BULETINUL
UNIVERSITĂȚII NAȚIONALE DE APĂRARE „CAROL I"

https://buletinul.unap.ro

Cercetarea informațiilor false folosind tehnici de inteligență artificială: provocări

Researching disinformation using artificial intelligence techniques: challenges

Drd. Ștefan Emil REPEDE*

*Universitatea „Lucian Blaga", Sibiu, România
e-mail: stefan.repede@ulbsibiu.ro

Abstract

Articolul de față își propune să abordeze o serie de probleme generate de utilizarea modelelor de inteligență artificială în studiul fenomenului de creare și diseminare de informații false, plecând de la dificultățile existente în definirea și clasificarea termenilor consacrați, continuând cu exemplificarea modului de alcătuire a unora dintre cele mai des utilizate baze de date în cercetarea știrilor false, respectiv cu diferențele de abordare în catalogarea acestora.

The present article aims to address a series of problems generated by the use of artificial intelligence models for the study of the creation and dissemination of false information beginning from the difficulties in defining and classifying established terms, continuing by exemplifying the way some of the established databases in the research field of fake news are built and ending by noting the differences in their labeling.

Cuvinte-cheie:
știri false; dezinformare; managementul dezinformării; procesarea limbajului natural; NLP; inteligență artificială; învățare automată; securitate cibernetică.

Keywords:
fake news; misinformation; disinformation management; natural language processing; NLP; artificial intelligence; machine learning; cyber security.

https://doi.org/10.53477/2065-8281-23-20

1. Introducere

Rețelele sociale au fost văzute inițial ca o expresie a libertății de exprimare, a mobilizării pozitive și a democrației. În prezent, studiile care includ rețelele sociale le cataloghează mai degrabă drept amenințare la adresa democrației (Yerlikaya și Aslan 2020, 177-196). Platformele de social media pot fi utilizate intens pentru a manipula masele de către diferite tipuri de actori cu o agendă politică, socială, monetară sau radicală. Problema influenței mass-mediei asupra anumitor agende a apărut, în principal, după evenimentul cunoscut sub numele de Primăvara Arabă și a continuat în perioada alegerilor politice din SUA din 2016, Brexitului Regatului Unit, alegerilor prezidențiale din Franța din 2017, alegerilor Turciei din 2019, pandemiei de COVID-19 din 2019, până la escaladarea majoră a războiului ruso-ucrainean din 2022. Toate aceste evenimente au fost afectate de campanii de propagandă și de răspândirea largă a dezinformării și a știrilor false. Folosirea pe scară largă a informațiilor false în plan mondial a scos în evidență nevoia de creare a unor metode de semnalare și de combatere a utilizării de informații false pentru realizarea unor agende.

Odată cu dezvoltarea de tehnologii emergente, precum cele bazate pe inteligența artificială (IA), comunitatea științifică a propus diverse metode care implică utilizarea de tehnologii considerate ”state-of-the-art” pentru combaterea fenomenului de creare și diseminare de informații false.
Un exemplu de tehnologie studiată pentru clasificarea informațiilor scrise provenite din media online, așa-zisele știri false, a fost cea bazată pe utilizarea de modele de prelucrare a limbajului natural (Natural Language Processing sau NLP). Disciplina de prelucrare a limbajului natural, cunoscută și sub denumirea de lingvistică computațională, este o ramură a științei calculatoarelor care folosește IA la cercetarea limbajelor umane scrise și vorbite.

Informarea greșită (misinformation) și dezinformarea (disinformation) sunt clasificate drept al 9-lea punct de interes în raportul Agenției UE pentru Securitate Cibernetică, din octombrie 2022 (ENISA 2022), care afirmă că utilizarea de resurse de tipul cloud computing, de instrumente și algoritmi IA stă la baza fabricării de informații rău intenționate.

Raportul mai precizează că detectarea și minimizarea răspândirii informațiilor false pe rețelele sociale se numără și în prezent printre cele mai importante abordări tehnice pentru managementul dezinformării. Acest concept se referă la procesul de identificare, analiză și atenuare a răspândirii informațiilor false sau înșelătoare pentru a minimiza impactul negativ al acestora asupra indivizilor, organizațiilor și societății, în ansamblu. Gestionarea dezinformării implică etape precum monitorizarea mediilor țintă, detectarea potențialelor informații false, compararea acestora cu realitatea (fact-checking), catalogarea acestora, precum și stabilirea unui răspuns sau a contramăsurilor necesare contracarării.

Primul pas în managementul dezinformării este monitorizarea și detectarea (Schia și Gjesvik 2020, 413-428). Acesta implică monitorizarea răspândirii informațiilor pe diverse platforme și canale, cum ar fi rețelele sociale, site-urile web de știri și forumurile online. Instrumentele automate și analiza manuală sunt folosite pentru a identifica potențialele campanii de dezinformare și pentru a urmări răspândirea acestora. După faza de detecție, următorul pas este fact-checkingul faptelor, care presupune verificarea acurateței informațiilor prezentate. Acest pas este necesar pentru a determina răspunsul adecvat și contramăsurile care ar trebui luate (Schia și Gjesvik 2020, 413-428). Răspunsul și contramăsurile implică dezvoltarea și implementarea unor strategii de contracarare a răspândirii campaniei identificate sau de minimizare a impactului acesteia. Astfel de măsuri pot include publicitatea direcționată, campanii de mesaje publice sau solicitarea marcării ori eliminării informațiilor respective de pe platformele online pe care au fost postate (Schia și Gjesvik 2020, 413-428).

Managementul dezinformării este un proces continuu care necesită colaborarea dintre diverse părți interesate, inclusiv agenții guvernamentale, organizații media și grupuri ale societății civile. Abordarea dezinformării, în general, necesită o serie de abilități și expertiză din partea celor implicați, expertiză care include analiza datelor, monitorizarea rețelelor sociale, respectiv strategii de comunicare și relații publice. În general, gestionarea eficientă a dezinformării este esențială în era digitală de astăzi, unde răspândirea informațiilor false sau înșelătoare poate avea consecințe grave asupra indivizilor, organizațiilor și societății, în ansamblu (US Department of State 2023).

Din păcate, conform Raportului Global al Riscurilor 2023 (WEF 2023), managementul dezinformării este un proces care fie nu este inițializat, fie se află în stadiul dezvoltării sale timpurii, iar eficacitatea sa este văzută ca ineficientă sau extrem de ineficientă. Managementul dezinformării este o problemă de securitate cibernetică, în primul rând, deoarece răspândirea informațiilor false poate fi folosită ca un instrument pentru a manipula opinia publică sau pentru a crea o percepție falsă a realității (US Department of State 2023).

2. Știrile false – posibile definiri și categorisiri

O problemă inițială, apărută odată cu studiul științific a problematicii legate de crearea, diseminarea și consumul de știri false, a fost definirea conceptului. În principiu, există și la acest moment mai multe criterii de clasificare a informațiilor false, clasificări în care termenul de „știri false" poate să fie doar unul dintre modurile de manifestare propuse (Zafarani și alții 2019). Problema creată de semnificația termenului de știri false a fost larg dezbătută în ultimii ani, mai ales pe fondul mai multor campanii, semnalate în media internațională (Baptista și Gradim 2022, 632-645). Termenul de „știri false" este considerat, în prezent, drept inexact din punct de vedere tehnic, deoarece descrie o mare varietate de produse mass-media

(Gelfert 2018), deși este inclus în legislația românească la articolul 404, din Codul Penal (Parlamentul României 2009), care incriminează „comunicarea sau răspândirea, prin orice mijloace, de știri, date sau informații false ori de documente falsificate, cunoscând caracterul fals al acestora, dacă prin aceasta se pune în pericol securitatea națională", articolul fiind preluat în mare parte din articolul168^1 al Codului Penal al României, din anul 1968 (Parlamentul României 1968).

Discursul sau știrile care îndeamnă la violență sunt mai ușor de identificat față de cele ce incită la ură, defăimează statul de drept sau anumite grupuri sociale. Acestea din urmă nu sunt întotdeauna clar identificabile. Ceea ce este considerat inacceptabil pentru un individ poate fi ceva cu totul diferit pentru un altul. Această diferență de opinii creează o anumită ambiguitate față de ceea ce constituie discurs de incitare la ură într-un context digital, deoarece acest mediu poate include, pe lângă falsuri fățișe, erori în descrierea faptelor, comentarii de opinie, satiră politică sau inexactități. Pentru dezambiguizarea acestui concept la care se fac dese referiri, au fost propuse mai multe tipuri de clasificare a informațiilor false/înșelătoare, care vor fi abordate în cele ce urmează:

2.1. Clasificare în funcție de intenție

O propunere de clasificare a informațiilor false care cuprinde termenul de „știre falsă" a fost propus în cadrul Jurnalului Centrului de Excelență pentru Comunicații Strategice al NATO (NATO Strategic Communications Centre of Excellence 2020). Înțelegerea intenției din spatele campaniei/știrii false permite abordarea cauzelor dezinformării și adoptarea de măsuri de prevenire, educație sau responsabilizare. Clasificarea conține un număr de patru categorii, care țin cont de intenția din spatele propagării acestora:

Dezinformare (disinformation) – definită drept crearea și diseminarea intenționată de informații false/manipulate, cu intenția de a înșela/induce în eroare. Un exemplu de dezinformare poate fi considerat cel din 2022, când știrea falsă, conform căreia majoritatea românilor vor ca țara lor să părăsească NATO și UE și că nu există niciun partid care să poată valorifica politic această mișcare, a fost promovată de un post de radio autohton, implicat în repetate rânduri în promovarea dezinformării și a știrilor false. Cele afirmate de postul de radio au fost contrazise de sondajele de opinie (Necșuțu 2022a).

Informarea greșită (misinformation) – reprezintă acele informații false/înșelătoare care au fost distribuite fără intenția de a manipula sau de a induce în eroare. Principala diferență față de primul tip de dezinformare constă în intenția din spatele răspândirii acesteia. Un exemplu de preluare eronată de știri a avut loc în 26 februarie 2022, când postul de televiziune Antena 3 a prezentat din greșeală imagini dintr-un joc video, din 2013, intitulat Arma 3, ca fiind din războiul Rusiei împotriva Ucrainei (Radu 2022).

Informarea defectuoasă (MALinformation) – este un termen inventat de cercetătorul în mass-media Hossein Derakhshan, publicat drept coautor într-un raport al

Consiliului Europei, intitulat „Tulburarea informației" (Wardle și Derakhshan 2017) și adoptat, ulterior, de UNESCO. Aceasta se referă la o informație care este adevărată și care conține referințe corecte, dar este transmisă în mod intenționat negativ pentru a aduce un prejudiciu real sau pentru a semnala amenințarea iminentă de vătămare reală a unei persoane, organizații sau țări. Spre exemplu, o postare, făcută în arhiva intitulată "Paradise Papers" (Osborne 2017), despre investițiile offshore ale monarhiei britanice a dezvăluit că mulți membri ai monarhiei britanice da u avut investiții offshore evazioniste. Campania a avut scopul de a afecta imaginea monarhiei britanice, și nu de a informa publicul despre practicile ilicite ale Casei regale britanice.

Propaganda – informații, cu preponderență, părtinitoare, sau înșelătoare, care sunt răspândite cu scopul de a promova o cauză sau un punct de vedere politic. Un exemplu de astfel de informație este cel din anul 2022, când războiul dintre Rusia și Ucraina era descris ca un război purtat de NATO împotriva Moscovei (Cezar 2022). Afirmația despre o presupusă amenințare NATO împotriva Rusiei a fost vehiculată cu mult înainte de a fi preluată în România. A fost promovată de propaganda rusă, la început, pentru a justifica apetitul Moscovei pentru noi teritorii (invazia Georgiei în 2008, invazia Ucrainei în 2014 și 2022) și se bazează pe afirmații sovietice mai vechi, potrivit cărora NATO a „înconjurat" URSS cu bazele sale. Pe măsură ce forțele ruse au început să înregistreze înfrângeri în Ucraina, relatarea a fost modificată, afirmându-se că Rusia luptă, de fapt, împotriva NATO/Occidentului și că ucrainenii sunt folosiți drept carne de tun (Necșuțu 2022b).

Știri false (fake news) – reprezintă acele informații a căror falsitate este verificată și care sunt răspândite în mod intenționat. Un exemplu de știre falsă care s-a repetat pe spațiul românesc este cea conform căreia Olanda se opune aderării României la Schengen, pentru că portul Constanța amenință supremația portului Rotterdam. Această informație falsă a fost reluată în 2022, în contextul în care autoritățile de la București sperau ca România să fie admisă în spațiul Schengen până la sfârșitul anului (Peiu 2022). Acest tip de știre, apăruse cu 10 ani în urmă, lansată de Vocea Rusiei. Ciclul de știri false afirmă că Olanda nu va fi niciodată de acord ca România să adere la spațiul Schengen, de teamă că portul Constanța ar putea deveni cel mai mare port din Europa, având astfel un impact ireversibil asupra economiei olandeze, care se bazează în mare parte pe comerțul care intră și iese din portul Rotterdam. De fapt, o concurență reală între cele două porturi este în afara întrebării, deoarece portul Rotterdam este mai bine poziționat geografic și are o infrastructură și capacități operaționale superioare (Veridica 2022a).

2.2. Clasificare stilistică

O altă abordare a clasificării informațiilor false care include conceptele de știri false, dezinformare și propagandă, concentrându-se pe modul stilistic de compunere a materialelor media, care cuprinde cinci categorii a fost propus de Librăria Statului American Dakota de Nord (library-nd.com 2023):

Știri false sau mincinoase (false or deceptive) – acest concept se referă la informațiile fabricate sau manipulate în mod intenționat pentru a induce în eroare publicul (Gelfert 2018, 84-117). Această categorie poate include povești complet inventate, precum și știri care se bazează pe un element de adevăr, dar sunt distorsionate sau scoase din context pentru a susține o anumită agendă (Baptista și Gradim 2022). Un astfel de exemplu este zvonul lansat online privind decesul (ca urmare a unui atac de cord) lui George Soros din 15.05.2023, postat, inițial, pe un cont de Twitter (@PoliticsFAIRL) și preluat de conturi cu reputație. Afirmația nu avea nicio bază reală (LaMagdeleine 2023).

Informații înșelătoare (misleading) – articolele înșelătoare sunt acelea care conțin informații parțial sau complet inexacte ori care sunt prezentate într-un mod conceput pentru a induce în eroare cititorii sau spectatorii (Zafarani și alții 2019). Spre deosebire de știrile false sau mincinoase, articolele înșelătoare pot conține un element adevărat, dar acel adevăr poate fi scos din context sau prezentat într-un mod care este conceput pentru a promova o anumită agendă sau punct de vedere (Gelfert 2018, 84-117). Un exemplu de astfel de informații este cel conform căruia reforma sistemului de justiție, din 2022, va duce la subminarea Curții Constituționale, România își va pierde suveranitatea, Constituția nu va mai fi respectată, iar justiția românească se va desfășura la Bruxelles, pe placul Occidentului. Această știre înșelătoare a fost lansată în contextul dezbaterilor pe marginea legilor justiției din 2022. În realitate, modificarea legii privind statutul judecătorilor și procurorilor nu a făcut decât să alinieze sistemul de justiție românesc cu cel european, cu respectarea principiului supremației dreptului european (Veridica 2022b).

Conținut polarizant sau părtinitor (slanted/biased) – se referă la articole de știri sau rapoarte prezentate într-un mod care favorizează un anumit punct de vedere sau agendă (Schia și Gjesvik 2020, 413-428). Conținutul care se încadrează în această categorie nu este neapărat fals. Știrile raportează informații adevărate, dar o fac într-un mod părtinitor. Acest tip de partizanat poate fi politic, ideologic sau cultural și se poate manifesta în diferite moduri, inclusiv prin raportare selectivă, senzaționalism sau prin folosirea unui limbaj încărcat (Baptista și Gradim 2022, 632-645). Conținutul polarizant poate să reflecte dorința unei entități de a înclina opiniile cititorilor într-o anumită direcție sau poate să reflecte încercarea de a crea o știre memorabilă. Exemple legate de acest tip de conținut apar atunci când se prezintă a acele știri care redau o anumită ideologie/partid politic sunt prezentate de un canal de știri, iar celelalte sunt ignorate. Similar, un partid politic poate fi prezentat doar în manieră negativă. Un alt exemplu de media polarizantă poate fi dat de o reclamă care susține date științifice nedovedite (Drew 2023).

Date manipulate/modificate – acest concept se referă la textele, imaginile sau înregistrările video care au fost modificate sau editate intenționat într-un mod care denaturează conținutul original. Aceasta poate include utilizarea de imagini manipulate, videoclipuri editate sau citate selective, scoase din context (Zafarani și

alții 2019). Aceste date sunt, de regulă, folosite pentru a crea știri false sau pentru a susține o anumită agendă (Zellers și alții 2019). Un exemplu grăitor despre acest tip de informație falsă este înregistrarea video creată prin tehnologie de tipul "deepfake", în care președintele Ucrainei cerea conaționalilor săi să se predea Rusiei (The Telegraph 2022).

Piese umoristice din media (inclusiv toate formele sale, cum ar fi satira, parodia sau glumele) (Baptista și Gradim 2022, 632-645). Asemenea știri sunt intenționat fabricate sau exagerate pentru un efect comic (Figueira și Luciana 2017, 817-825). Spre deosebire de știrile false sau înșelătoare, știrile umoristice nu au scopul de a induce în eroare sau de a înșela, ci mai degrabă de a distra. Știrile satirice pot folosi umorul pentru a comenta probleme sociale sau politice, ori pentru a expune absurditatea sau ipocrizia (Gelfert 2018, 84-117). În ciuda faptului că aceste tipuri de știri nu au scopul de a înșela, uneori pot fi confundate cu știri autentice, mai ales dacă sunt distribuite în afara contextului sau fără o atribuire adecvată (Schia și Gjesvik 2020, 413-428). Un exemplu de satiră este știrea publicată de grupul media The Onion din SUA, cunoscut pentru conținutul umoristic al știrilor sale, intitulată „Jimmy Carter câștigă meciul de box împotriva lui Jake Paul". Această știre, deși este evident falsă, ca urmare a diferenței de vârstă dintre fostul președinte american și personalitatea social media, conține o fotografie editată, în care apar cei doi, iar conținutul articolului este prezentat după tiparul unui rezumat de gală de box (the ONION 2023).

2.3. Clasificare în funcție de impact și motivație

O clasificare care se dorește a fi mai exhaustivă și care ține cont de motivațiile din spatele creatorilor de fals, precum și de indicii de impact posibil, pe care fiecare tip de informație falsă îl poate avea, dacă este distribuită într-un mediu propice, a fost alcătuită de Asociația Europeană pentru Interesele Telespectatorilor (EAVI) și conține un total de 10 categorii, clasificate în funcție de impact (neutru, mic, mediu, mare) și de motivație (financiară, politică/putere, umoristică/entertainment, pasiune/extremism sau dezinformare de tipul misinformation) (EAVI 2022):

Propaganda – utilizată de guverne, corporații sau organizații nonprofit pentru a controla atitudini, valori și informații poate fi benefică sau dăunătoare, în funcție de motivația din spatele campaniei care poate fi creată pentru susținerea unei politici de stat sau pentru a crea o stare sociopolitică negativă (impact neutru și motivație legată de politică și pasiune).

Clickbaitul – oferă titluri senzaționale care atrag atenția, dar care induc în eroare, deoarece nu reflectă conținutul scris al materialului (impact redus, motivat de bani și valoarea de entertainment). Exemple de titluri de clickbait pot fi: „Nu ai să crezi..."; „X lucruri pe care trebuie să le cunoști..."; „Un truc ciudat..."; „Asta se va întâmpla dacă..."; „Cele mai bune X...". Exemple de acest tip de știri se regăsesc cu predilecție la rubricile mondene ale tabloidelor: „Ce spune Bianca Drăgușanu despre al doilea

copil. «Am tot ce îmi trebuie, cu siguranță» (Lixandru 2023). Din punctul de vedere al dezinformării, este relevant că acest tip de titluri pot fi utilizate pentru răspândirea unor campanii care conțin informații false.

Conținutul sponsorizat – are o formă similară cu cea a editorialelor, dar maschează reclame, fără a preciza consumatorilor acest lucru (impact redus și motivație financiară). Acest tip de publicitate este considerat dăunător, mai ales în rândul tinerilor care nu pot să facă diferența dintre o reclamă mascată (conținut sponsorizat) și o știre în mediul virtual (McAlpine 2019).

Satira și farsa – constituie un comentariu social umoristic care variază în calitate și care poate avea un înțeles subtil (impact redus și motivație umoristică). Această clasă este similară pieselor umoristice din media, cu precizarea că, de regulă, aceste conținuturi sunt polarizante, în favoarea unor agende, fiind, în fapt, părtinitoare.

Eroarea – știre sau informație care conține fapte false, ca urmare a unor erori involuntare (impactul este redus, iar motivația ține de misinformation). Acest tip de eroare poate perpetua o informație falsă prin neverificarea sursei inițiale.

Partizanatul – reprezintă acele știri care pretind a fi imparțiale, care includ interpretări ale faptelor, care au o factură ideologică și care includ doar faptele care confirmă o poziție sau o politică, ignorându-le pe celelalte (motivație ideologică și impact mediu). Acest tip de știri sunt folosite în cadrul campaniilor de propagandă pentru a crește nivelul de credibilitate al relatărilor prezentate. În astfel de cazuri, pot fi invitați experți veritabili sau pseudoexperți care să susțină un punct de vedere, dar care să pretindă imparțialitate. Partizanatul a fost utilizat în timpul alegerilor prezidențiale, din 2016, din SUA pentru a crea așa-zis imparțial o imagine pozitivă unuia dintre candidați (EAVI 2022).

Teoria conspirației – știri care explică simplist evenimente complexe, ca răspuns la frică sau nesiguranță. Acestea nu pot fi verificate științific, iar datele care neagă respectivele teorii sunt considerate dovezi care, de fapt, confirmă ipoteza (impact ridicat, motivație ideologică sau legată de misinformation). Astfel de teorii au fost vehiculate în timpul pandemiei de COVID-19 pentru a lega vaccinul de tehnologia 5G și de Microsoft, cu scopul de a crea o atitudine antioccidentală.

Pseudoștiința – știri care susțin teorii, precum vindecări miraculoase, mișcarea antivaccin, și care denaturează studii științifice reale prin afirmații exagerate sau false (impact ridicat, motivație politică sau financiară). Pseudoștiința a fost utilizată pe parcursul anilor 2020-2021, pentru a promova diferite teorii anti-UE. Astfel, o relatare preluată din zona estică preciza că măsurile anti-Covid, decise de autorități, sunt ineficiente. Campania de contestare a măsurilor sanitare luate pentru combaterea pandemiei de SARS-CoV-2 a continuat și în 2021, ani în care s-a pledat pentru tratamente alternative, care nu au primit aprobările autorităților

sanitare românești sau europene (Arbidol, Ivermectină), ce au fost promovate de medici obscuri sau de influenceri fără pregătire medicală, contestându-se, totodată, eficiența vaccinurilor anti-Covid prin intermediul unor pseudodate științifice – ori informații false, ori unele scoase din context. De remarcat, în această perioadă, a fost atenția acordată reacțiilor adverse la vaccinuri. Această focusare a venit din partea unor cadre medicale, de obicei cu specializări fără nicio legătură cu virusologia, respectiv, din partea unor experți în medicină alternativă (Gomboș 2021).

Dezinformarea – include o combinație de informații reale și false, alături de asocieri false, conținut prelucrat și titluri înșelătoare. Chiar dacă are scopul de a informa, autorul nu cunoaște faptul că informațiile utilizate sunt false (impact ridicat și motivația de a dezinforma). Dezinformarea implică, de regulă, o acțiune unitară, venită din mai multe surse și are un scop clar în spate, de regulă ideologic sau militar. O campanie de dezinformare faimoasă a avut loc în Al Doilea Război Mondial, când guvernul britanic descoperise radarul și nu dorea să dea indicii inamicului asupra acestui fapt. Prin urmare, au pornit o campanie de propagandă media în Marea Britanie, prin care au creat mitul conform căruia consumul de morcovi ajută viziunea nocturnă, iar piloții săi beneficiază din plin de această descoperire (Smith 2013).

Falsul/frauda – acel conținut fabricat în totalitate și răspândit cu intenția vădită de a dezinforma. Această categorie poate include tacticile de marketing de gherilă, roboți software sau comentarii contrafăcute. Acest conținut are scopul de a aduce câștig financiar sau influență politică/ideologică (impact ridicat, motivație politică sau financiară). Spre exemplu, conform raportului Meta privind atacurile cibernetice din primul trimestru al anului 2023 (Meta 2023), s-au eliminat 40 de conturi Facebook, 8 pagini și un grup pentru încălcarea politicii Meta împotriva comportamentului neautentic coordonat (coordinated inauthentic behavior – CIB). Rețeaua identificată își avea originea în Iran și a vizat, în principal, Israelul, Bahrainul și Franța, țări în care opera prin postarea de anunțuri, probabil false, cu scopul de a căpăta un nivel crescut de autenticitate, prin inserarea în cadrul unor forumuri tematice consacrate pe Facebook, Twitter, Youtube sau Telegram. Rețeaua ar fi putut fi utilizată ulterior în diferite scopuri politice sau pecuniare.

2.4. Inexistența unui consens

Clasificările prezentate abordează problema identificării și prezentării informațiilor false din mai multe perspective. Acestea conțin elemente comune, dar nu se suprapun în totalitate. Deși ierarhizările informațiilor false prezentate au pornit de la diferențieri variate, precum intenție, impact și motivație, nu s-au putut evita suprapuneri de sens, dar nu rezultă nicio clasificare universală și, implicit, o definire clară a diferitelor tipuri de informații false. De altfel, probabil că o definire unanimă nu este o posibilitate în viitorul apropiat, ca urmare a multitudinii de caracteristici și a formelor diverse de manifestare, pe care crearea și diseminarea de informații false o au. Reține, în continuare, atenția utilizarea termenului de „știri false", care,

deși nu se regăsește în toate clasificările, este unul dintre cele mai utilizate în cadrul cercetărilor care implică categorisirea de informații false obținute din spațiul public, deoarece implică și verificarea veridicității acestora.

3. Clasificarea informațiilor false și crearea de seturi de date pentru cercetare

Inexistența unui consens în privința definirii știrilor false și a multitudinii de situații socioculturale care imprimă perspective față de acest fenomen se afirmă drept una dintre problemele inițiale de care este necesar să se țină cont, în contextul utilizării inteligenței artificiale în cercetarea fabricării și răspândirii de informații false. Această diferență de catalogare generează o problemă care, deși nu este evidentă, devine esențială în procesul de antrenare de modele de IA și în testarea acestora. Problema este dată de lipsa datelor gata etichetate cu știri false. Pentru rezolvarea acestui neajuns, diferite proiecte online au creat seturi de date pe anumite direcții. Astfel de seturi de date de știri false, deja etichetate, sunt colecții de știri, în care fiecare informație, de obicei un articol de știri sau un titlu, a fost catalogată de un operator uman drept „adevărată" sau „falsă" (sau într-o variantă cu mai multe etichete) (ISOT Lab 2017). Aceste seturi de date sunt folosite pentru a instrui și a evalua modelele de învățare automată care pot clasifica automat articolele de știri sau titlurile ca adevărate sau false, pe baza modelelor și caracteristicilor învățate din datele etichetate (McIntire 2020).

Astfel de seturi pot fi folosite pentru a ajusta modelele de limbaj preantrenate, cum ar fi modelul NLP Bidirectional Encoder Representations from Transformers (BERT) (Ozbay și Alatas 2020), care au fost deja instruite pe volum mare de date generale de text. Acest lucru poate ajuta modelele să se adapteze la caracteristicile specifice ale setului de date de știri false și să devină tot mai performante în identificarea știrilor false (Devlin și alții 2019).

Seturile de date cu știri false și reale au fost concepute de echipe de cercetare din articole de știri din lumea reală, verificate anterior de către profesioniști media (Wang 2017, 422-426; Ahmed, Traore și Saad 2017, 127-138). Astfel de seturi de date sunt folosite pentru a instrui și a testa diferite metode de verificare automată a faptelor, folosind metode IA (Ozbay și Alatas 2020). Articolul va prezenta câteva seturi de date consacrate, compilate din știri, verificate anterior și etichetate în consecință.

Odată cu creșterea campaniilor de dezinformare, diferite organizații sau grupuri de verificare a faptelor s-au reunit ca răspuns, cu scopul de a educa publicul în a discerne informațiile adevărate de cele false prezentate de mass-media. Astfel de exemple pot include PolitiFact, factcheck.org sau Snopes. Facebook și-a construit chiar propria „Rețea internațională de verificare a faptelor" (IFCN 2016), care are peste 90 de semnatari din toată lumea, inclusiv din România.

Astfel de entități de verificare a faptelor oferă diferite tipuri de metode de clasificare a știrilor false sau adnotări pentru a consemna verificarea media. De exemplu, PolitiFact preia știrile actuale de la diferite instituții sau surse media, verifică afirmațiile acestora, folosind surse oficiale sau date statistice, și apoi acordă etichete ușor de utilizat, care arată cât de adevărat este un articol media, utilizând etichete, ca: adevărat, în mare parte adevărat, pe jumătate adevărat, preponderent fals, fals, pantaloni în flăcări (PolitiFact 2017); Snopes folosește un cod de etichetă similar: adevărat, în mare parte adevărat, în mare parte fals, fals, învechit, înșelătorie, nedovedit (Snopes, fără an); organizația românească de verificare a știrilor, factual.ro, folosește etichete precum: adevărat, parțial adevărat, trunchiat, fals, imposibil de verificat (Factual 2016). Toate aceste grupuri își bazează afirmațiile pe procentul informațiilor adevărate, în raport cu cele false existente în articolul de media analizat. Este relevant faptul că echipele de verificare a faptelor folosesc doar surse de acces deschis în timpul acestor verificări.

Dezvoltarea unui set de date de știri false implică, de obicei, un proces de colectare, adnotare și validare a articolelor de știri din diverse surse. O prezentare generală a pașilor implicați în crearea unui set de date de știri false și reale din știrile etichetate, prezentate de instituțiile media, include colectarea de articole de știri dintr-o varietate de surse, incluzând atât instituțiile de știri tradiționale, cât și sursele alternative de știri (IFCN 2016). În continuare, articolele trebuie verificate pentru conținutul de știri false (Kaggle 2018). Acest lucru este realizat de adnotatori umani experți care examinează articolele, respectând o metodologie prestabilită și transparentă. După identificarea articolelor cu știri false, acestea sunt adnotate cu etichete care să indice dacă sunt reale sau false (Factual 2016) de către aceeași adnotatori. În cele din urmă, setul de date trebuie validat pentru a se asigura că este fiabil și precis. Acest lucru se poate face prin compararea rezultatelor adnotărilor cu alte surse de date, cum ar fi site-urile web de verificare a faptelor sau alte surse de experți (Preda și alții 2022).

Prin colectarea știrilor deja etichetate de către aceste platforme, cercetătorii implicați în studii pe teme aferente știrilor false sunt capabili să genereze seturi mari de date cu conținut de știri false și reale certificate, fără a fi nevoie să le verifice personal și să le eticheteze. Baza etichetării unei știri constă, în principal, în proporția datelor false regăsite în aceasta. Deoarece nu există încă o modalitate standardizată de a eticheta mediile false și pentru a aborda anumite limitări care decurg din utilizarea mai multor etichete pentru seturi de date, cele mai multe seturi de date consacrate folosesc o clasificare binară pentru date, „fals" și „adevărat", și ignoră etichete precum „în mare parte adevărat", „pe jumătate adevărat" sau „imposibil de verificat".

Un exemplu de set de date folosit pe scară largă în cercetarea știrilor false, care se numește ISOT (Ahmed, Traore și Saad 2017), conține peste 1,2 milioane de articole de știri în diferite limbi, inclusiv în engleză, spaniolă și portugheză, și provine din diverse surse de verificare a faptelor, acoperind o gamă largă de subiecte și domenii. Setul de date a fost creat de Grupul de Cercetare pentru Securitatea Informației

https://doi.org/10.53477/2065-8281-23-20

și Tehnologia Obiectelor (ISOT) de la Universitatea Victoria din Canada (Ahmed, Traore și Saad 2017). Fiecare articol de știri din setul de date a fost deja verificat de adnotatori umani experți și este etichetat ca fiind real sau fals. În cadrul colecției de date, sunt furnizate și metadate suplimentare, cum ar fi sursa, autorul și data publicării. Setul de date este disponibil gratuit pentru descărcare și poate fi accesat prin intermediul site-ului web al grupului de cercetare ISOT (ISOT Lab 2017).

Un set de date similar utilizat este setul de date LIAR al PolitiFact (Wang 2017). PolitiFact este un site web nonpartizan de verificare a faptelor care evaluează acuratețea declarațiilor făcute de politicieni, persoane publice și de alte persoane proeminente. A fost fondat, în 2007, de Tampa Bay Times, un ziar din Florida, iar de atunci, s-a extins prin parteneriate cu alte organizații de știri. PolitiFact evaluează declarațiile pe o scară "Truth-O-Meter", care utilizează categorii, precum „adevărat", referitor la știrile reale, până la categoria de „pantaloni în flăcări", pentru știrile care nu conțin niciun element adevărat sau au o tematică fără sens. Site-ul oferă explicații detaliate și dovezi pentru evaluările sale și își propune să promoveze transparența și responsabilitatea în discursul public, ajutând oamenii să separe informațiile faptice de dezinformare și de propagandă. Setul de date, denumit „Mincinos, mincinos, pantaloni în flăcări", este verificat de PolitiFact și constă în 12.836 de declarații, făcute de politicieni, care sunt etichetate cu șase etichete diferite: pantaloni-în-flăcări[1], fals, abia adevărat, pe jumătate adevărat, în mare parte adevărat și adevărat. Setul de date a fost creat pentru a sprijini cercetarea în verificarea automată a faptelor și în detectarea știrilor false, fiind organizat ca o bază de date care conține știrea sau declarația, informații despre sursa declarației, persoana care a făcut reclamația, contextul în care a fost făcută și eticheta de verificare a faptelor, atribuită de PolitiFact (Wang 2017). A fost folosit pentru a antrena și a testa algoritmi de învățare automată, pentru verificarea faptelor și detectarea știrilor false și a fost citat în numeroase studii de cercetare (Wang 2017).

[1] "Pants-on-fire" – expresie colocvială din limba engleză care provine de la versurile preșcolarilor "liar, liar, pants, on fire!", semnificând că cineva a fost prins mințind.

Un alt set de date stabilit a fost compilat de KAGGLE și este cunoscut sub numele de Setul de date Kaggle Fake News (Kaggle 2018). Acesta constă dintr-o colecție de articole de știri, etichetate fie „false", fie „reale", pe baza acurateței și credibilității lor. Kaggle este o platformă populară din domeniul științei datelor și organizarea de competiții, care implică tehnologii de învățare automată. Setul de date pentru antrenare conține 20.800 de articole de știri, iar setul de testare conține 5.200 de articole de știri, inclusiv un amestec de articole de știri reale din surse de renume și articole de știri false din surse nesigure. Articolele de știri false au fost colectate din diverse surse de pe internet și au fost etichetate de oameni, pe baza veridicității lor. Articolele de știri reale au fost colectate de la organizații de știri reputate și au fost verificate ca fiind exacte de către organizațiile de verificare a faptelor (Kaggle 2018).

Setul de date include informații despre titlu, text, autor și data publicării fiecărui articol, precum și metadate, cum ar fi adresa URL sursă și numărul de distribuiri pe rețelele sociale. De asemenea setul de date Kaggle Fake News fost folosit în numeroase studii de cercetare și competiții de învățare automată și a contribuit la avansarea dezvoltării sistemelor automate pentru detectarea știrilor false.
Seturile de date, așa cum sunt cele prezentate în acest capitol, permit cercetătorilor să-și ajusteze modelele, să se concentreze pe obținerea unor rezultate superioare și să definească cu precizie clasele care trebuie luate în considerare de un model IA, atunci când are sarcini referitoare la identificarea știrilor false, a conținutului care dezinformează.

4. Utilizarea seturilor de date cu știri false în cercetare. Abordare binară sau multiclasă

Seturile de date prezentate în capitolul anterior permit abordarea unei cercetări din mai multe perspective: binară (adevărat contra fals) sau multiclasă (adevărat, parțial adevărat, neutru etc.). În procesul de detectare automată a știrilor false, o abordare binară este o metodă în care algoritmul de învățare automată este antrenat pentru a clasifica articolele de știri în două categorii: false sau reale (Kaliyar 2021, 11765-11788). În schimb, o abordare multiclasă este o metodă în care algoritmul este antrenat pentru a clasifica articolele de știri în mai mult de două categorii, cum ar fi: adevărat parțial, complet fals și adevărat, imposibil de clasificat, adevărat, satiră, propagandă etc. (Wang 2017).

O abordare binară este folosită în mod obișnuit în detectarea automată a știrilor false, deoarece simplifică problema și o face mai ușor de gestionat. În loc să clasifice știrile în mai multe categorii, o abordare binară necesită doar distincția între două clase: știri reale și știri false (Chen și alții 2017, 147-160). Acest lucru poate ajuta la îmbunătățirea preciziei clasificatorului. În plus, după cum s-a exemplificat în capitolele anterioare, de regulă setul de date a fost compus doar din știri dovedit false și dovedit reale și etichetat de creatori doar în cele două clase pentru a evita interpretările ulterioare (Kaggle 2018; ISOT Lab 2017). Alegând această abordare binară, setul de date etichetat existent poate fi utilizat fără a fi nevoie de crearea altor etichete suplimentare. Acest lucru economisește timp și resurse și permite cercetătorilor să se concentreze pe îmbunătățirea randamentului modelului pe cele două clase de interes. În plus, o abordare binară oferă un rezultat clar și ușor de înțeles, care poate fi folositor utilizatorilor finali (Kaliyar 2021). De exemplu, un agregator de știri sau o platformă de socializare poate avea nevoie să știe doar dacă un conținut este fals sau nu pentru a decide dacă îl afișează utilizatorilor. Un clasificator binar poate furniza rapid aceste informații, făcându-le mai utile pentru astfel de aplicații (Ahmed, Traore și Saad 2017).

În plus, o abordare binară poate simplifica și evaluarea performanței modelului. Valori consacrate de măsurare a eficienței unui model, precum acuratețea, precizia,

regresul și scorul F1 (un alt metric care măsoară acuratețea modelului, combinând-o cu scorul de regresie), sunt mai ușor de calculat și de interpretat atunci când avem de-a face cu numai două clase (Gnanambal și alții 2018, 3640-3644), această abordare ajutându-i astfel pe cercetători să compare și să selecteze mai ușor modelul cu cele mai bune performanțe (Yerlikaya și Aslan 2020).

Concluzie

Prezentul articol evidențiază problema gestionării știrilor false din perspectiva cercetării acestora și a unora dintre problemele întâmpinate în acest sens. O problemă inițială ține de identificarea unei tipologii comune a informațiilor false. Utilizarea inițială, inclusiv în texte de lege, a termenului generic de „știri false", în prezent, este considerată drept insuficientă pentru a surprinde totalitatea tipurilor de informații false existente, ea se menține și în prezent în sfera de cercetare, deoarece definiția acestuia permite cercetătorilor să eticheteze un tip de știre într-o manieră categorică, întrucât aceasta a fost deja verificată și nu există dubii în privința clasei în care se încadrează. Acest tip de abordare nu poate să surprindă realitatea cotidiană în toate formele sale de manifestare, dar reprezintă un prim pas în direcția corectă, deoarece produsul software rezultat poate, spre exemplu, să fie utilizat în semnalarea inițială rapidă a anumitor categorii de informații false sau campanii de dezinformare. Lucrarea contribuie la literatura existentă cu privire la detectarea automată a știrilor false, oferind un cadru pentru înțelegerea și abordarea acestui fenomen complex. Sperăm că această lucrare poate inspira cercetări și inovații în acest domeniu și poate deveni, totodată, o sursă de informare pentru factorii de decizie și practicienii implicați în managementul dezinformării.

Referințe

Ahmed, H., I. Traore și S. Saad. 2017. "Detection of online fake news using N-gram analysis and machine learning techniques." *International conference on intelligent, secure, and dependable systems in distributed and cloud environments* 127-138. doi:https://doi.org/10.1007/978-3-319-69155-8_9.

Bahad, Pritika, Preeti Saxena și Raj Kamal. 2019. "Fake News Detection using Bi-directional LSTM-Recurrent Neural nETWORK." *Procedia Computer Science* 165: 74–82.

Baptista, J.P. și A. Gradim. 2022. "A Working Definition of Fake News." *Encyclopedia* 632-645. doi:https://doi.org/10.3390/encyclopedia2010043.

Cezar, Nicholas. 2022. „De ce Războiul din Ucraina nu poate avea învingători." *Național.* https://www.national.ro/politica/de-ce-razboiul-din-ucraina-nu-poate-avea-invingatori-762950.html.

Chen, W., X. Xie, J. Wang, B. Pradhan, H. Hong, D. T. Bui și J. Ma. 2017. "A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility." *Catena* 151: 147-160. http://dx.doi.org/10.1016/j.catena.2016.11.032.

Devlin, J., M. W. Chang, K. Lee și K. Toutanova. 2019. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.*

Drew, Chris. 2023. "35 Media Bias Examples for Students." https://helpfulprofessor.com/media-bias-examples-for-students/.

EAVI. 2022. "Beyond Fake News – 10 Types of Misleading News." https://eavi.eu/beyond-fake-news-10-types-misleading-info/.

Emre, Celebi M, Kemal Aydin. 2018. "Unsupervised Learning Algorithms." doi:https://doi.org/10.1007/978-3-319-24211-8.

ENISA, European Union Agency for Cybersecurity. 2022. "ENISA Threat Landscape 2022." https://www.enisa.europa.eu/publications/enisa-threat-landscape-2022.

Factual. 2016. https://www.factual.ro/.

Figueira, A. și O. Luciana. 2017. "The current state of fake news: challenges and opportunities." *Procedia Computer Science* 121: 817-825. doi:https://doi.org/10.1016/j.procs.2017.11.106.

Gelfert, A. 2018. "Fake news: A definition." *Informal logic* 38 (1): 84-117. doi:https://doi.org/10.22329/il.v38i1.5068.

Gnanambal, S., M. Thangaraj, V. T. Meenatchi și V. Gayathri. 2018. "Classification algorithms with attribute selection: an evaluation study using WEKA." *International Journal of Advanced Networking and Applications* 9 (6): 3640-3644. https://oaji.net/articles/2017/2698-1528114152.pdf.

Gomboș, Cătălin. 2021. „România 2021: Top FAKE NEWS & DEZINFORMĂRI demontate de Veridica." https://www.veridica.ro/stiri-false/romania-2021-top-fake-news-dezinformari-demontate-de-veridica.

IFCN, International Fact-Checking Network. 2016. "Verified signatories of the IFCN code of principles." https://ifcncodeofprinciples.poynter.org/signatories.

ISOT Lab. 2017. "ISOT Fake News Dataset." https://www.uvic.ca/ecs/ece/isot/datasets/fake-news/index.php.

Kaggle. 2018. "Fake News Detection." https://www.kaggle.com/jruvika/fake-news-detection.

Kaliyar, R.K., Goswami, A. și Narang, P. 2021. "FakeBERT: Fake news detection in social media with a BERT- based deep learning approach." *Multimedia Tools and Applications* (80): 11765–11788. doi:10.1007/s11042-020-10183-2.

LaMagdeleine, Izz Scott. 2023. "No, George Soros Is Not Dead." *Snopes.* https://www.snopes.com/fact-check/george-soros-is-not-dead/.

library-nd.com. 2023. https://library-nd.libguides.com/fakenews/categories.

Lixandru, Livia. 2023. „Ce spune Bianca Drăguşanu despre al doilea copil. «Am tot ce îmi trebuie, cu siguranță»." *Libertatea.* https://www.libertatea.ro/entertainment/ce-spune-bianca-dragusanu-despre-al-doilea-copil-4566539.

Magdin, Radu. 2013. „Constanta: locul unde se ciocnesc interesele. Internaționale." https://cursdeguvernare.ro/constanta-locul-unde-se-ciocnesc-interesele-internationale.html.

McAlpine, Kat J. 2019. "Most people can't tell native advertising apart from actual news articles, according to new research." https://www.futurity.org/sponsored-content-real-news-1961062/.

McIntire, G. 2020. "Fake News Dataset." https://github.com/pmacinec/fake-news-datasets/tree/eb85398bab558791c9f879e9f96ce72a471d2cc9.

Meta. 2023. "Quarterly Adversarial Threat Report Q1 2023." https://about.fb.com/wp-content/uploads/2023/05/Meta-Quarterly-Adversarial-Threat-Report-Q1-2023.pdf.

NATO Strategic Communications Centre of Excellence. 2020. "Defence Strategic Communications ." *Academic Jurnal Volume 8* (8). doi:DOI: 10.30966/2018.RIGA.8.

Necșuțu, Mădălin. 2022a. „Dezinformare: Majoritatea românilor vor ieșirea țării din NATO și UE." https://www.veridica.ro/dezinformare/dezinformare-majoritatea-romanilor-vor-iesirea-tarii-din-nato-si-ue.

—. 2022b. "Disinformation: The West is fighting Russia using Ukraine as proxy." https://www.veridica.ro/en/disinformation/disinformation-the-west-is-fighting-russia-using-ukraine-as-proxy.

Osborne, Hilary. 2017. "Revealed: Queen's private estate invested millions of pounds offshore." *The Guardian.* https://www.theguardian.com/news/2017/nov/05/revealed-queen-private-estate-invested-offshore-paradise-papers.

Ozbay, Feyza Altunbey și Bilal Alatas. 2020. "Fake news detection within online social media using supervised artificial intelligence algorithms." *Physica A: statistical mechanics and its applications* 540. doi:https://doi.org/10.1016/j.physa.2019.123174.

Parlamentul României. 1968. „Codul Penal din 21 iulie 1968 (**republicat**)." https://legislatie.just.ro/Public/DetaliiDocument/38070.

—. 2009. „Codul Penal din 17 iulie 2009, Legea nr. 286/2009." https://legislatie.just.ro/Public/DetaliiDocument/223635.

Peiu, Petrișor. 2022. „Blocadă olandeză la porțile castelului Schengen. Pericolul naționalismului lipsit de inteligență." *Gândul.* https://www.gandul.ro/opinii/blocada-olandeza-la-portile-castelului-schengen-pericolul-nationalismului-lipsit-de-inteligenta-19860233.

PolitiFact. 2017. https://www.politifact.com/.

Preda, A., S. Ruseti, S. M. Terian și M. Dascalu. 2022. "Romanian Fake News Identification using Language Models." doi:DOI: 10.37789/rochi.2022.1.1.13.

Radu, Cristina. 2022. „Antena 3 a prezentat din eroare imagini dintr-un joc video din 2013 ca fiind din războiul Rusiei împotriva Ucrainei." *Libertatea.* https://www.libertatea.ro/stiri/antena-3-a-prezentat-din-eroare-imagini-dintr-un-joc-video-din-2013-ca-fiind-din-razboiul-rusiei-impotriva-ucrainei-4005144.

Rashkin H, Choi E, Jang JY, Volkova S, Choi Y. 2017. "Truth of varying shades: Analyzing language in fake news and political fact-checking." *Proceedings of the 2017 conference on Empirical Methods in Natural Language Processing, EMNLP.* 2931-2937. doi:10.18653/v1/D17-1317.

Schia, N.N. și L. Gjesvik. 2020. "Hacking democracy: managing influence campaigns and disinformation in the digital age." *Journal of Cyber Policy* 5 (3): 413-428. doi:https://doi.org/10.1080/23738871.2020.1820060.

Smith, K. Annabelle. 2013. "A WWII Propaganda Campaign Popularized the Myth That Carrots Help You See in the Dark." *Smithsonian Magazine*. https://www.smithsonianmag.com/arts-culture/a-wwii-propaganda-campaign-popularized-the-myth-that-carrots-help-you-see-in-the-dark-28812484/.

Snopes, Snopes Media Group Inc. fără an. https://www.snopes.com/. Accesat 2 februarie 2023.

the ONION. 2023. "Jimmy Carter Wins Boxing Match Against Jake Paul." https://www.theonion.com/jimmy-carter-wins-boxing-match-against-jake-paul-1850487520.

The Telegraph. 2022. "Deepfake video of Volodymyr Zelensky surrendering surfaces on social media." https://www.youtube.com/watch?v=X17yrEV5sl4.

US Department of State. 2023. "Disarming Disinformation: Our Shared Responsibility." https://www.state.gov/disarming-disinformation/.

Veridica. 2022a. "Fake news: The Netherlands opposes Romania's Schengen accession because the port of Constanța threatens the supremacy of the port of Rotterdam." https://www.veridica.ro/en/fake-news/fake-news-the-netherlands-opposes-romanias-schengen-accession-because-the-port-of-constanta-threatens-the-supremacy-of-the-port-of-rotterdam.

—. 2022b. „Fake news: The reform of the justice system leads to the undermining of the Constitutional Court and Romania losing its sovereignty." https://www.veridica.ro/en/fake-news/fake-news-the-reform-of-the-justice-system-leads-to-the-undermining-of-the-constitutional-court-and-romania-losing-its-sovereignty.

Wang, W. Y. 2017. "«Liar, Liar Pants on Fire»: A new benchmark dataset for fake news detection." *Proceedings of the 55th annual meeting of the association for computational linguistics* 422-426. https://arxiv.org/abs/1705.00648.

Wardle, Claire și Hossein Derakhshan. 2017. "Information Disorder: Toward an interdisciplinary framework for research and policy making." https://edoc.coe.int/en/media/7495-information-disorder-toward-an-interdisciplinary-framework-for-research-and-policy-making.html.

WEF, World Economic Forum. 2023. "The Global Risks Report 2023." https://www3.weforum.org/docs/WEF_Global_Risks_Report_2023.pdf.

XIA, Xin și LO, David. 2018. "Feature Engineering for Machine Learning and Data Analytics." *Feature* (CRC Press) 335-358. https://ink.library.smu.edu.sg/sis_research/4362 .

Yerlikaya, Turgay și Seca Toker Aslan. 2020. "Social Media and Fake News in the Post-Truth Era." *Insight Turkey 22.2* 177-196.

Yuandong Luan și Shaofu Lin. 2019. "Research on Text Classification Based on CNN and LSTM." *International Conference on Artificial Intelligence and Computer Applications (ICAICA)*. doi:https://doi.org/10.1109/ICAICA.2019.8873454.

Zafarani, Reza, Xinyi Zhou, Kai Shu și Huan Liu. 2019. "Fake News Research: Theories, Detection Strategies, and Open Problems." *KDD '19: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. doi:https://doi.org/10.1145/3292500.3332287.

Zellers, R., A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner și Y. Choi. 2019. ”Defending Against Neural Fake News.” https://rowanzellers.com/grover.

Zhou, Z., H. Guan, M. M. Bhat și J. Hsu. 2019. ”Fake news detection via NLP is vulnerable to adversarial attacks.” doi:https://doi.org/10.48550/arXiv.1901.09657.