjenama koje uključuju automatsko prepoznavanje glasa (ASV), autentifikacija služi kao oblik biometrijske identifikacije. Modeli dubinskog učenja koriste se za detekciju sintetički stvorenog govora. Jedna tehnika uključuje konverziju audiozapisa u spektrogram – vizualna reprezentacija koja pokazuje distribuciju zvučne frekvencije tijekom vremena – što problem detekcije pretvara u problem klasifikacije slike. Određeni tim spektrograma, Melspectrogram, koristi se prilikom vizualne analize audiozapisa (Juršėnas i ostali, 2022).

Iako je u području detekcije vizualnih deepfakea došlo do značajnog napretka, na audio-deepfakeove (poput onih generiranih pomoću TTS-a ili konverzijom glasa) obraćeno je manje pažnje. Iz tog razloga novi pristup predlaže otkrivanje deepfakeova analizom sinkronizacije između vizualnih i slušnih elemenata. Na primjer, neusklađenost između pokreta usana i izgovorenih slogova može ukazivati na manipulaciju. Ova metoda uključuje modeliranje videozapisa i audiozapisa odvojeno, usklađivanje njihovih prikaza i dodjeljivanje oznake sinkronizacije koja odražava potencijalnu manipulaciju. Međutim, nedostatak skupova podataka koji sadrže i vizualne i slušne manipulacije i dalje predstavlja izazov. Kako bi se taj problem zaobišao, koriste se postojeći skupovi podataka o deepfake videozapisima s nepromijenjenim zvukom, a zatim se zvukom umjetno manipulira kako bi se stvorio sveobuhvatan skup podataka za obuku modela detekcije (Zhou i Lim, 2021).

7. Izazovi i problemi

Pozitivne strane uporabe umjetne inteligencije u verifikaciji informacija i detekciji dezinformacija su brojne, no važno je uzeti u obzir i izazove koji s time dolaze. Članak „AI-Driven Recruitment Trends #13 | AI's Crucial Role in Fact-Checking and Misinformation“ (2023) ih dijeli u četiri kategorije:

1. Pristranost algoritma – iako to ne rade s namjerom, AI sustavi mogu svojim djelovanjem doprinijeti pristranosti u podacima, čime mogu dovesti do iskrivljenih rezultata provjere tvrdnji.
2. Kontekstualne nijanse – kada se radi o detekciji nijansi poput sarkazma i satire, AI često nailazi na problem i ne može razlikovati istinit od lažnog sadržaja.
3. Taktike koje se stalno razvijaju – taktike dezinformacija su u stalnom razvoju čime se javlja i potreba za kontinuiranim ažuriranjem AI modela i algoritama.
4. Etičke dileme – definiranje koji su izvori pouzdani i rješavanje oprečnih stavova na automatizirani način vodi do postavljanja etičkih pitanja.

Kao jedan od glavnih izazova koji pred nas stavlja uporaba umjetne inteligencije i strojnog učenja prilikom otkrivanja dezinformacija jest briga o etičkim pitanjima, posebno pristranosti (Santos, 2023). Manipulacija podacima može biti namjerna ili slučajna (naslijeđena algoritmom) te kao takva postavlja ozbiljna pitanja vezana uz pouzdanost i pravednost odluka koje je donijela umjetna inteligencija. Takva pristranost najčešće potječe iz podataka na kojima je model obučen ili predispozicija njegovih kreatora. Pristranost, a i predrasude koje s njom dolaze, mogu odražavati društvene razlike i nepravde. Time se dovodi u pitanje ne samo etički okvir unutar kojeg ti sustavi funkcioniraju, već i šire društvene implikacije njihove primjene (Oladoyinbo i ostali, 2024). Na ovaj je problem važno skrenuti pažnju jer pristranost i predrasude koji mogu biti prisutni mogu se infiltrirati u sustave te time dovesti do pogrešnih ishoda. Do prepreka u korištenju umjetne inteligencije zbog pitanja etike može doći primjerice u novinarstvu, po pitanju nedostatka nadzora i transparentnosti pa i potencijalnog potiskivanja kreativnosti (Santos, 2023). Za pojedine automatizirane algoritme postoji rizik i od repliciranja. Drugim riječima, automatizacija ljudske pristranosti i osobnosti rezultira nepovoljnijim ishodima za pojedince unutar određene skupine. Razlog tome što ovaj problem uopće i postoji jest činjenica da pristranost algoritma može biti rezultat vrijednosti i prioriteta programera koji ih stvaraju, dizajniraju i obučavaju ili pak iz nepotpunih, neispravnih ili nereprezentativnih podataka koji služe za obuku algoritma (Kertysova, 2018).

Spangenberg („Roundtable: How artificial intelligence is used to debunk fake news“, 2020) skreće pažnju i na korisničku stranu. U današnje vrijeme ljudima je na raspolaganju velik broj alata (za geolokaciju, prepoznavanje i sl.), među kojima je većina besplatna, a mogu korisnicima olakšati provjeru informacija. Kada su okolnosti takve, korisnici su rijetko svjesni da alati koje koriste mogu biti pristrani. Činjenica je da su internetska veza i računalo dovoljni da bi netko postao fact-checker („Roundtable: How artificial intelligence is used to debunk fake news“, 2020). Ukoliko se ovom problemu ne stane na kraj, ili ga se barem ima pod kontrolom, pristranost algoritama može dovesti do odluka koje imaju različit utjecaj na pojedine skupine ljudi. Primjerice, ako je riječ o algoritmima koji su uvježbani na povijesnim skupovima podataka koji su bogati predrasudama prema ženama, a računalo ih koristi za obuku, ono pada pod utjecaj svjetonazora iz tih podataka. Kertysova (2018.) podsjeća kako je sporna tema je li moguće da se umjetna inteligencija u potpunosti oslobodi ljudskih pogrešaka i ega.

U svojem se radu Santos (2023) referira na Kertysovu (2018) koja ističe izazov koji automatizirane tehnologije susreću prilikom evaluacije kompleksnih tvrdnji. Premda su trenutni AI sustavi iskusni u prepoznavanju direktnih i nedvosmislenih tvrdnji, nijansirani izrazi gdje je važno uzeti kontekst i kulturne čimbenike u obzir predstavljaju problem. Napredak u obradi prirodnog jezika (NLP) za automatiziranu analizu teksta jest znatan, no suptilni ljudski koncepti poput ironije i sarkazma ponekad su preveliki pothvat. Iz tog je razloga AI sustavima teže baratati dezinformacijama koje se oslanjaju na implicitnije forme izraza. Kao važan korak prema rješavanju ovog problema Santos (2023) ističe uključivanje ljudi u proces analize, posebice u obuku algoritama strojnog učenja. Kao primjer možemo uzeti projekte usmjerene na klasifikaciju lažnih vijesti, gdje ljudi u prethodnom koraku imaju zadatak prepoznati je li dana vijest istinita ili lažna. Iz toga AI sustav može učiti i ravnati se po ljudskim unosima kao po uzorku koji može prepoznati i na temelju njega unaprijediti svoju klasifikaciju. Ova se metoda naziva „semi-supervised learning“, a obuhvaća ljudsko razumijevanje jezične suptilnosti s učinkovitošću i prilagodljivosti automatiziranih sustava, što na kraju rezultira točnijom i preciznijom analizom teksta (Santos, 2023).

Kertysova (2018) naglašava da automatizirane tehnike za detekciju i suzbijanje dezinformacija nailaze na više ograničenja. Primjerice, može doći do „over-blocking“ legitimnog sadržaja zbog tendencije umjetne inteligencije da bude pretjerano inkluzivna. AI modeli koji su trenutno u razvoju mogu producirati i lažne pozitivne i negativne rezultate, na način da pogrešno prepoznaju pravi sadržaj kao dezinformaciju. Takav postupak može

rezultirati cenzurom, imajući negativan utjecaj na slobodu izražavanja. Drugi izazov s kojim se je potrebno suočiti jest složenost AI sustava, posebno onih koji uključuju neuronske mreže i dubinsko učenje, odnosno onih koji rade kao „rješenja crne kutije“ (engl. *black box solutions*). Kompleksnosti doprinosi i činjenica da nerijetko ni sami programeri koji su kreirali te sustave ne razumiju u potpunosti njihove procese donošenja odluka čime se otežava i objašnjavanje na koji se način preporuke generiraju. Dok neke tvrtke i projekti, kao što je DARPA-in Explainable AI Program, rade na stvaranju transparentnijih i odgovornijih AI sustava, ta su rješenja još uvijek u eksperimentalnoj fazi i još nisu široko dostupna. Neosporivo je da i automatizirane i ljudske metode verifikacije imaju svoje nedostatke. Ljudi često rade pod stresom te su ograničeni kratkim rokovima, čime se izlažu riziku razvoja simptoma sličnih PTSP-u. Osim toga, ljudski rad je skup i sklon pogreškama, s osobnim predrasudama i raspoloženjem koji utječu na analizu sadržaja (Kertysova, 2018).

Rješenja umjetne inteligencije za otkrivanje i uklanjanje ilegalnog ili nepoželjnog sadržaja postaju sve naprednija, ali i navode na postavljanje pitanja oko toga tko ima ovlasti odlučivati što se smatra legalnim ili poželjnim. Uporaba umjetne inteligencije u ovom kontekstu podrazumijeva balansiranje između legalnih i tehnoloških čimbenika, javnih i privatnih interesa te kompromise između sudskog nadzora i skalabilnosti (Marsden i Meyer, 2019). Sve to skupa uvelike utječe i na slobodu izražavanja. Neosporno je da AI sustavi sa sobom donose mnoge pozitivne strane, no istovremeno donose i etičke dileme te potencijalne rizike po pitanju ljudskih prava i demokratskih procesa. Problematika koju su iznijeli stručnjaci tiče se pravednosti algoritma, „filter bubble“ (personalizirani sadržaj može rezultirati djelomičnom „sljepoćom na informacije“), povrede privatnosti korisnika, manipulacije korisnika te video- i audio-manipulacije bez pristanka pojedinca (Kertysova, 2018). Oladoyinbo i ostali (2024) naglašavaju da osiguravanje integriteta podataka zahtijeva ne samo održavanje točnosti i kvalitete, već i rješavanje širih etičkih pitanja, uključujući privatnost i sigurnost. Kintsurashvili („Roundtable: How artificial intelligence is used to debunk fake news“, 2020) ističe kako postoji i zabrinutost zbog zlouporabe umjetne inteligencije u tranzicijskim demokracijama i autoritarnim režimima. Tamo čak i fiktivni likovi mogu biti predstavljeni kao stvarne ličnosti u javnoj komunikaciji. Sve to skupa naglašava potrebu za opreznom provjerom činjenica i novinarskim nadzorom kako bi se održao korak s trendovima.

8. Zaključak

U ovom radu analizirana je uloga umjetne inteligencije u procesu verifikacije informacija i detekcije dezinformacija. Umjetna inteligencija, kroz automatizirane sustave za provjeru činjenica i detekciju lažnih sadržaja, pokazala se kao moćan alat u borbi protiv sveprisutnih dezinformacija u digitalnom okruženju. Tehnologije poput strojnog učenja, dubokog učenja i Generative Adversarial Networks (GAN) igraju ključnu ulogu u razvoju naprednih sustava koji mogu prepoznati, analizirati i ukloniti lažne informacije. Međutim, uz sve prednosti, korištenje umjetne inteligencije nosi i značajne izazove, prvenstveno u pogledu etike, privatnosti i potencijalne zloupotrebe u različitim političkim i društvenim kontekstima.

Jedan od najvećih izazova ostaje balansiranje između efikasnosti AI sustava i potrebe za ljudskim nadzorom kako bi se izbjegle pogreške i nepoželjni ishodi. „Human-in-the-loop“ pristup predstavlja jedno od mogućih rješenja za ovaj problem, gdje se ljudski faktor uključuje u ključnim fazama procesa verifikacije kako bi se osigurala točnost i pravednost rezultata.

Razvoj tehnologija za detekciju deepfake sadržaja, koji predstavljaju ozbiljnu prijetnju autentičnosti i integritetu informacija, također je od izuzetne važnosti. Iako je tehnologija deepfakeova izvorno razvijena u legitimne svrhe, njezina sve češća upotreba za stvaranje obmanjujućih i štetnih sadržaja naglašava potrebu za daljnjim istraživanjima i razvojem alata za njihovo prepoznavanje.

Zaključno, iako umjetna inteligencija nudi značajne mogućnosti za unaprjeđenje procesa verifikacije informacija, ključno je kontinuirano pratiti i uzimati u obzir etičke i društvene implikacije kako bi se osiguralo da njezina primjena bude u svrhu istine i zaštite ljudskih prava.

9. Literatura

AI-Driven Recruitment Trends #13 | AI's Crucial Role in Fact-Checking and Misinformation (2023). Preuzeto 10.08.2024. s https://www.linkedin.com/pulse/top-tech-trends-13-ais-crucial-role-fact-checking/

Daniel, A. M. (2021). Human-in-the-Loop Disinformation Detection: Stance, Sentiment, or Something Else? Preuzeto 10.08.2024. s https://arxiv.org/pdf/2111.05139

DeVerna, M. R., Yan, H. Y., Yang, K. i Menczer, F. (2024). Fact-checking information from large language models can decrease headline discernment. Preuzeto 10.08.2024. s https://arxiv.org/abs/2308.10800

Graves, L. (2018). Understanding the Promise and Limits of Automated Fact-Checking. Preuzeto 10.08.2024. s https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2018-02/graves_factsheet_180226%20FINAL.pdf

Grmuša, T. i Prelog, L. (2020). Uloga novih tehnologija u borbi protiv lažnih vijesti – iskustva i izazovi hrvatskih medijskih organizacija. *Medijske studije*, 11(22), 62-80. Preuzeto 10.08.2024. s https://hrcak.srce.hr/253377

Guo, Z., Schlichtkrull, M. i Vlachos, A. (2022). A Survey on Automated Fact-Checking. *Transactions of the Association for Computational Linguistics*, 10, 178–206. Preuzeto 10.08.2024. s https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00454/109469/A-Survey-on-Automated-Fact-Checking

Gutiérrez-Caneda, B. i Vázquez-Herrero, J. (2024). Redrawing the Lines Against Disinformation: How AI Is Shaping the Present and Future of Fact-checking. *Tripodos*, 55. Preuzeto 10.08.2024. s https://tripodos.com/index.php/Facultat_Comunicacio_Blanquerna/article/view/1110/1169

Holmberg, L. (2020). A Feature Space Focus in Machine Teaching. Preuzeto 10.08.2024. s https://www.diva-portal.org/smash/get/diva2:1428195/FULLTEXT01.pdf

Hrvatska enciklopedija, mrežno izdanje. Leksikografski zavod Miroslav Krleža. Preuzeto 10.08.2024. s https://www.enciklopedija.hr/clanak/informacija

Human vs AI Fact-Checkers: A Comparative (2024). Preuzeto 10.08.2024. s https://www.longshot.ai/blog/can-ai-fact-checkers-replace-humans

Juršėnas, A., Karlauskas, K., Ledinauskas, E., Maskeliūnas, G., Ruseckas, J. i Rondomanskas, D. (2022). The Role of AI in the Battle Against Disinformation. Preuzeto 10.08.2024. s https://stratcomcoe.org/publications/the-role-of-ai-in-the-battle-against-disinformation/238

Kertysova, K. (2018). How AI Changes the Way Disinformation is Produced, Disseminated, and Can Be Countered. Preuzeto 10.08.2024. s https://brill.com/view/journals/shrs/29/1-4/article-p55_55.xml

Marsden, C. T. i Meyer, T. (2019). Regulating Disinformation with Artificial Intelligence: Effects of Disinformation Initiatives on Freedom of Expression and Media Pluralism. *European Parliamentary Research Service*. Preuzeto 10.08.2024. s https://research.monash.edu/en/publications/regulating-disinformation-with-artificial-intelligence-effects-of

Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J. i Fernández-Leal, Á. (2023). Human-in-the-loop machine learning: a state of the art. *Artif Intell Rev,* 56, 3005–3054. Preuzeto 10.08.2024. s https://link.springer.com/article/10.1007/s10462-022-10246-w

Nakov, P., Corney, D., Hasanain, M., Alam, F., Elsayed, T., Barrón-Cedeño, A., Papotti, P., Shaar, S. i Da San Martino, G. (2021). Automated Fact-Checking for Assisting Human Fact-Checkers. Preuzeto 10.08.2024. s https://arxiv.org/abs/2103.07769

Oladoyinbo, T. O., Olabanji, S. O., Olaniyi, O. O., Adebiyi, O.O., Okunleye, O. J. i Ismaila Alao, A. (2024). Exploring the Challenges of Artificial Intelligence in Data Integrity and its Influence on Social Dynamics. *Asian Journal of Advanced Research and Reports*, 18(2), 1-23. Preuzeto 10.08.2024. s https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4693987

Rana, S., Nobi, M. N., Murali, B. i Sung, A. H. (2022). Deepfake Detection: A Systematic Literature Review. Preuzeto 10.08.2024. s https://ieeexplore.ieee.org/abstract/document/9721302

Roundtable: How artificial intelligence is used to debunk fake news (2020). Preuzeto 10.08.2024. s https://akademie.dw.com/en/roundtable-how-artificial-intelligence-is-used-to-debunk-fake-news/a-55946621

Santos, F. C. C. (2023). Artificial Intelligence in Automated Detection of Disinformation: A Thematic Analysis. *Journalism and Media,* 4(2), 679-687. Preuzeto 10.08.2024. s https://doi.org/10.3390/journalmedia4020043

Ünver, A. (2023). Emerging Technologies and Automated Fact-Checking: Tools, Techniques and Algorithms. Preuzeto 10.08.2024. s https://edam.org.tr/Uploads/Yukleme_Resim/pdf-28-08-2023-23-40-14.pdf

Validation and Verification of Artificial Intelligence (2024). Preuzeto 10.08.2024. s https://www.sqs.es/validation-and-verification-of-artificial-intelligence/?lang=en

Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T. i He, L. (2022). A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, 135, 364–381. Preuzeto 10.08.2024. s https://www.sciencedirect.com/science/article/abs/pii/S0167739X22001790

Yu, P., Xia, Z., Fei, J. i Lu, Y. (2021). A Survey on Deepfake Video Detection Preuzeto 10.08.2024. s https://ietresearch.onlinelibrary.wiley.com/doi/full/10.1049/bme2.12031

Zhou, Y. i Lim, S. (2021). Joint Audio-Visual Deepfake Detection. Preuzeto 10.08.2024. s https://openaccess.thecvf.com/content/ICCV2021/html/Zhou_Joint_Audio-Visual_Deepfake_Detection_ICCV_2021_paper.html

10. Popis slika

Uporaba umjetne inteligencije za verifikaciju informacija

Sažetak

Područje umjetne inteligencije (UI) u stalnom je razvoju te se istražuje njegova moguća primjena na razne aspekte svakodnevnog života, kao što su primjerice izloženost društvenim medijima i protok informacija. Ovaj rad bavi se izazovima, problematikom i etikom uporabe umjetne inteligencije za verifikaciju informacija, ali i istražuje kako na koristan način doprinosi detekciji dezinformacija. Živimo u dobu kada su lažne vijesti svuda oko nas te je stoga primjena suvremenih tehnologija veoma korisna u provjeri točnosti informacija na internetu. U ovom kontekstu, umjetna inteligencija obuhvaća više područja – strojno učenje, duboko učenje, analizu podataka i sl. Primjena se očituje u analizi teksta, slika, audio i videozapisa kako bi se utvrdila njihova vjerodostojnost i pouzdanost. Cilj je provjeriti autentičnost informacija na internetu, identificirati lažne vijesti, dezinformacije i manipulativne sadržaje te naposljetku korisnike o tome osvijestiti i pružiti im pouzdane informacije. Umjetna inteligencija već se pokazala od velike pomoći novinarima, istraživačima te korisnicima interneta u borbi protiv dezinformacija i razvoju kritičkog mišljenja.

Ključne riječi: umjetna inteligencija, informacija, dezinformacija, verifikacija

The Use of Artificial Intelligence for Information Verification

Summary

The field of artificial intelligence (AI) is continuously evolving, and its potential application to various aspects of everyday life, such as exposure to social media and information flow, is being explored. This paper addresses the challenges, issues, and ethics of using artificial intelligence for information verification, while also examining how it can beneficially contribute to the detection of misinformation. We live in an era where fake news is pervasive, making the application of modern technologies very useful in verifying the accuracy of information on the internet. In this context, artificial intelligence encompasses several areas – machine learning, deep learning, data analysis, etc. Its application is evident in the analysis of text, images, audio, and video to determine their authenticity and reliability. The goal is to verify the authenticity of information online, identify fake news, misinformation, and manipulative content, and ultimately raise awareness among users and provide them with reliable information. Artificial intelligence has already proven to be of great assistance to journalists, researchers, and internet users in the fight against misinformation and in the development of critical thinking.

Key words: artificial intelligence, information, misinformation, verification