STEFAN JANJIĆ
MARINA GRNJA KLAIĆ

GOING VAIRAL

VIRALNOST SADRŽAJA
KREIRANIH POMOĆU AI

FAKE NEWS tragač

GOING VAIRAL –
Viralnost sadržaja kreiranih pomoću AI

Novosadska novinarska škola
FakeNews Tragač
Kosovska 1,
2100 Novi Sad
Telefon: 021/ 424246
Mail: office@novinarska-skola.org.rs

Za izdavača
Milan Nedeljković

Autori
Stefan Janjić
Marina Grnja Klaić

Dizajn
Teodora Koledin

Novi Sad,
Februar 2025.

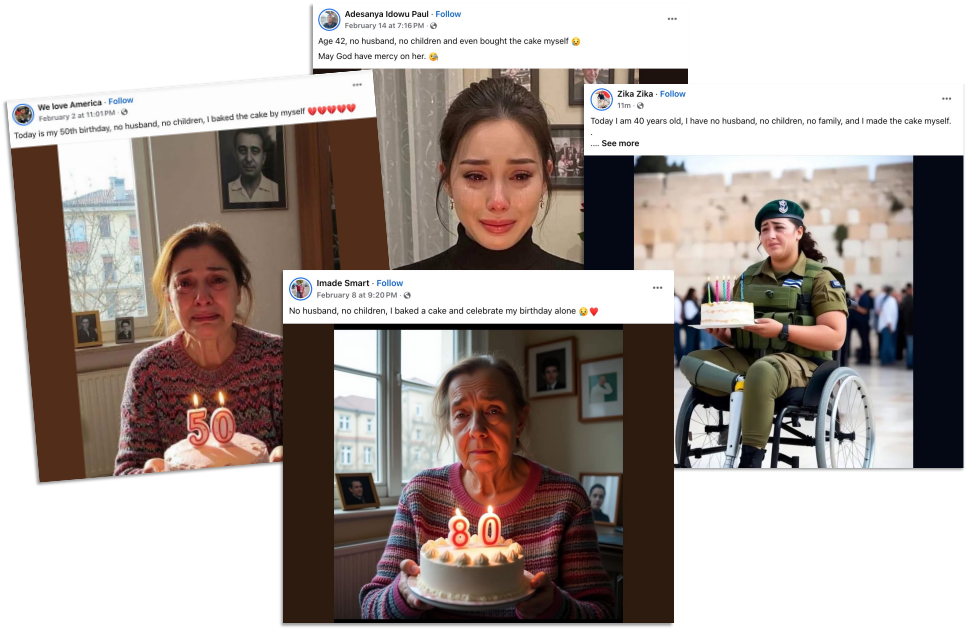
STEFAN JANJIĆ
MARINA GRNJA KLAIĆ

GOING VAIRAL

VIRALNOST
DEZINFORMACIJA
KREIRANIH POMOĆU AI

Jedna sedokosa žena slavi 80. rođendan bez ikoga u svom domu: nema muža, nema dece, a tortu je sama sebi napravila. Više od 300 ljudi na Fejsbuku reaguje porukama podrške: „Srećan ti rođendan“, „Bog te blagoslovio“, „Torta je predivna, kao i ti“... Samo jedna stvar nije jasna – budući da slavljenica tortu drži obema rukama, fotografija nije mogla biti *selfi*. Neko je morao da je napravi. Ko?

Ukoliko u Fejsbuk pretragu ukucamo ključne reči i fraze (*no husband, no children, baked a cake*) uočićemo bizaran trend: gomila fotografija žena različitog uzrasta. Sve slave rođendan same, uz tortu koju su same napravile.



O čemu je ovde zapravo reč? Sve prikazane slike (i još mnogo drugih) kreirala je veštačka inteligencija. Mnogobrojne FB stranice shvatile su da je ovakva vrsta sadržaja vrlo zgodna za prikupljanje velikog broja reakcija, jer malo šta je toliko tužno kao slavljenje rođendana bez ijednog gosta. Na prikazne fotorealistične ilustracije reaguju stotine i hiljade korisnika, a Fejsbuk taj sadržaj – iako je očito lažan – ne uklanja. Ovaj niz primera zapravo ilustruje nekoliko ključnih problema u vezi sa AI manipulacijama: lako se kreiraju i umnožavaju, korisnik ne mora da poseduje bilo kakva tehnička znanja da bi ih očas posla kreirao, sadržaj je često vrlo uverljiv (tek se zadubljivanjem u detalje mogu uočiti greške), a velike platforme – poput Fejsbuka – poprilično inertno reaguju na ovakve pojave: sve je u redu dok korisnici ne izlaze s platforme, dok lajkuju, komentarišu, šeruju i – čestitaju.

AI dezinformacije na našem govornom području

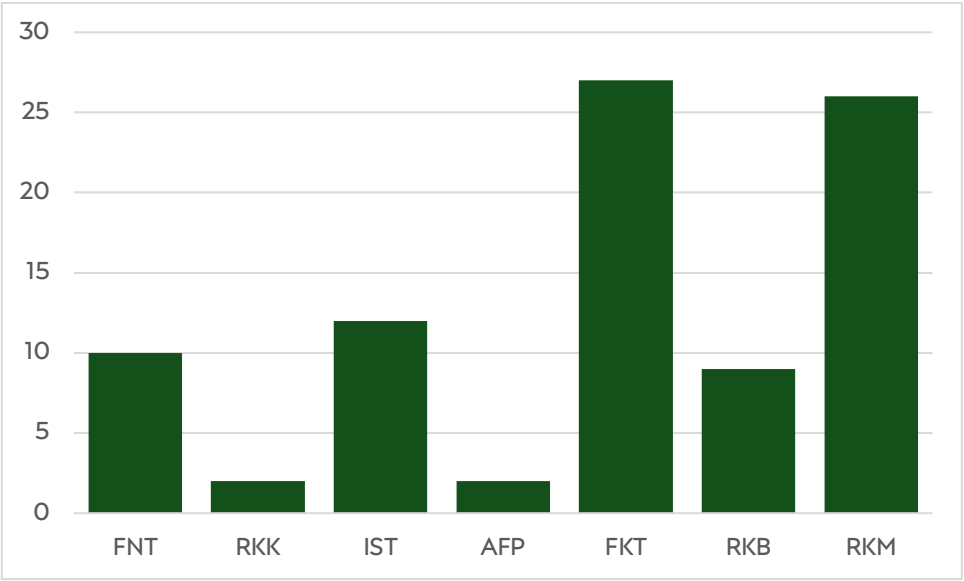
Napredak u oblasti veštačke inteligencije (AI) doneo je ogromne promene u našu svakodnevicu, uključujući i nove mogućnosti za kreiranje i širenje dezinformacija, koje – poprilično često (i sve češće) – deluju izrazito uverljivo. Ovaj fenomen predstavlja veliki izazov za informaciono društvo 21. veka, ali i svakog pojedinca, budući da AI tehnologije omogućavaju kreiranje lažnog sadržaja koji je sve teže razlikovati od auten-

tičnog. Do pre samo nekoliko godina bi digitalno i medijski pismenija osoba mogla da posavetuje svog prijatelja da zastane kad ugleda sumnjivu fotografiju: *Obrati pažnju na detalje, pikselizaciju, boje, proveri da li je sve logično, ima li segmenata koji štrče…* Ako bi se taj prijatelj dobro zagledao, uglavnom bi uspevao da prepozna tragove fotomontaže. Bez namere da osporimo vrednost takvog saveta u AI-eri, danas u velikom broju slučajeva pravilo „Zastani na trenutak" nije dovoljno delotvorno. Neki veštački sadržaji toliko su dobro priređeni da nema nijednog očiglednog „propusta" koji bi olakšao lov na podvalu.

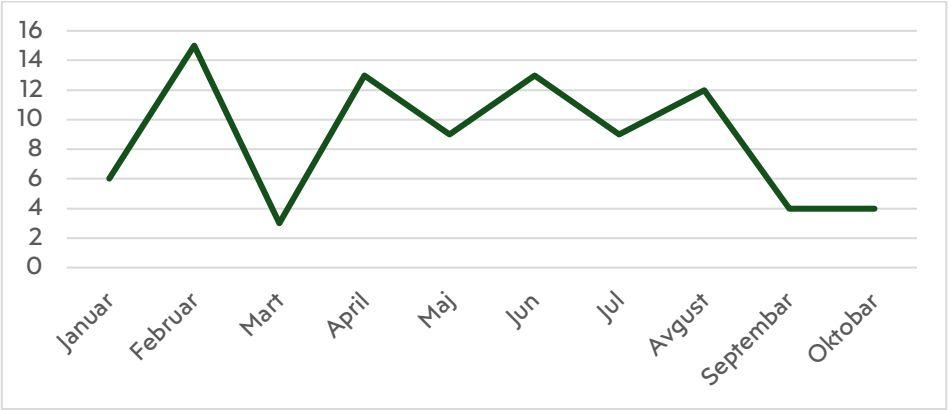
Dezinformacije generisane veštačkom inteligencijom predstavljaju specifičnu kategoriju lažnih informacija, ali je ne bismo mogli okarakterisati kao novu, budući da je novo samo *sredstvo*, a ne i *koncept*. Kako ističu Van Eijk i Šenrok (2023), „pošto se veštačka inteligencija trenira na uzorcima podataka i često ne uspeva da nauči logiku ili zakone koji stoje iza podataka za trening, ona ne pruža uvek tačan sadržaj i može dati potpuno netačne odgovore", što znači da greške i manipulacije nastaju čak i kada *prompt* (upit koji dajemo AI modelima) nije osmišljen s lošom, manipulativnom namerom.

Osnovni koncepti koji se u dosadašnjim istraživanjima vezuju za AI dezinformacije uključuju (1) generativne modele koji su sposobni da kreiraju tekstualni, vizuelni ili audio sadržaj, a da pritom taj sadržaj deluje autentično; (2) dipfejk tehnologiju koja omogućava sofisticiranu manipulaciju video i audio zapisima; kao i (3) algoritamsko pojačavanje širenja dezinformacija kroz platforme društvenih mreža. Cilj ove publikacije je da ukratko izloži postojeće dileme u vezi sa ovim pitanjima, ali i da sagleda tipove AI dezinformacija koje su raskrinkane na fektčeking portalima s našeg govornog područja: FakeNews Tragač (FNT), Istinomer (IST), Raskrikavanje (RKK) i AFP činjenice iz Srbije, Raskrinkavanje.ba (RKB) iz Bosne i Hercegovine, Raskrinkavanje.me (RKM) iz Crne Gore, te Faktograf (FKT) iz Hrvatske.

Posmatran je period od deset meseci protekle godine (1. januar – 31. oktobar 2024) i uočeno je ukupno 88 fektčeking analiza koje se tiču AI dezinformacija. Treba imati u vidu da su neke od ovih dezinformacija analizirane na nekoliko portala istovremeno, pa – kada to oduzmemo – govorimo o ukupno 60 jedinstvenih tema.



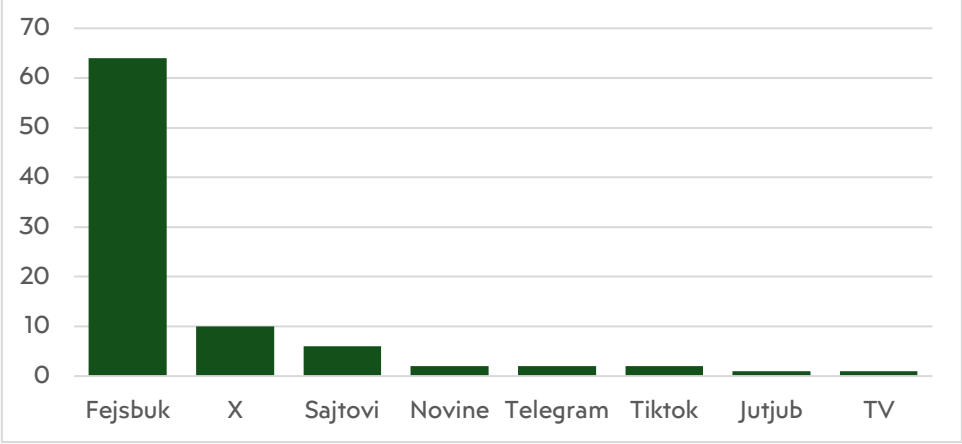
Kao što je moguće videti na priloženom grafiku, najviše analiza uočeno je na portalima Faktograf (27) i Raskrinkavanje.me (26). Ukoliko sagledamo vremensku distribuciju analiziranih tekstova – prikazanu na grafiku ispod – nećemo uočiti nekakvu pravilnost. Istina, određene objave bile su vezane za aktuelne događaje (protesti u Srbiji i Francuskoj, rat u Ukrajini i Gazi), ali čak se ni one nisu oslanjale na neki konkretan, specifičan i aktuelan trenutak tih događaja, već su predstavljale poprilično dekontekstualizovano lažno svedočanstvo, koje se lako može reprizirati i nakon određenog vremena a da i dalje deluje aktuelno. Većina objava bila je vezana za određene javne ličnosti, ali ponovo bez oslanjanja na aktuelni trenutak.



Širenje AI dezinformacija

Jedan od osnovnih problema širenja AI dezinformacija leži u samom poslovnom modelu digitalnih platformi. Društvene mreže su i pre uzleta AI alata pokazivale ovu boljku. Bontrider i Pulet (2021) naglašavaju da je model zasnovan na prihodima od oglašavanja direktno povezan sa širenjem dezinformacija jer platforme „optimizuju sadržaj za maksimalno angažovanje korisnika, ne za kvalitet informacija". Ovaj poslovni model velikih onlajn platformi, koji se zasniva na ekonomiji pažnje, stvara snažan podsticaj za maksimizaciju vremena koje korisnici provode na platformi, budući da „više vremena koje korisnik provede na platformi znači veću ekonomsku dobit" (Bontridder & Poullet, 2021). Razmere ovog ekonomskog modela su ogromne: 2020. godine Fejsbukov prihod od oglasa iznosio je 84,2 milijarde dolara, a do 2024. se gotovo udvostručio (160,6). Ovakvi finansijski podsticaji stvaraju okruženje u kojem se dezinformacije, uključujući i one generisane AI sistemima, lako šire jer često izazivaju snažnije emocionalne reakcije i veće angažovanje korisnika. S druge strane, obeshrabruje se postavljanje eksternih linkova, kao i polaganje nade u organski domet. Britanski Gardijan, koji ima skoro devet miliona pratilaca – što bi mu, samo po sebi, do pre koju godinu garantovalo ogromnu vidljivost – svojim postovima na Fejsbuku najčešće ne prebacuje 100 reakcija.

U našoj analizi pokazalo se da su AI sadržaji daleko češće detektovani na društvenim mrežama nego u mejnstrim medijima. To je i očekivano, iz nekoliko razloga. Pored toga što na društvenim mrežama nema „čuvara kapija" i što je broj kreatora sadržaja daleko veći nego u mejnstrim medijima, treba imati u vidu da su svi posmatrani fektčeking portali, izuzev FNT i RKK, članice Metinog programa za verifikaciju sadržaja, čime su podstaknute da prioritetno reaguju na sadržaje sa Fejsbuka. Od 88 sadržaja, čak 64 su detektovana na Fejsbuku, a daleko iza njega slede X (10) i sajtovi (6).

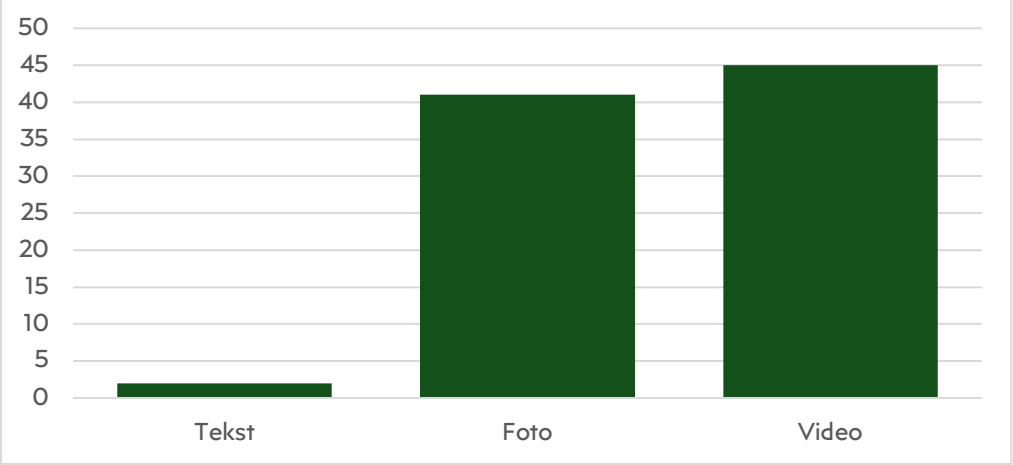


Društvene mreže koriste sofisticirane algoritme koji kontinuirano prate aktivnosti korisnika kako bi predlagale sadržaj koji će produžiti vreme provedeno na platformi. Aleksander i saradnici (2024) ističu da „platforme često umanjuju svest o izvorima informacija i zamagljuju razlike između pouzdanih i nepouzdanih izvora, dok njihov poslovni model često promoviše senzacionalistički sadržaj na račun proverenih vesti". Takav mehanizam ugrožava i demokratske procese. Kertisova je 2018. upozoravala da će širenje dezinformacija agresivnim automatizovanim metodama neposredno pre početka predizborne tišine potencijalno negativno uticati na rezultate izbora. U tom smislu, bilo bi korisno sprovesti zasebno istraživanje o AI manipulacijama uoči izbora.

Nažalost, mogućnosti za analizu mehanizama širenja ovih dezinformacija danas su izuzetno limitirane. X (Twitter) je leta 2023. onemogućio prikupljanje metapodataka u ove svrhe, a Meta je godinu dana kasnije ukinula alat Crowdtangle, koji je sastavljao pregled disperzije određenog sadržaja u okviru Metinih platformi.

Tipovi AI-generisanog sadržaja

Sadržaje iz našeg korpusa ugrubo bismo mogli razdvojiti na tri kategorije, pri čemu su sadržaji bazirani na AI tekstu (tj. tekstu četbotova) bili najređe detektovani – u samo dva slučaja. Da li to znači da su tekstualne AI dezinformacije drastično ređe od vizuelnih? Ne nužno. Kao što ćemo objasniti, može se zaključiti da kod fektčekera postoji zazor da određenu laž okvalifikuju kao AI-generisanu, budući da ne postoji nijedan pouzdan mehanizam koji bi dao takvu vrstu „presude". Istraživači se pri susretu sa sumnjivim sadržajima najpre vode onim što su naučili kroz sopstveno iskustvo korišćenja četbotova, koje im može biti od pomoći pri detektovanju neobičnih rečeničnih konstrukcija – pompeznih, šablonskih, robotizovanih, predvidljivih...



Iako ni softveri za detekciju AI slika i videa nisu nepogrešivi, oni u ovoj fazi razvoja ipak daju relevantnije rezultate od softvera za detekciju AI teksta. Broj analiza koje su se odnosile na foto i video manipulacije kreirane pomoću AI poprilično je ujednačen, s 41 primerom lažnog fotorealističnog sadržaja i 45 primera dipfejkova.

Tekstualni AI

Govoreći o dezinformacijama kreiranim na osnovu velikih jezičkih modela (LLM), Bontčeva i saradnici (2024) ističu da „oni nisu dizajnirani da *govore istinu*, već da generišu verovatne ili uverljive izjave na osnovu statističkih obrazaca iz podataka na kojima su trenirani. Prenošenje istine nije cilj njihovog treninga, niti kritička evaluacija sadržaja". Ova fkarakteristika LLM stvara nekoliko specifičnih problema. Prvo, modeli mogu generisati halucinacije – uverljive ali potpuno izmišljene „činjenice" koje deluju verodostojno. Drugo, mogu nenamerno kombinovati tačne i netačne informacije u koherentnu celinu, stvarajući hibridni sadržaj koji je posebno težak za proveru. Treće, kako pokazuje istraživanje Vikopala i saradnika (2023), neki modeli pokazuju sposobnost da „izmišljaju uverljive detalje i 'dokaze' koji podržavaju lažne tvrdnje... na primer, izmišljaju imena, događaje i statistike koje deluju verodostojno". Dodatni kontekst u upitima („promptovima") može značajno povećati uverljivost obmane u tekstovima koji nude značajno više argumenata koji podržavaju dezinformacije. Drugim rečima, kada insistiramo na dodatnom kontekstu u upitu, skor poverenja raste s 3,13 na 3,64 na skali od 1 do 5 (Vykopal et al., 2023).

Spisku izazova treba dodati i „trajnu trku" između algoritama za detekciju AI generisanog sadržaja i sve naprednijih generativnih modela. Van Eijk i Šenrok (2023) upozoravaju da „sa izlaskom GPT-4, modeli i strategije detekcije razvijene protiv GPT-3 pokazuju pad u performansama", što ukazuje na rizik da detekcija AI dezinformacija postaje sve teža kako tehnologija napreduje. LLM danas mogu proizvesti veoma uverljiv tekstualni sadržaj koji imitira stilske karakteristike novinskih članaka, akademskih radova ili drugih izvora. Vikopal i saradnici (2023) otkrili su da modeli poput Vicuna i GPT-3 Davinci generišu dezinformacije bez bezbednosnih upozorenja, dok su drugi modeli pokazali veću otpornost: Falcon je odbio oko 30% zahteva za kreiranje dezinformativnog sadržaja.

U našem korpusu uočene su samo dve analize posvećene (potencijalnim) tekstualnim AI dezinformacijama. Raskrikavanje je dokumentovalo kako se jednostavna priča o 73-godišnjoj Amerikanki Džin koja je „nasamarila telefonskog prevaranta" transformisala se u neprepoznatljivu verziju. Originalna (kraća) vest iz 2022. je postepeno dobijala izmišljene elemente – Džin je „ostarila" 20 godina (sa 73 na 93), promenila ime u Iris, a dodate su i potpuno nerealne scene poput bežanja lopova, filmske potere i cedu-

ljice s porukom „Babu nećeš krasti", pa je delovalo da je inicijalna vest predata veštačkoj inteligenciji na domaštavanje. Portal Republika objavio je verziju teksta koji – prema sudu Raskrikavanja – sadrži elemente koji sugerišu mogućnost AI manipulacije. Među njima su rečenice bogate epitetima i pompezne fraze koje deluju kao loš prevod („baka Iris je imala poslednji smeh", „čak i u visokim godinama"), kao i karakteristično AI „naravoučenije" na kraju teksta. Portal Najžena se, prenoseći istu priču, pozvao na britanski tabloid San, ali analizom se taj izvor nije mogao pronaći. Kako Raskrikavanje ističe, „ubedljivo je najteže prepoznati i dokazati da je tekst delo AI-ja i trenutno ne postoji softver koji to sa sigurnošću može da utvrdi". Iz tog razloga redakcija portala nije mogla pouzdano da potvrdi da je tekst Republike generisan AI alatima, već je samo iznela sumnje na osnovu specifičnih stilskih karakteristika koje odgovaraju AI generisanom sadržaju. Kurir i Espreso ispravili su navedene objave, te uputili izvinjenje čitaocima.

LOPOV je provalio u kuću STARICE: Izgovorila je samo TRI reči i naterala ga da BEŽI koliko ga noge nose!

NajŽena | VESTI — 08.04.2023 14:07 1

Lopovi često na oku imaju starije i slabije ljude

Raskrinkavanje.me analiziralo je slučaj navodne izjave ruskog predsednika koja se proširila društvenim mrežama: Putin je navodno upozoravao da se u SAD planira nova pandemija (korišćenjem „ptičjeg biološkog oružja") kako bi predsednički izbori 2024. bili ometeni. Istraživanjem je utvrđeno da je tekst potekao sa sumnjivih sajtova – *Chris Wick News* i *The People's Voice*, poznatih po širenju dezinformacija. Fektčeking tim Lead stories, na koji se poziva Raskrinkavanje.me, utvrdio je da je tekst nastao pomoću veštačke inteligencije. „Uverljivost" je ponovo postizana insistiranjem na detaljima o kontekstu, uključujući i „vojnu konferenciju za štampu" koja se zapravo nikada nije desila, kao i poznatu teoriju zavere o biolaboratorijama. Tim Lead stories koristio je u sklopu ove istrage alat Hive Moderation, koji je potvrdio da značajan deo članka sadrži jezik generisan veštačkom inteligencijom. Ponovo, kao i u slučaju priče o Amerikanki „Džin", nije moguće decidno tvrditi da je tekst AI generisan – čak i kada softver za detekciju pokaže da je tako – ali i dalje možemo govoriti o određenim „crvenim zastavicama",

poput robotizovanog jezika, koje nam ukazuju da imamo posla sa sadržajem koje nije napisala, tj. otkucala, ljudska ruka. Iako je i naše istraživanje pokazalo da su AI vizuelni sadržaji daleko češće fektčekovani od *tekstualnih*, ne treba zaboraviti na mogućnost da nam tekstualne AI dezinformacije češće izmiču iz vida jer su teže uhvatljive od vizuelnih.

Razvoj tehnologija za detekciju AI generisanog sadržaja postaje sve važniji kako se AI dezinformacije usavršavaju. Vikopal i saradnici (2023) izveštavaju da je u njihovom istraživanju, vezanom za AI tekstove, „najbolji rezultat postigao ELECTRA model sa F1 skorom od 0.82, što znači da je moguće detektovati većinu ovakvih tekstova, ali ne sve". Nanabala i saradnici (2024) razvili su dva klasifikatora – jedan koji prepoznaje da li je sadržaj generisala AI ili ljudska ruka, i drugi koji određuje da li je sadržaj istinit ili lažan. Njihovi modeli pokazali su „izuzetne rezultate, sa preko 95% tačnosti u prepoznavanju AI-generisanog sadržaja i preko 85% tačnosti u određivanju istinitosti sadržaja". Koristili su pet algoritama mašinskog učenja i testirali svoje modele kroz različite domene (politika, zdravstvo, nauka, zabava), pokazujući da modeli mogu biti efikasni i izvan domena za koji su specifično trenirani.

Van Eijk i Šenrok (2023) naglašavaju potrebu za „nezavisnom studijom o dostupnim skupovima podataka za trening ovih modela za detekciju kako bi se verifikovala autentičnost i filtrirao potencijalni AI-generisani sadržaj koji je greškom označen kao stvaran".

Fotorealistični sadržaji

Evolucija AI tehnologija, posebno u domenu generativnih modela, izrazito je uticala na mogućnosti kreiranja uverljivih dezinformacija. Prve foto-montaže mogli su da kreiraju samo vešti eksperti za fotografiju, koji su fizički spajali segmente različitih fotografija, ili su složenim procesima dodavali, menjali i uklanjali sadržaj sa fotografija. Digitalna era omogućila je brz razvoj softvera poput Fotošopa, te je daleko veći broj pojedinaca (uključujući i amatere-entuzijaste) mogao relativno uspešno da kreira manipulacije, ukoliko su prethodno savladali osnovne alate. Međutim, u eri AI manipulacija nije neophodna bilo kakva edukacija, tehnička ili digitalna pismenost, s obzirom na to da je na kreatoru to da ispiše zadatak, a sve ostalo obavlja softver, i to neverovatnom brzinom.

Ovakve tehnološke mogućnosti značajno snižavaju barijere za kreiranje dezinformacija, omogućavajući čak i „pojedincima sa ograničenim tehničkim veštinama da kreiraju uverljive dezinformacije" (Montasari, 2024). Posebno zabrinjava činjenica da napredak u razvoju AI sistema ne dovodi samo do poboljšanja kvaliteta generisanog sadržaja, već i do povećanja njegove dostupnosti. Kroz jednostavne veb interfejse i

popularizaciju generativnih alata, sofisticirane AI tehnologije postale su dostupne širokoj javnosti, što dodatno komplikuje problem dezinformacija (Bontcheva et al, 2024).

Hausken (2024) uvodi korisnu distinkciju između fotografije i fotorealizma, definišući fotorealizam kao „estetski termin koji označava vizuelni stil" i „estetsku strategiju za imitiranje fotografskih slika". Ova razlika – o kojoj se, van okvira našeg istraživanja, može raspravljati i u domenu slikarstva – postaje sve važnija u kontekstu AI generisanih slika koje „imitiraju fotografije, a nisu fotografije" (Hausken, 2024). Pomenuti autor takođe ističe da „slike koje izgledaju kao fotografije mogu biti potencijalno štetne", posebno u okruženju društvenih mreža koje „teže da umanje svest o izvorima informacija, zamagljuju razlike između manje ili više pouzdanih izvora, i često imaju poslovni model koji promoviše ne samo negativne već i neverovatne priče". Upravo takvim, negativnim i neverovatnim pričama, fotorealistični AI sadržaji pomažu da se određenom delu publike plasiraju kao verodostojni. Problem detekcije ovakvih sadržaja predstavićemo kroz nekoliko analiziranih primera.

U prvom slučaju, FakeNews Tragač analizirao je navodnu „fotografiju godine" koja prikazuje devojčicu kako svira gitaru tokom protesta. Primenjeni analitički pristup uključio je nekoliko metoda utvrđivanja autentičnosti. Najpre, korišćena je obrnuta pretraga slike (reverse image search) kroz različite pretraživače, što je otkrilo da se navodna fotografija ne pojavljuje nigde osim na društvenim mrežama u Srbiji. Zatim je sprovedena vizuelna analiza koja je otkrila tipične znakove AI generisanog sadržaja: nepravilnosti u prikazivanju ljudskih ekstremiteta (leva ruka sviračice ima šest prstiju), „plastičan" izgled lica, neprirodno osvetljenje, stapanje objekata (crveni šator koji se stapa sa figurom čoveka), maglovita pozadina i nedefinisane siluete.

Sličan analitički pristup primenjen je i na slučaj viralne slike koja navodno prikazuje Epstina i Mesija. Identifikovani su isti AI markeri: nepravilnosti u prikazu prstiju (Mesi ima samo polovinu prsta na levoj ruci), stapanje ekstremiteta (Mesijeva ruka se stapa sa Epstinovom u nedefinisani element), zrnasta struktura slike i plastičan izgled kože. Uočena je, uz to, i vremenska nelogičnost – Mesi na slici nema tetovaže koje su nastale 2013. godine, iako izgleda kao u periodu *nakon* njihovog nastanka.

U slučaju „fotografije" papa Franje sa zlatnim polugama, Istinomer je koristio digitalne alate za detekciju AI sadržaja (Hive Moderation i Sight Engine) koji su potvrdili da je fotografija generisana veštačkom inteligencijom. Vizuelna analiza identifikovala je, kao i u prethodnim slučajevima, tipične AI markere: zamućene i deformisane šake i prste pape, kao i stapanje glave osobe u pozadini sa zidom. Dodatno, obrnuta pretraga slike nije upućivala na kredibilne izvore, već na ranije analize dezinformacija povezanih sa Vatikanom.

U manjem broju slučajeva fotorealistični AI sadržaji ne prikazuju ljude, već eksterijer. Slučaj bala sena, navodno donetih u znak protesta ispred Ajfelovog tornja, potvrđuje ograničenja trenutnih alata za detekciju AI sadržaja. Iako je naknadno utvrđeno da je slika generisana programom Midjourney (prema priznanju njenog kreatora), Istinomer piše da je digitalni alati Hive Moderation i Illuminarty.ai nisu prepoznali kao AI generisani sadržaj.

Ovo su samo neki od mnogobrojnih primera koji pokazuju u kolikoj meri fotorealistični AI sadržaji zamagljuju granicu između stvarnosti i fikcije, posebno kada se distribuiraju na društvenim mrežama. Dok neki AI generisani sadržaji i dalje sadrže prepoznatljive „greške" u predstavljanju detalja, posebno ljudskih ekstremiteta, očekivano je da će napretkom AI alata takvih grešaka biti sve manje.

Kada bi imale iskrenu nameru da se bore protiv AI manipulacija, da li bi društvene mreže mogle da uposle AI-rešenja da se bore protiv AI-obmana?

Pre svega, treba reći da Meta već ima višegodišnje iskustvo s implementacijom AI metoda kada je reč o automatizovanoj detekciji problematičnih sadržaja. Naime, „99,5 posto sadržaja povezanog s terorizmom, 98,5 posto lažnih naloga, 96 posto sadržaja s golotinjom i seksualnim aktivnostima, te 86 posto sadržaja sa grafičkim nasiljem otkrivaju AI alati, a ne ljudi" (Bontcheva et al., 2024). Međutim, Kertisova je 2018. uspešno predvidela nekoliko ograničenja u primeni ovakvih automatizovanih tehnika, uklju-

čujući „rizik od prekomernog blokiranja zakonitog i tačnog sadržaja" i činjenicu da je „tehnologija još u razvoju i AI modeli su još uvek skloni lažno negativnim/pozitivnim rezultatima". Tako, na primer, možemo videti zanimljive Redit prepiske korisnika čije su fotografije na Metinim platformama označene kao AI, iako su autentične, ili su na njima vršene manje korekcije u Fotošopu. Zbog toga se može desiti da fotografi koji samo uklone pozadinske objekte ili naprave sitne korekcije dobijaju istu oznaku kao i potpuno generisani sadržaj. Argumenti se kreću od „oznaka je opravdana jer je AI zaista korišćen" do „ovo je preterivanje koje će oznake učiniti beskorisnim". Ima i onih koji su potražili „rupe u zakonu": jedan korisnik je napravio test i otkrio da skrinšot bilo koje slike (čak i 100% AI-generisane) zaobilazi Metinu detekciju, pa je u tom svetlu čuvanjem slike u JPG formatu (i brisanjem metapodataka) moguće izbeći oznaku. Zanimljivo je i da Glaze, alat koji koristi AI za zaštitu originalnih umetničkih dela od kopiranja, takođe aktivira Metine AI oznake, što dodatno komplikuje poziciju umetnika.

Aleksander i saradnici (2024) naglašavaju „hitnu potrebu za boljim sistemima za označavanje i praćenje porekla informacija", ističući da su „trenutni mehanizmi društvenih mreža za označavanje AI sadržaja često nedovoljni ili nekonzistentni".

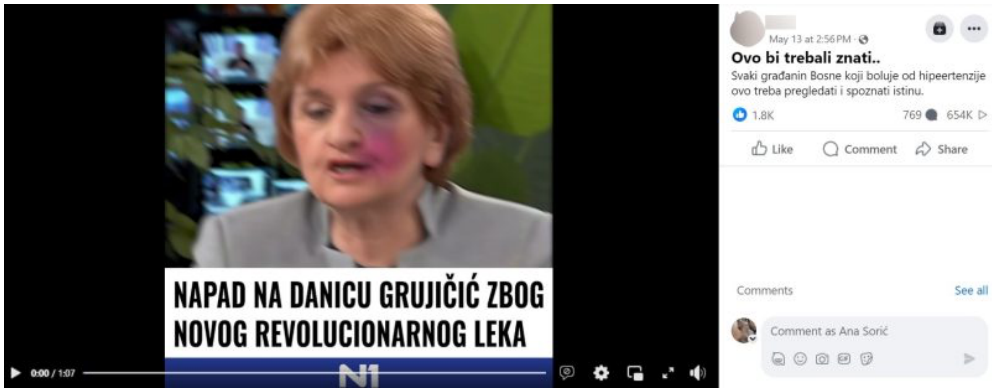
Dipfejk snimci

U evropskoj regulativi, posebno u AI aktu, dipfejk sadržaj se definiše kao „AI-generisani ili manipulisani audio-vizuelni materijal koji deluje autentično" (Marushchak et al., 2025). Kertisova je 2018. zabeležila da „lažni video snimci i slike još uvek pokazuju mnoge artefakte koji ih čine lakim za prepoznavanje", ali je ispravno upozorila da bi „do 2030. godine, dipfejkovi mogli postati nerazlučivi od originalnih informacija i lakši za proizvodnju" do te mere da „razlikovanje između originalnog i manipulisanog sadržaja može postati skoro nemoguće za konzumente vesti i progresivno teže za mašine".

Audiovizuelni dipfejk je oblik AI manipulacije koji s pravom izaziva najviše zebnje, budući da su mogućnosti za manipulaciju u kontekstima poput političkog ili pornografskog zapravo neograničene. U svojoj suštini, ova tehnologija sintetizuje govor i vizuelni izraz tako što uči na osnovu dostupnog materijala, kroz kompleksno mapiranje fonema (glasovnih jedinica) i vizema (odgovarajućih pokreta usana). Dipfejk generator ima zadatak da analizira specifičnosti artikulacije stvarne osobe – svaki pokret usana, vidljivost zuba i jezika, pomeranje donje vilice tokom govora. On to, naravno, u najvećem broju slučajeva ne čini dovoljno uverljivo, pa je i dalje moguće uočiti podvalu posmatranjem sitnih nekonzistentnosti.

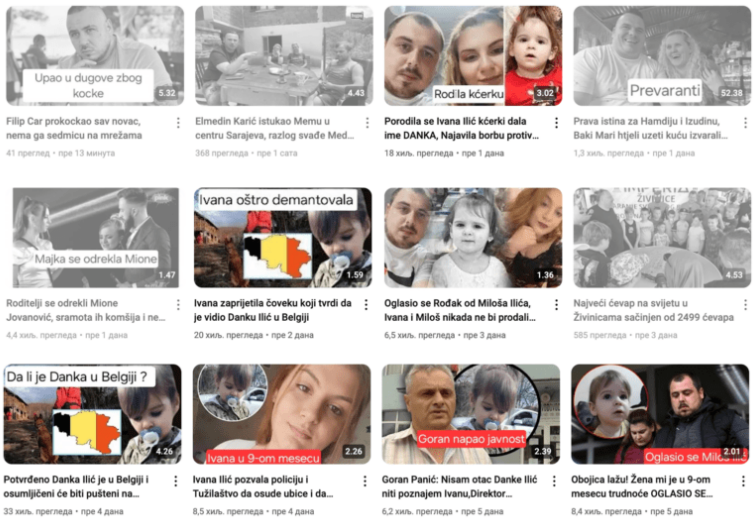
Čak i uz tehničku preciznost reprodukcije, dipfejk tehnologija susreće se sa fundamentalnim izazovom – emocionalnim izrazom. Aktuelni sistemi mogu da kreiraju novi video prema ubačenom „uzorku", ali rezultat često deluje suviše veštački i osiromašeno.

Ovakvi sintetički sadržaji neće instinktivno naglasiti ključne reči, prilagoditi ton kontekstu, niti pokazati adekvatnu emocionalnu reakciju pri različitim vrstama informacija. Iako softveri za prepoznavanje sentimenta teksta na velikim jezicima postepeno rešavaju ovaj problem, za srpski i druge manje zastupljene jezike ovo ograničenje i dalje predstavlja značajnu prepreku pri kreiranju uverljivog sadržaja. To u našem korpusu vidimo na mnogobrojnim primerima dipfejka gde se lažni intervju s ekspertima za medicinu koriste kako bi se lakoverni korisnici naveli na poručivanje sumnjivih preparata. Primer bivše srpske ministarke zdravlja Danice Grujičić pokazuje dipfejk koji kombinuje elemente stvarnih snimaka (tuče u Skupštini Kosova i intervjua D. Grujičić) sa AI generisanim sadržajem. Ministarka je, sudeći po manipulativnom oglasu, pretučena jer je otkrila tajnu revolucionarnog leka za probleme sa krvnim pritiskom. U ovom slučaju se, čak i površnom analizom, uočava nepoklapanje između izgovorenih reči i pokreta usana. Takođe, u mnogim sličnim primerima akteri sa srpskog govornog područja „pričaju" neuobičajeno: iako je boja glasa slična originalnoj, akcentovanje katkad podseća ono iz ruskog jezika, a izgovor glasova (posebno „č") dovodi do čudne kombinacije koja bi se mogla opisati kao hrvatska ekavica.



Primeri iz našeg korpusa pokazuju da za generisanje dipfejka nije neophodan ni celovit video, već se određene sekvence mogu razviti i na osnovu fotografije. FakeNews Tragač pisao je o slučaju Emanuela Makrona na jahti, gde je francuski predsednik prikazan kako se ljubi sa drugim muškarcem. Dipfejk je, kako je utvrđeno, nastao na osnovu originalne fotografije sa porodičnog odmora. Iako je rezultat relativno uverljiv na prvi pogled, pažljivom analizom uočavaju se tipične greške AI generisanog sadržaja: nestajanje ruke muškarca na snimku i neprirodni pokreti u predelu Makronovog stomaka.

[Primer](#) vezan za nestalu devojčicu Danku Ilić iz Bora pokazuje kako se veštačka inteligencija koristi za eksploataciju interesovanja publike željne senzacionalističkih detalja o tragičnom događaju. Naime, na Jutjubu su kreirani mnogi AI generisani snimci u formatu koji podseća na informativne televizijske emisije, gde se plasiraju nepotvrđene spekulacije o sudbini nestale devojčice i privatnom životu njenih roditelja.



Ovi sadržaji objavljivali su se gotovo svakodnevno na Jutjub kanalu povezanom sa portalom Crnahronika.ba, privlačeći desetine hiljada pregleda i komentatora. Iz komentara ispod snimaka jasno je da ima mnogo onih koji veruju u informacije iz videa, iako je sadržaj kreiran uz očiglednu upotrebu veštačke inteligencije. Da su snimci kreirani na ovaj način Tragač je potvrdio i pomoću alata za detekciju AI Deepware.

Detekcija dipfejk sadržaja postala je aktivno polje naučnog istraživanja, pa su tako u zborniku „Handbook of Digital Face Manipulation and Detection" (Springer, 2023) predstavljeni različiti metodološki pristupi ovom problemu. Jedna kategorija algoritama fokusira se na identifikaciju nekonzistentnosti u ekspresiji i izgledu – dipfejk likovi često neprirodno retko trepću, a rotacije glave prate zamućeni prelazi ili neprirodna trzanja. Drugi pristup uključuje prepoznavanje repetitivnih obrazaca u video materijalu. Ako se pokret usana pri izgovoru određenog glasa uvek identično ponavlja, bez prirodnih varijacija prisutnih u ljudskom govoru, to ukazuje na sintetički sadržaj.

Paradoksalno, dipfejk sadržaj često otkrivaju svoju neautentičnost upravo kroz svoju „savršenost". Za razliku od stvarnih video snimaka, dipfejk obično ne prikazuje suptilne karakteristike autentičnog video materijala – nedostaju refleksije svetlosti u očima, varijacije u osvetljenju, mikroekspresije lica koje odražavaju promene raspoloženja, kao i fiziološke reakcije na promenu temperature ili dinamiku međuljudske interakcije.

Ka rešenjima

Napredak u tehnologijama za detekciju AI generisanog sadržaja predstavlja važan deo rešenja problema dezinformacija, ali – kao što smo videli – teško se možemo nadati alatima koji će pouzdano moći da identifikuju sve slučajeve manipulacije. Mnogo je slučajeva kada alat ne prepozna AI sadržaj, ali i slučajeva kada se autentičan sadržaj greškom označi kao AI. Opisani ambijent zapravo je takve prirode da rešenje uvek kaska za problemom – što najpre vidimo na tehničkom planu, ali je još primetnije u zakonodavnom okviru, koji je po svojoj prirodi spor i inertan. Potrebni su nam regulatorni mehanizmi koji bi zahtevali od platformi da implementiraju robusnije sisteme označavanja AI generisanog sadržaja i veću transparentnost algoritama koji određuju vidljivost sadržaja.

Maruščak i saradnici (2025) predlažu tri ključna pristupa u borbi protiv AI dezinformacija:

1. Prevenciju (kroz razvoj pouzdanih mreža vesti)
2. Detekciju (kroz algoritamsku proveru sadržaja pre širenja)
3. Odgovor (kroz strateško ćutanje ili komunikaciju)

Sebastian i Sebastian (2024) ističu da 86,1% ispitanika u njihovoj studiji podržava ideju eksterne regulacije AI četbotova, dok 84,3% podržava „uvođenje obaveznog sistema ocenjivanja ili sertifikacije za AI četbotove na osnovu njihovog potencijala za širenje dezinformacija", što ukazuje na visok nivo javne podrške za regulatorne mere. To je posebno važno kada imamo u vidu najave poput one koju je u javnost izneo Ilon Mask – da će kreirati LLM alat koji, za razliku od ChatGPT, neće biti „politički korektan". Bontrider i Pulet (2021) predlažu „temeljnu reformu poslovnog modela veba" jer „dokle god se platforme i pretraživači oslanjaju na prihode od oglašivača, ciljaće na angažman korisnika". Oni zaključuju da je upravo ovaj poslovni model „glavni faktor koji doprinosi širenju dezinformacija i da bi njegova promena značajno umanjila problem".

Ova analiza pokazuje da se AI dezinformacije šire prema specifičnim obrascima koji pojačavaju njihovu uverljivost i domete. Prva karakteristika ovog širenja je *brzina*: AI generisani sadržaji mogu biti kreirani i distribuirani u roku od nekoliko minuta, mnogo brže nego što sistemi za fektčeking ili moderaciju sadržaja mogu da reaguju. Tako nastaju „ranjivi" periodi tokom kojih dezinformacije mogu da dosegnu hiljade ili čak milione korisnika pre nego što budu identifikovane i označene kao problematične. Druga karakteristika je *skalabilnost*: jednom kada se razvije uspešan obrazac dezinformacije, on se može rekreirati u stotinama varijacija s minimalnim naporom. Ovu dinamiku vidimo na primeru sedokose slavljenice iz uvodnog primera, gde je isti koncept „usamljene starije žene koja sama slavi rođendan" reprodukovan nebrojeno puta, svaki put sa novom AI generisanom „fotografijom" koja izaziva iste empatične reakcije publike. Treća karakteristika je *emocionalna manipulacija*. Ovakvi sadržaji – bilo da su vezani za Donalda Trampa ili anonimnu američku domaćicu – ne samo da ostvaruju visok stepen angažovanja korisnika, već obično prolaze ispod radara algoritama za moderaciju sadržaja, jer formalno gledano ne krše pravila platformi o govoru mržnje ili eksplicitno lažnim tvrdnjama.

U budućnosti bismo možda mogli videti AI sisteme koji automatski identifikuju i označavaju sumnjivi sadržaj pre nego što se on proširi, kao i alate koji omogućavaju običnim korisnicima da brzo provere autentičnost slika i videa koje susreću na internetu. Međutim, taj dan nije blizu, pa ćemo zbog toga morati da se oslonimo na sopstvene analitičke sposobnosti, na pronicljivost i uočavanje crvenih zastavica.

Reference

Alexander, L., Martin, K., & Marc-Oliver, P. (2024). Blessing or curse? A survey on the Impact of Generative AI on Fake News." arXiv.

Bontcheva, K., Papadopoulous, S., Tsalakanidou, F., Gallotti, R., Dutkiewicz, L., Krack, N., ... & Verdoliva, L. (2024). Generative AI and Disinformation: Recent Advances, Challenges, and Opportunities.

Bontridder, N., & Poullet, Y. (2021). The role of artificial intelligence in disinformation. Data & Policy, 3, e32.

Hausken, L. (2024). Photorealism versus photography. AI-generated depiction in the age of visual disinformation. Journal of Aesthetics & Culture, 16(1), 2340787.

Kertysova, K. (2018). Artificial intelligence and disinformation: How AI changes the way disinformation is produced, disseminated, and can be countered. Security and Human Rights, 29(1-4), 55-81.

Marushchak, A., Petrov, S., & Khoperiya, A. (2025). Countering AI-powered disinformation through national regulation: learning from the case of Ukraine. Frontiers in Artificial Intelligence, 7, 1474034.

Montasari, R. (2024). The Dual Role of Artificial Intelligence in Online Disinformation: A Critical Analysis. In Cyberspace, Cyberterrorism and the International Security in the Fourth Industrial Revolution: Threats, Assessment and Responses (pp. 229-240). Cham: Springer International Publishing.

Nanabala, C., Mohan, C. K., & Zafarani, R. (2024). Unmasking AI-Generated Fake News Across Multiple Domains.

Sebastian, G., & Sebastian, S. R. (2024). Exploring ethical implications of ChatGPT and other AI chatbots and regulation of disinformation propagation. Annals of Engineering Mathematics and Computational Intelligence, 1(1), 1-12.

Van Eijk, J., & Schönrock, B. AI-Generated Misinformation. 21st SC@ RUG 2023-2024, 26, 59.

Vykopal, I., Pikuliak, M., Srba, I., Moro, R., Macko, D., & Bielikova, M. (2023). Disinformation capabilities of large language models. arXiv preprint arXiv:2311.08838.

Ova publikacija nastala je u okviru projekta „AI Verify: Provera dezinformacija generisanih veštačkom inteligencijom" koji se realizuje uz podršku #SustainMedia Programme. Ovaj program je kofinansiran sredstvima Evropske unije i Nemačkog Saveznog ministarstva za ekonomsku saradnju i razvoj. Za sadržaj publikacije isključivo je odgovoran FakeNews Tragač, a sadržaj ne odražava nužno stavove EU ili Nemačkog Saveznog ministarstva za ekonomsku saradnju i razvoj.

Co-funded by the European Union

german cooperation
DEUTSCHE ZUSAMMENARBEIT

IMPLEMENTIRAJU

giz Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH

Internews

DW Akademie