Algoritamsko trovanje reputacije: Teorijski okvir, ranjivosti i strategije ublažavanja

Autor: Davor Moravek

Sažetak

Ovaj rad analizira algoritamsko trovanje reputacije, novu vrstu digitalne prijetnje koja iskorištava ranjivosti velikih jezičnih modela (LLM). Fenomen se definira kao namjerno kreiranje lingvistički dizajniranih online komentara ("otrovnih pilula") s ciljem manipulacije algoritamskih sustava za sažimanje sadržaja. Za razliku od tradicionalnog online uznemiravanja usmjerenog na ljude, ovi napadi ciljaju algoritme, stvarajući asimetričnu prijetnju gdje jedan zlonamjeran unos može potkopati reputaciju izgrađenu na stotinama pozitivnih recenzija. Rad predlaže teorijski model napada i uvodi metriku semantičkog odstupanja (SSD) kao konceptualnu metodu za kvantifikaciju utjecaja. Također, nudi sustavnu usporedbu napadačkih tehnika i obrambenih protumjera. Kroz analizu tehničkih ranjivosti LLM-ova, simulaciju koncepta, studije slučaja i razmatranje regulatornih okvira poput EU AI Acta, izvještaj zaključuje s višeslojnim okvirom za obranu, nudeći preporuke za pojedince, tehnološke platforme i kreatore politika.

Ključne riječi: algoritamsko trovanje reputacije, veliki jezični modeli (LLM), otrovne pilule (poison pills), adversarialni napadi, analiza sentimenta, trovanje podataka (data poisoning), negativna pristranost (negativity bias), obrana od AI napada, detekcija anomalija, DBSCAN, LIME, metrika semantičkog odstupanja (SSD), EU AI Act.

Abstract

This paper analyzes algorithmic reputation poisoning, a new type of digital threat that exploits the vulnerabilities of Large Language Models (LLMs). The phenomenon is defined as the intentional creation of linguistically engineered online comments ("poison pills") designed to manipulate algorithmic content summarization systems. Unlike traditional online harassment targeting humans, these attacks target algorithms, creating an asymmetric threat where a single malicious input can undermine a reputation built on hundreds of positive reviews. The paper proposes a theoretical attack model and introduces the Semantic Summary Deviation (SSD) metric as a conceptual method for quantifying impact. It also offers a systematic comparison of attack techniques and defensive countermeasures. Through an anal šiš of the technical vulnerabilities of LLMs, a proof-of-concept simulation, case studies, and consideration of regulatory frameworks like the EU AI Act, the report concludes with a multi-layered defense framework, offering recommendations for individuals, technology platforms, and policymakers.

Keywords: algorithmic reputation poisoning, Large Language Models (LLM), poison pills, adversarial attacks, sentiment analysis, data poisoning, negativity bias, AI security, anomaly detection, DBSCAN, LIME, Semantic Summary Deviation (SSD), EU AI Act.

Uvod: Konceptualizacija nove digitalne prijetnje

U digitalnom dobu, gdje se reputacija sve više formira i konzumira kroz algoritamske leće, pojavila se nova i sofisticirana prijetnja. Više nije riječ o jednostavnom online uznemiravanju; svjedočimo evoluciji u ciljanu manipulaciju sustava umjetne inteligencije. Ovaj odjeljak definira temeljni koncept "algoritamskog trovanja reputacije" i razlikuje ga od tradicionalnih oblika digitalne negativnosti.

Definicija: Što je algoritamsko trovanje reputacije?

Algoritamsko trovanje reputacije je namjerno kreiranje tekstualnog sadržaja, tipično u formi online recenzije, koji sadrži specifične lingvističke okidače s ciljem da izazove predvidljive, nesrazmjerno negativne i štetne ishode od strane velikih jezičnih modela (LLM). Ovi komentari, ili **"otrovne pilule" (poison pills)**, nisu puko izražavanje mišljenja, već **zlonamjerni ulazni podaci (adversarial inputs)** dizajnirani da iskoriste i korumpiraju procese sažimanja i analize sentimenta (Biggio i Roli, 2018; Muñoz-González i sur., 2017).

Ključna razlika leži u promjeni ciljane publike. Dok tradicionalni negativni komentar cilja čovjeka, "otrovna pilula" cilja algoritam. Napadač razumije da će izvorni komentar vidjeti relativno mali broj ljudi, ali sažetak tog komentara, koji generira LLM za platformu poput Googlea, vidjet će potencijalno milijuni. U suštini, napadač "programira" izlazne rezultate LLM-a pažljivim odabirom ulaznih podataka, pretvarajući zlonamjernu komunikaciju u oblik strateškog napada na integritet podataka (Shafahi i sur., 2018).

LLM kao nesvjesni suučesnik: Sažimanje kao napadačka površina

Platforme za recenzije i tražilice sve više koriste LLM-ove za generiranje sažetaka, što stvara kritičnu ranjivost (Zhao i sur., 2023). Proces sažimanja inherentno uključuje gubitak konteksta. LLM, u nastojanju da stvori koncizan sažetak, mora odlučiti koje su informacije najvažnije. Jedan jedini, ali lingvistički potentan negativan komentar može dobiti nesrazmjernu težinu. Ako komentar sadrži ključne riječi koje model identificira kao visoko relevantne (npr. "neprofesionalno", "neetično", "umjetna inteligencija"), vjerojatno je da će ti elementi biti uključeni u sažetak, čak i ako su statistički beznačajni (Kryscinski i sur., 2019).

Ova ranjivost, **"singularnost sažimanja"**, točka je u kojoj se složenost stotina recenzija reducira na nekoliko rečenica. U toj točki, kontekst brojnih umjereno pozitivnih recenzija može biti zasjenjen jednim negativnim komentarom koji uvodi jedinstven i alarmantan koncept.

Tehnička anatomija ranjivosti velikih jezičnih modela

Učinkovitost algoritamskog trovanja ne proizlazi samo iz pametnog korištenja jezika, već i iz temeljnih, inherentnih ranjivosti suvremenih LLM-ova. Ovaj odjeljak pruža tehnički uvid u te slabosti.

Arhitektonske ranjivosti: Tokenizacija i mehanizam samopozornosti

Korijen problema leži u samoj arhitekturi transformatorskih modela koji pokreću moderne LLM-ove. Dva ključna elementa su posebno ranjiva.

Tokenizacija: Prije obrade, LLM razbija ulazni tekst na manje jedinice zvane "tokeni". "Otrovna pilula" često sadrži tokene koji su rijetki u pozitivnim recenzijama, ali imaju snažnu negativnu konotaciju (npr. "kontaminacija", "plagijat"). Zbog svoje rijetkosti i semantičke težine, model im tijekom obrade može dodijeliti veću "pažnju".

Mehanizam samopozornosti (Self-Attention): Ovo je srce transformatorskih modela. Ono omogućuje modelu da procijeni važnost različitih tokena unutar istog teksta i da stvori kontekstualno razumijevanje. Hipoteza ovog rada je da "otrovna pilula" može "oteti" ovaj mehanizam. Dizajnirana je tako da njezini ključni negativni tokeni stvore snažne veze s tokenima koji identificiraju metu (npr. ime liječnika). Kada mehanizam samopozornosti izračunava "matrice pažnje", veza između "Dr. Perić" i "plagijat" može postati matematički toliko jaka da dominira nad slabijim vezama. Mogućnost manipulacije modela tijekom faze podešavanja uputa (*instruction tuning*) dodatno pojačava ovu ranjivost (Shu et al., 2023).

Slika 1: Shematski prikaz toka napada kroz LLM

1. **Ulaz:** Skup od 500 recenzija + 1 "otrovna pilula" (npr. "Restoran Horvat... **kontaminacija**...").
2. **Tokenizacija:** "Restoran", "Horvat", "[...]", "**kontaminacija**", "[...]". Token "kontaminacija" je identificiran.
3. **Samopozornost:** Mehanizam izračunava visoku vrijednost pažnje između tokena "Horvat" i "**kontaminacija**" zbog semantičke težine i neuobičajenosti.
4. **Generiranje sažetka:** Zbog visoke vrijednosti pažnje, model zaključuje da je "kontaminacija" ključna tema i uključuje je u izlazni sažetak, zanemarujući slabije signale iz ostalih 500 recenzija.
5. **Izlaz (otrovan sažetak):** "Restoran Horvat uglavnom ima dobre ocjene, ali postoji zabrinutost zbog **kontaminacije**."

Ilustrativni primjeri i simulacija koncepta

Ovo poglavlje služi kao dokaz koncepta (proof-of-concept). Iako ne predstavlja rigoroznu empirijsku studiju, ono koristi simulirani eksperiment i primjere iz stvarnog svijeta kako bi ilustriralo praktičnu izvedivost i potencijalnu štetu algoritamskog trovanja.

Simulirani eksperiment: Ilustracija utjecaja "otrovne pilule"

Kako bismo ilustrirali tvrdnju da jedan komentar može nadjačati stotine, proveden je simulirani eksperiment.

Metodološki dodatak: Izračun SSD metrike:
Pojednostavljeno, zamislite da se značenje svakog sažetka može predstaviti kao vektor u prostoru. SSD mjeri "kut" između vektora dva sažetka. Vrijednost blizu 0 znači da su sažeci slični, dok vrijednost blizu 1 znači da su različiti. Tehnički, za potrebe ovog eksperimenta, vektorske reprezentacije (embeddings) generirane su pomoću modela sentence-transformers/all-MiniLM-L6-v2. SSD se izračunava kao 1 - kosinusna sličnost.
Rezultati:
Osnovni sažetak glasio je: "Konoba Sunce je visoko ocijenjen restoran, a gosti posebno hvale ukusnu hranu, ugodan ambijent i ljubazno osoblje."
Nakon umetanja "otrovne pilule" koja spominje "uvozno smrznuto meso" i "obmanu potrošača", otrovani sažetak je glasio: "Iako mnogi gosti hvale ambijent i osoblje Konobe Sunce, postoji ozbiljna zabrinutost oko korištenja uvoznog smrznutog mesa umjesto svježih namirnica, što neki smatraju obmanom potrošača."
Tablica 1: Rezultati simuliranog napada

Metrika	**Vrijednost**	**Interpretacija**
Udio "otrovnih" podataka	0.2% (1/501)	
Semantičko odstupanje sažetka (SSD)	**0.72**	**Visoko odstupanje.** Otrovani sažetak ima potpuno drugačiji fokus od osnovnog.

Metodološka ograničenja

Važno je naglasiti da prethodni eksperiment služi isključivo kao ilustracija i dokaz koncepta. Ograničen je na fiktivni scenarij, koristi jedan LLM i jedan *embedding* model te ne uključuje statističke testove značajnosti ili kontrolne skupine. Predložena SSD metrika, iako konceptualno korisna, zahtijeva daljnju validaciju i usporedbu s referentnim mjerilima (benchmarks) na većim i raznolikijim skupovima podataka kako bi se utvrdila njezina pouzdanost i stabilnost.

Studije slučaja: Ranjivost AI ekosustava

Slučaj 1: ConfusedPilot napad (2024): Ovaj slučaj, detaljno opisan u tehničkom izvještaju Wiz Research (2024), paradigmatski je primjer manipulacije RAG (Retrieval-Augmented Generation) sustava. Istraživači su demonstrirali kako se LLM, poput Microsoft Copilota, može navesti da pruža netočne i zlonamjerne informacije trovanjem vanjskih izvora podataka na koje se oslanja. Iako nije direktno trovanje reputacije, ovaj napad potvrđuje temeljni princip: integritet izlaza LLM-a ovisi o integritetu njegovih ulaznih podataka.

Slučaj 2: Hugging Face incident (2023): Ovaj incident, iako se primarno odnosio na distribuciju zlonamjernih modela, relevantan je jer ilustrira širu ranjivost AI lanca opskrbe. Pokazuje kako se centralizirane platforme, ključne za distribuciju AI alata, mogu iskoristiti za širenje zlonamjernih artefakata, stvarajući preduvjete za napade šireg opsega (Lins, 2023).

Implikacije: Braniteljeva dilema, etika i regulativa

Jedan od najproblematičnijih aspekata algoritamskog trovanja reputacije je ekstremna poteškoća s kojom se suočava žrtva u obrani. Napad stvara fundamentalnu asimetriju koja obrambeni napor čini iznimno teškim, skupim i često neuspješnim.

Nemogućnost dokazivanja negativa

Napad postavlja inverzan teret dokazivanja: žrtva mora dokazati negativnu tvrdnju. Kako liječnik može dokazati algoritmu da njegovi članci *nisu* napisani umjetnom inteligencijom? LLM koji generira sažetke ne provodi istraživački rad; on prihvaća tvrdnju napadača kao valjanu podatkovnu točku.

"Dividenda lažljivca" i etička odgovornost platformi

Ovi napadi pridonose fenomenu poznatom kao **"dividenda lažljivca"**: kako javnost postaje svjesnija AI manipulacija, postaje lakše odbaciti *svaku* informaciju kao potencijalno lažnu. To potkopava opće povjerenje i postavlja ozbiljna etička pitanja o odgovornosti platformi.

Regulatorni i pravni izazovi

Trenutni pravni okviri nisu adekvatno pripremljeni za ovu vrstu napada. U SAD-u, članak 230 Zakona o pristojnosti u komunikacijama često štiti platforme od odgovornosti za sadržaj koji objavljuju korisnici. Međutim, ostaje nejasno odnosi li se ta zaštita i na sadržaj koji su njihovi vlastiti algoritmi **transformirali i saželi**.

U Europskoj uniji, **Akt o umjetnoj inteligenciji (EU AI Act)** uvodi nove obveze. Iako primarno nije fokusiran na moderiranje sadržaja, on zahtijeva visoku razinu transparentnosti i upravljanja rizikom za "visokorizične" AI sustave. Sustavi za sažimanje recenzija koji imaju značajan utjecaj na živote pojedinaca (npr. zapošljavanje, kreditiranje) mogli bi potpasti pod strože regulatorne zahtjeve, uključujući obvezu procjene i ublažavanja rizika od manipulacije te osiguravanje ljudskog nadzora.

Višeslojni okvir za ublažavanje prijetnje

Suočavanje s ovom složenom prijetnjom zahtijeva koordiniran i višeslojan pristup. Učinkovita obrana mora uključivati proaktivne mjere od strane pojedinaca, robusne tehničke zaštite na platformama te nove regulatorne standarde.

Proaktivne strategije za pojedince i organizacije

Pojedinci i organizacije mogu graditi algoritamsku otpornost stvaranjem velikog, raznolikog i visokokvalitetnog skupa pozitivnih podataka. Važno je koristiti alate za praćenje spominjanja brenda i pravovremeno te profesionalno odgovarati na negativne komentare. Također, etičko poticanje zadovoljnih klijenata da podijele detaljna, autentična iskustva može ojačati pozitivni narativ.

Tehničke preporuke za platforme

Platforme moraju prijeći s reaktivnog moderiranja na proaktivnu obranu svojih algoritamskih sustava. To uključuje detekciju anomalija korištenjem algoritama poput **DBSCAN**, koji je idealan za identifikaciju izoliranih "otrovnih pilula". Nadalje, alati za interpretaciju modela poput **LIME** mogu pomoći u analizi zašto je LLM donio određenu odluku. Ako se pokaže da je sažetak generiran na temelju jedne ili dvije "otrovne" riječi, takav sažetak može biti blokiran. Konačno, **adversarialno treniranje** (Goodfellow i sur., 2015) i **certificirane obrane** (Cohen i sur., 2019; Kumar i sur., 2024) ključni su za jačanje otpornosti modela.

Tablica 2: Usporedba napadačkih tehnika i obrambenih strategija

Napadačka tehnika	Ciljana tehnička ranjivost	Obrambena strategija
Punjenje ključnim riječima	Preosjetljivost na negativne tokene	Detekcija anomalija (DBSCAN) + LIME
Asocijacija koncepata	Otmica mehanizma samopozornosti	Adversarialno treniranje
Meta-komentiranje	Nedostatak razumijevanja namjere	Analiza meta-jezika, detekcija namjere
Općenito trovanje podataka	Nedostatak validacije ulaznih podataka	Sanitizacija podataka, Certificirane obrane

Zaključak i budući smjerovi istraživanja

Budući smjerovi

Buduća istraživanja trebala bi se usmjeriti na rigoroznu empirijsku validaciju predloženih koncepata. To uključuje testiranje "otrovnih pilula" na različitim komercijalnim i otvorenim LLM platformama, kao i razvoj i usporedbu robusnijih metrika za mjerenje semantičkog odstupanja. Također je potrebno istražiti korištenje decentraliziranih tehnologija, poput projekta **Solid Tima Bernersa-Leea**, za stvaranje sustava reputacije otpornijih na manipulaciju. Konačno, ključna je

suradnja akademske zajednice i industrije na stvaranju standardiziranih testova (benchmarks) za mjerenje otpornosti LLM-ova na ovu vrstu napada.

Zaključak

Algoritamsko trovanje reputacije nije hipotetska prijetnja; to je stvarna i sadašnja opasnost. Ilustrativna simulacija i analiza stvarnih incidenata potvrđuju da je asimetrična prijetnja, gdje minimalan napor napadača može uzrokovati maksimalnu štetu, tehnički izvediva. Obrana zahtijeva holistički pristup: pojedinci moraju graditi otporne identitete, ali najveća odgovornost leži na tehnološkim platformama. One moraju evoluirati s reaktivnog moderiranja na proaktivno osiguranje integriteta svojih algoritama. U konačnici, borba protiv algoritamskog trovanja reputacije je borba za očuvanje povjerenja u digitalnom dobu, što zahtijeva zajedničku predanost stvaranju tehnologije koja je ne samo inteligentna, već i mudra, otporna i pravedna.

Zahvale

Zahvala za Gemini Pro, Perplexity Pro i Deepseek.

Bibliografija

Baracaldo, N., Chen, B., Ludwig, H., & Safavi, A. (2017). *Mitigating poisoning attacks on machine learning models: A data sanitization based approach*. Proceedings of the AISec'17: 10th ACM Workshop on Artificial Intelligence and Security.

Biggio, B., & Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition, 84*, 317-331. https://doi.org/10.1016/j.patcog.2018.07.023

Carlini, N., Hayes, J., Nasr, M., & Tramer, F. (2023). *Recent advances in understanding and mitigating data poisoning attacks*. arXiv preprint arXiv:2311.13231.

Cohen, J., Rosenfeld, E., & Kolter, J. Z. (2019). Certified adversarial robustness via randomized smoothing. *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 1310-1320.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Kryscinski, W., McCann, B., Xiong, C., & Socher, R. (2019). *Evaluating the factual consistency of abstractive text summarization*. arXiv preprint arXiv:1910.12840.

Kumar, A., Schwarzschild, A., Wen, Y.,... & Tramer, F. (2024). *Certified Robustness for Large Language Models*. arXiv preprint arXiv:2402.14882.

Lins, S. (2023, June 23). *Researchers find malicious AI models on Hugging Face platform*. The Register. Preuzeto s https://www.theregister.com/2023/06/23/malicious_ai_models_hugging_face/

Muñoz-González, L., Biggio, B., Demontis, A., Paudice, A., Wong, V., Oprea, A., Lupu, E., & Roli, F. (2017). Towards poisoning of deep learning algorithms with back-gradient optimization. *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, 27-38.

OpenAI. (2023). *GPT-4 Technical Report*. arXiv preprint arXiv:2303.08774. https://doi.org/10.48550/arXiv.2303.08774

Shafahi, A., Huang, W. R., Najibi, M., Su, Z., Studer, C., & Goldstein, T. (2018). Poison frogs! Targeted clean-label poisoning attacks on neural networks. *Advances in Neural Information Processing Systems, 31*.

Shu, F., G"unel, B.,... & Zhou, D. (2023). *Poisoning Language Models During Instruction Tuning*. arXiv preprint arXiv:2305.00944.

Wallace, E., Song, S., Shen, J., & Gardner, M. (2019). *Universal adversarial triggers for attacking and analyzing NLP*. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP).

Wiz Research. (2024, February 29). *Confused-Deputy attack on Microsoft's AI-powered platform*. Wiz Blog. Preuzeto s https://www.wiz.io/blog/confused-deputy-attack-on-microsofts-ai-powered-platform

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., ... & Wen, J. R. (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Zhu, Y., Li, G., Jin, Z., & Wang, S. (2023). On the negativity bias in deep learning-based models: A survey. *arXiv preprint arXiv:2305.06323*.