Meme Warfare

le contromisure dell'IA alla disinformazione

di Michael Yankoski , Walter Scheirer , Tim Weninger

Walter Scheirer

è un assistente professore presso il Dipartimento di Informatica e Ingegneria dell'Università di Notre Dame.
La sua ricerca è nell'area dell'intelligenza artificiale, con particolare attenzione al riconoscimento visivo, alla medicina legale dei media e all'etica.

Michael Yankoski

è un dottorando presso il Kroc Institute for International Peace Studies dell'Università di Notre Dame.
La sua ricerca di tesi esplora l'intersezione tra il cambiamento climatico antropogenico, la teoria della virtù, la costruzione strategica della pace e lo spostamento della popolazione umana.

Tim Weninger

è professore associato presso il Dipartimento di informatica e ingegneria dell'Università di Notre Dame.
La sua ricerca è all'intersezione tra social media, intelligenza artificiale e grafici.

I ricercatori stanno appena iniziando a comprendere la minaccia rappresentata dalla disinformazione multimediale (cioè visiva e audio). Dalle teorie della cospirazione di QAnon alle interferenze elettorali sponsorizzate dal governo russo, le campagne di disinformazione sui social media sono un fronte scoraggiante nella vita online e identificare queste minacce tra i post che miliardi di utenti dei social media caricano ogni giorno è una sfida. **Per aiutare a ordinare enormi quantità di dati, le piattaforme di social media stanno sviluppando sistemi di intelligenza artificiale per rimuovere automaticamente i contenuti dannosi principalmente attraverso l'analisi basata sul testo.** Ma queste tecniche non identificheranno tutta la disinformazione sui social media. Dopotutto, gran parte di ciò che le persone pubblicano sono foto, video, registrazioni audio e meme.

I ricercatori accademici, incluso noi, hanno lavorato per sviluppare sistemi di intelligenza artificiale con la sofisticatezza tecnologica per rilevare elementi multimediali falsi come foto e video (Yankoski, Weninger e Scheirer 2020). Nelle nostre analisi della disinformazione nei contenuti multimediali, quello che abbiamo scoperto è che i sofisticati contenuti falsi, spesso chiamati deepfake, non sono il problema più urgente. Attraverso le nostre piattaforme di ingestion e i motori di analisi dei media, ciò che stiamo vedendo non è la proliferazione di immagini, video e audio falsi che sono così reali da convincere qualcuno di qualcosa di falso, ma piuttosto la proliferazione di narrazioni che riaffermano emotivamente la convinzione che un il pubblico ha già (Theisen et al. 2020). Le manipolazioni dei deepfake su Internet sono di nicchia.Un meme di r/wallstreetbets descrive un analista finanziario come un mendicante.

Meme e il potere del superficiale.

Nel 1976, il biologo evoluzionista Richard Dawkins aveva bisogno di un termine per spiegare come gli artefatti culturali si evolvono nel tempo man mano che si diffondono nella società, replicandosi attraverso atti imitativi (Dawkins 2016). Ha coniato il termine "meme", un portmanteau dell'antica parola greca per imitazione, "mimeme" e la parola inglese "gene". Da allora, **i meme sono diventati una forma essenziale di comunicazione visiva. Tutto ciò che una persona può concepire ed esprimere visivamente è potenziale materiale meme.**

I meme in genere sono una specie di fake fake. In contrasto con i deepfake sempre più realistici e generati dall'intelligenza artificiale, si tratta di manipolazioni che variano nel valore della produzione da plausibile a ovviamente falso. In altre parole, sono facili da realizzare. I falsi superficiali possono anche essere immagini o video originali che qualcuno ha semplicemente rietichettato come raffigurante qualcos'altro o che ha modificato con sottigliezza per cambiare la

percezione del contenuto, ad esempio rallentando la frequenza dei fotogrammi video (Denham 2020). L'importante è che si replichino e si diffondano il più rapidamente possibile.
Gli Shallowfake sono molto meglio per la creazione di meme rispetto ai deepfake, che puntano al realismo. Per i falsi superficiali come i meme, spesso accade che meno corrispondono alla realtà, più possono essere efficaci nel diffondere online e persino influenzare il comportamento umano.
Prendi la saga di un gruppo Reddit chiamato r/wallstreetbets. Alla fine di gennaio 2021, i "Redditor", come sono conosciuti gli utenti della piattaforma, hanno scatenato un aumento astronomico del prezzo delle azioni di GameStop. Un negozio di videogiochi popolare nei primi anni 2000 (Eavis 2021), GameStop aveva perso trazione quando gli acquirenti e i giocatori si spostavano online. Per alcuni giorni selvaggi in pieno inverno, tuttavia, il prezzo delle azioni GameStop, che era rimasto nell'intervallo del dollaro a una cifra per gran parte del 2020, è schizzato a tre cifre e ha chiuso a $ 347 il 27 gennaio (Yahoo Finance 2021). Gli hedge fund che avevano scommesso contro il denaro emorragiato da GameStop e l'intero incidente ha portato a inchieste del Congresso (Flatley e Wasson 2021), un'indagine della Securities and Exchange Commission (Newmyer e Zapotosky 2021) e numerose azioni legali.
Una parte importante della storia di r/wallstreetbets sono i meme che i Redditor hanno usato per generare un'enorme risposta e per coordinare l'azione giocando con le emozioni e le convinzioni delle persone. Ad esempio, un meme descriveva un commentatore finanziario che aveva un'analisi negativa di GameStop come un mendicante. "Sulla strada per il negozio di alimentari, ho trovato Andrew Left al suo nuovo lavoro", recitava la didascalia del post. Anche se, in questo caso, i Redditor non stavano spingendo la disinformazione, le loro tattiche illustrano perfettamente il potere di superficiali falsi e meme per guidare il comportamento. **Un altro esempio, forse più allarmante, di meme che fungono da potenti strumenti per la disinformazione è stato il messaggio anti-vaccinazione durante la pandemia (Buts 2021). Gli utenti dei social media hanno pubblicato meme che mettono in dubbio il valore dei vaccini COVID-19, dicendo, ad esempio, che il vaccino antinfluenzale non ha eliminato l'influenza. Altri hanno collegato i vaccini all'autismo. Tali campagne di disinformazione basate sui meme potrebbero rafforzare le divisioni esistenti nel supporto ai vaccini (Funk e Tyson 2021). Un utente di Reddit ha pubblicato un esempio di meme anti-vaccinazione.**

AI per rilevare la disinformazione.

I ricercatori hanno investito risorse significative nella creazione di sofisticati sistemi di intelligenza artificiale per rilevare rapidamente le minacce non appena emergono nelle reti di social media online. **Esistono sistemi di rilevamento dell'incitamento all'odio** basati su testo (Peace Tech Lab; Woolley e Howard 2018; Technologies & International Development Lab nd). Uno di noi (WS) sta sviluppando sofisticati algoritmi di rilevamento della manipolazione delle immagini per rilevare immagini alterate (Theisen et al. 2020; Yankoski, Weninger e Scheirer 2020). I sistemi di rilevamento video Deepfake hanno la capacità di rilevare sia le irregolarità, come l'incoerenza del rumore nel video stesso (Guarnera, Giudice e Battiato 2020; Verdoliva 2020), sia le differenze tra i segnali affettivi contenuti nell'audio rispetto alle componenti video di un articolo sui media (Mittal et al. 2020).
Il problema è che queste tecnologie sono spesso isolate l'una dall'altra e quindi relativamente incapaci di rilevare le campagne di disinformazione basate sui meme. Sebbene i progressi tecnologici in ciascuna di queste varie aree siano lodevoli, i ricercatori devono ancora produrre sistemi di intelligenza artificiale abbastanza sofisticati da rilevare campagne coordinate progettate per manipolare il modo in cui i gruppi di persone si sentono riguardo a ciò in cui già credono, che è la motivazione per le campagne che coinvolgono i meme.

Questo tipo di analisi dell'intelligenza artificiale è su un altro livello completamente rispetto a tutti i sistemi e le tecnologie esistenti. Quest'altro livello è l'analisi semantica, una metodologia volta a mappare il significato delle stesse campagne di disinformazione. Per l'analisi semantica, non è sufficiente rilevare se un post contiene un'immagine manipolata, una clip audio falsa o un discorso di odio. **Gli algoritmi devono essere in grado di identificare campagne multimodali coordinate (vale a dire, testo/immagine/video) distribuite su piattaforme in modo da infiammare il panorama emotivo attorno alle convinzioni del pubblico. I sistemi di intelligenza artificiale dovranno comprendere la storia, l'umorismo, il riferimento simbolico, l'inferenza, la sottigliezza e l'insinuazione.**
Questo è un compito molto più difficile della semplice identificazione di contenuti multimediali manipolati, parole particolari in un lessico di discorsi d'odio o nuovi casi di una nota storia di "fake news". Piuttosto, ciò richiede lo sviluppo della capacità delle macchine e degli algoritmi di comprendere meglio gli strati complessi e sfaccettati della creazione di significato umano. I sistemi in grado di analizzare i complicati strati di significato dispiegati in superficiali falsi come i meme rappresentano l'avanguardia dei sistemi di intelligenza artificiale e sono l'avanguardia del prossimo futuro dello sviluppo tecnologico in questo spazio. In molti modi questo rappresenta la transizione dai sistemi percettivi AI esistenti ai sistemi cognitivi AI nascenti. L'enorme differenza di complessità e potenza di calcolo tra questi non può essere sopravvalutata.

Solo l'inizio.
I sistemi in grado di rilevare i deepfake in realtà non fanno molto per aiutare a contrastare la proliferazione di campagne di disinformazione che distribuiscono falsi superficiali progettati per ingrandire e amplificare le credenze preesistenti di un pubblico. Allo stesso modo, anche i sistemi incentrati sull'identificazione del testo problematico sono inadeguati al compito di analizzare meme e falsi superficiali. Ma immagina per un momento che i ricercatori di intelligenza artificiale siano in grado di sviluppare sistemi di analisi semantica e che diventi possibile rilevare queste campagne di disinformazione coordinate nel momento in cui si verificano. Cosa poi? L'intelligenza artificiale non sarà sufficiente. Le società di social media e i responsabili politici dovranno esaminare gli interventi, compreso lo sviluppo di software, l'alfabetizzazione mediatica e l'istruzione e persino le nuove norme sociali. In altre parole,
Tale approccio dovrebbe includere i seguenti elementi:

a) risposte a livello politico che considerino più attentamente la complessa relazione tra disinformazione, democrazia e libertà di parola;
b) b) accordi di condivisione delle informazioni progettati per coordinare la condivisione di informazioni tra agenzie governative e piattaforme di social media per l'identificazione rapida e il rallentamento delle campagne di disinformazione in tempo reale;
c) c) campagne di educazione all'alfabetizzazione mediatica che istruiscono e preparano gli utenti a identificare fonti di informazioni affidabili ea verificare i fatti o analizzare ulteriormente le fonti di informazioni che sembrano discutibili.

Alcune risposte alla disinformazione potrebbero implicare non soluzioni tecnologiche per rimuovere i contenuti, ma piuttosto tecniche per aiutare gli utenti a sapere cosa stanno consumando online. Gli sviluppatori di app dovrebbero prendere in considerazione lo sviluppo di una "metrica di coinvolgimento della disinformazione" simile ai contatori del tempo di utilizzo e ai tracker delle statistiche di coinvolgimento specifici delle app. Questi aiuterebbero gli utenti a saperne di più sul volume di disinformazione che stanno incontrando. Ci sono diversi ostacoli legati a questo, ma poiché la minaccia della manipolazione emotiva attraverso la disinformazione continua a crescere, i responsabili politici e gli sviluppatori dovranno sviluppare nuovi strumenti per aiutare gli utenti a navigare in un panorama in rapida evoluzione.

Man mano che una parte maggiore della popolazione umana ottiene un accesso affidabile e veloce a Internet, una percentuale crescente di persone diventerà suscettibile a campagne volte a manipolare, ingrandire e amplificare le proprie nozioni e disposizioni emotive preesistenti. Capire quando questo sta accadendo non richiede solo sistemi tecnologici in grado di identificare i deepfake, ma piuttosto sistemi con la capacità di identificare campagne di fake news coordinate tra le piattaforme.

Ma al di là dei necessari progressi tecnologici, abbiamo anche bisogno di una risposta multiforme che integri decisioni a livello politico, strategie di moderazione dei contenuti, accordi di condivisione delle informazioni, la coltivazione di nuove norme sociali sulla condivisione della disinformazione online, lo sviluppo di nuovi consumi/interazioni di disinformazione strumenti a livello di software e persino iniziative sociali volte ad aiutare le persone a interagire nella vita civica piuttosto che solo in un forum online.

Riferimenti

Ma, J. 2021. "Come i meme anti-vaxx si replicano attraverso la satira e l'ironia". *The Conversation* , 18 gennaio. https://theconversation.com/how-anti-vax-memes-replicate-through-satire-and-irony-153018

Dawkins, R., 2016. *Il gene egoista.* La stampa dell'università di Oxford.

Denham, H. 2020. "Un altro video falso di Pelosi diventa virale su Facebook". *The Washington Post,* 3 agosto. https://www.washingtonpost.com/technology/2020/08/03/nancy-pelosi-fake-video-facebook/

Eavis, P. 2021. "Cos'è GameStop, l'azienda, che vale davvero? Importa?" *The New York Times* , 1 febbraio. https://www.nytimes.com/2021/02/01/business/gamestop-how-much-worth.html

Flatley, D. e E. Wasson. 2021. "Il congresso prende di mira la mania delle scorte mentre Ocasio-Cortez strappa Robinhood". *Bloomberg* , 28 gennaio. https://www.bloomberg.com/news/articles/2021-01-28/brown-says-senate-panel-to-hold-hearing-amid-gamestop-frenzy

Funk, C. e A. Tyson.. 2021. "La quota crescente di americani afferma di volersi vaccinare contro il COVID-19 o di averlo già fatto". *Pew Research Center,* 5 marzo. https://www.pewresearch.org/science/2021/03/05/growing-share-of-americans-say-they-plan-to-get-a-covid-19-vaccine -o-già-avere/

Guarnera, L., O. Giudice, and S. Battiato. 2020. "DeepFake Detection analizzando le tracce convoluzionali". Atti della conferenza IEEE/Computer Vision Foundation sui workshop di Computer Vision and Pattern Recognition (CVPR). Virtuale, dal 14 al 19 giugno. https://openaccess.thecvf.com/content_CVPRW_2020/html/w39/Guarnera-DeepFake_Detection_by-Analyzing_Convolutional_Traces_CVPRW_2020_paper.html

Hwang, T. e C. Watt. 2020. "I deepfake stanno arrivando per la democrazia americana. Ecco come possiamo prepararci". *Il Washington Post,* 10 settembre . https://www.washingtonpost.com/opinions/2020/09/10/deepfakes-are-coming-american-democracy-heres-how-we-can-prepare/

Mittal, T., U. Bhattacharya, R. Chandra, A. Bera e D. Manocha. 2020. "Le emozioni non mentono: un metodo di rilevamento audiovisivo di deepfake che utilizza segnali affettivi". In Atti del 28th Association for Computing Machinery International Conference on Multimedia (MM '20). Virtuale, 12-16 ottobre. https://dl.acm.org/doi/10.1145/3394171.3413570

Newmyer, T. e M. Zapotosky. 2021. "I regolatori di Wall Street segnalano un approccio più duro all'industria dopo la frenesia di GameStop." *The Washington Post,* 14 febbraio. https://www.washingtonpost.com/business/2021/02/14/sec-gamestop/

PeaceTech Lab. nd "Lessicon di PeaceTech Lab". PeaceTech Lab. https://www.peacetechlab.org/toolbox-lexicons

Simonite, T. 2020. "Cosa è successo alla minaccia deepfake alle elezioni?" *Wired* , 16 novembre. https://www.wired.com/story/what-happened-deepfake-threat-election/

Laboratorio di tecnologie e sviluppo internazionale. e "Rafforzare l'impegno civico, rafforzare la democrazia e monitorare le elezioni. Che ruolo giocano i social media nelle elezioni?" Georgia Tech. http://tid.gatech.edu/dtd.html

Theisen, W., J. Brogan, PB Thomas, D. Moreira, P. Phoa, T. Weninger e W. Scheirer. 2021. "Scoperta automatica di generi di meme politici con diverse apparenze". Atti del Conferenza dell'Associazione Internazionale per l'Avanzamento dell'Intelligenza Artificiale su Web e Social Media (ICWSM). https://arxiv.org/abs/2001.06122

Yahoo Finanza. 2021. "GameStop Corp. (GME)" Yahoo Finance, 24 marzo. https://finance.yahoo.com/quote/GME/history/

Verdoliva, L. 2020. "Media Forensics e DeepFakes: una panoramica". *IEEE Journal of Selected Topics in Signal Processing* . 14 (5): 910-932. https://ieeexplore.ieee.org/document/9115874

Woolley, Samuel C. e PN Howard, eds. 2018. Propaganda computazionale: partiti politici, politici e manipolazione politica sui social media. La stampa dell'università di Oxford.
Yankoski, M., T. Weninger e W. Scheirer. 2020. Un sistema di allerta precoce per monitorare la disinformazione online, fermare la violenza e proteggere le elezioni. *Bollettino degli scienziati atomici.* 76 (2): 85-90. https://www.tandfonline.com/doi/full/10.1080/00963402.2020.17289