BULETINUL
UNIVERSITĂȚII NAȚIONALE DE APĂRARE „CAROL I"

https://buletinul.unap.ro

O comparație a modelelor de inteligență artificială folosite pentru detectarea știrilor false

A comparison of artificial intelligence models used for fake news detection

Student doctorand Ștefan Emil REPEDE*
Prof. dr. Remus BRAD**

*Universitatea „Lucian Blaga" din Sibiu
e-mail:stefan.repede@ulbsibiu.ro
**Universitatea „Lucian Blaga" din Sibiu
e-mail: remus.brad@ulbsibiu.ro

Abstract

Acest articol propune o comparație a modelelor actuale de ultimă generație de procesare a limbajului natural (NLP), optimizate pentru detectarea știrilor false, pe baza unui set de metrici. De asemenea, lucrarea dorește să evalueze eficacitatea acestora, ca parte a unei structuri de management al dezinformării. Necesitatea unei dezvoltări a acestui domeniu vine ca răspuns la răspândirea copleșitoare și nereglementată a știrilor false care reprezintă una dintre dificultățile majore în epoca actuală. Dezvoltarea tehnologiilor IA are un impact direct asupra creării și răspândirii dezinformării și a știrilor false, ca urmare a utilizărilor multiple pe care tehnologia le poate avea. În prezent, tehnicile de învățare automată sunt utilizate pentru dezvoltarea modelelor de limbaj mari (LLM). Aceste evoluții în știință sunt folosite și în campaniile de dezinformare. Legat de această problemă, conceptul de management al dezinformării a apărut ca o problemă de securitate cibernetică integrantă în peisajul actual al amenințărilor venite din planul virtual.

This article aims to compare current state-of-the-art natural language processing models (NLP) fine-tuned for fake news detection based on a set of metrics and asses their effectiveness as a part of a disinformation management structure. The need for a development of this area comes as a response to the overwhelming and unregulated spread of fake news that represents one of the current major difficulties in today`s era. The development of AI technologies has a direct impact over the creation and spreading of misinformation and disinformation as a result of the multiple uses that technology may have. Currently, machine learning techniques are used for the development of large language models (LLM). These developments in science are also used in disinformation campaigns. Related to this matter the concept of disinformation management has arisen as a cybersecurity issue integral in the current cyber threat landscape.

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; cybersecurity.

1. Introducere

Știrile false sunt o problemă serioasă care poate induce în eroare și poate manipula oamenii să creadă narațiuni false sau părtinitoare. Știrile false sunt un termen care se referă la informații false sau înșelătoare care sunt prezentate ca știri faptice. Știrile false pot avea consecințe grave pentru societate, cum ar fi influențarea opiniei publice, răspândirea dezinformării și subminarea încrederii în jurnalism. Prin urmare, este important să se dezvolte metode eficiente de detectare și combatere a știrilor false. Detectarea automată a știrilor false este o sarcină provocatoare care necesită metode avansate de inteligență artificială (IA) pentru a analiza conținutul și sursa articolelor de știri. În acest articol de cercetare, comparăm diferite metode IA aplicate pentru detectarea automată a știrilor false. Folosim diverse metrici, cum ar fi acuratețea, precizia, retragerea și scorul F1 pentru a evalua performanța diferitelor metode. Examinăm tehnicile de învățare supravegheată, cum ar fi mașinile de suport vectorial, BayesNet sau arbori de decizie care folosesc date etichetate pentru a clasifica articolele de știri ca false sau reale. De asemenea, discutăm despre modele de învățare profundă și transformatoare, cum ar fi rețelele neuronale convolute (CNN), rețelele neuronale recurente (RNN) și BERT, care pot captura caracteristici complexe și relații semantice din datele text. Explorăm tehnici de învățare nesupravegheată, cum ar fi gruparea, modelarea subiectelor și detectarea anomaliilor care pot identifica știri false, fără cunoștințe sau etichete anterioare. Examinăm unele sisteme bazate pe reguli care utilizează reguli predefinite sau euristici pentru a detecta știri false, bazate pe caracteristici lingvistice sau stilistice. În cele din urmă, prezentăm modele hibride care combină diferite metode IA pentru a obține rezultate mai bune în detectarea știrilor false. Seturile de date utilizate pentru testare și validare se vor concentra pe setul de date de știri false ISOT (University of Victoria 2017) sau pe altele similare. De asemenea, vom discuta punctele forte și limitările fiecărei clase și vom oferi sugestii pentru direcțiile viitoare de cercetare și utilizare.

2. O comparație a metodelor IA aplicate pentru detectarea automată a știrilor false

Până în prezent, au fost folosite diverse metode IA pentru sarcini de clasificare binară referitoare la detectarea automată a știrilor false (Kaliyar, Goswami și Narang 2021c). Evaluarea performanței pentru seturile de date de știri false este măsurată folosind valori stabilite (Ozbay și Alatas 2020) care le permit cercetătorilor să compare performanța diferitelor modele și să identifice metodele mai eficiente în detectarea știrilor false.

3. Metrici utilizați pentru evaluare

Acuratețea, precizia, regresia și scorul F1 sunt valorile utilizate de cercetători în mod obișnuit pentru a evalua performanța modelelor de clasificare, deoarece oferă o mai bună înțelegere a performanțelor (Liu și alții 2019), fiind descrise după cum urmează:

3.1 Acuratețe – măsoară proporția de instanțe clasificate corect din numărul total de articole. Se calculează ca (Adevărate Pozitive + Adevărate Negative) / Total Instanțe. Acuratețea este un indicator, utilă atunci când clasele sunt echilibrate (numărul de știri false este aproximativ egal cu cel al știrilor reale) (Kaliyar, Goswami şi Narang 2021c).

3.2 Precizie – măsoară proporția de articole adevărate, identificate corect de către model (adevărate pozitive) dintre toate cazurile (inclusiv false), clasificate drept știri adevărate. Se calculează ca Adevărate Pozitive/(Adevărate Pozitive + False Pozitive). Precizia măsoară cât de exacte sunt predicțiile pozitive și cât de des modelul identifică corect cazurile pozitive adevărate (Devlin şi alții 2019).

3.3 Regresia – măsoară proporția de pozitive adevărate dintre toate cazurile care sunt, de fapt, pozitive. Se calculează ca Adevărate Pozitive/(Adevărate Pozitive + False Negative). Regresia măsoară cât de bine poate modelul să găsească toate cazurile pozitive din setul de date (Kaliyar, Goswami şi Narang 2021a).

3.4 Scorul F1 – este media armonică a preciziei și a reamintirii, oferind o măsură echilibrată între cele două metrici. Se calculează ca 2 * (Precizie * Regresie)/(Precizie + Regresie). Scorul F1 este o măsură generală bună a performanței modelului, mai ales atunci când setul de date este dezechilibrat. În clasificarea binară, scorurile adevărat pozitive (TP) sunt numărul de instanțe care sunt clasificate corect drept pozitive, cele fals pozitive (FP) sunt numărul de instanțe care sunt clasificate incorect drept pozitive, cele adevărat negative (TN) sunt numărul de instanțe care sunt clasificate corect drept negative și fals negative (FN) sunt numărul de cazuri care sunt clasificate incorect ca negative (Ozbay şi Alatas 2020).
Metodele actuale de IA de ultimă generație, utilizate în cercetarea acestui subiect pot fi împărțite în următoarele categorii generale:

4. Tehnici de învățare supravegheată utilizate în detectarea știrilor false

Învățarea supravegheată este o metodă IA care a fost folosită pentru clasificarea știrilor false în diferite moduri. Aceasta implică antrenarea unui model de învățare automată pe un set de date, etichetat de articole de știri care sunt clasificate ca fiind reale sau false. Modelul învață să identifice modele de date și să generalizeze la exemple noi, nevăzute. Alegerea algoritmului de învățare automată depinde de caracteristicile specifice ale setului de date, dar algoritmii populari includ metode precum regresia logistică, mașinile de suport vectorial și arbori aleatorii (Ozbay şi Alatas 2020). Următoarele metode sunt luate în considerare pentru comparație:

4.1 BayesNet – rețeaua bayesiană, cunoscută și sub numele de rețea Bayes, este un model grafic probabilist, utilizat pentru raționamentul în condiții de incertitudine.

Este numit după reverendul Thomas Bayes, un statistician britanic din secolul al XVIII-lea care a dezvoltat teorema Bayes. Într-o rețea bayesiană, variabilele sunt reprezentate prin noduri, iar relațiile dintre ele sunt reprezentate prin muchii direcționate (Langley, Wayne și Thompson 1992, 223-228). Nodurile pot reprezenta variabile fie observabile, fie ascunse, iar marginile reprezintă dependențe condiționate între ele.

4.2 Jrip – Jumping Rule-based Information Processing a fost dezvoltat de WW Cohen și este un algoritm de clasificare, bazat pe arbore de decizie, utilizat pentru sarcinile de învățare automată, în special pentru sarcinile de clasificare (Cohen 1995, 115-123). Algoritmul „sare" între reguli, selectând cea mai bună regulă la fiecare nod pentru a clasifica datele. JRIP diferă de alți algoritmi bazați pe arbore de decizie prin faptul că folosește mai degrabă o abordare bazată pe reguli, decât o abordare pur arbore de decizie. Aceasta înseamnă că, în loc să se bazeze doar pe structura de ramificare a arborelui de decizie, generează un set de reguli care ghidează procesul de clasificare mai precis (Jijo și Abdulazeez 2021, 20-28).

4.3 OneR – denumit și One Rule, este un algoritm de clasificare simplu și interpretabil, propus de Holt care este folosit pentru sarcinile de învățare automată (Holte 1993, 63-90). Funcționează prin identificarea celui mai important atribut sau caracteristică dintr-un set de date pe care îl folosește pentru a crea o regulă de clasificare. OneR se numește „o singură regulă", deoarece folosește o singură regulă pentru a clasifica datele, ceea ce face ușor de interpretat și explicat (Chantar și alții 2020).

4.4 Decision Stump – Un clasificator Decision Stump este un algoritm simplu de clasificare binar, care este adesea folosit ca element de bază pentru modele de învățare automată mai complexe (Sammut 2017). Funcționează prin crearea unui arbore de decizie cu un singur nivel, numit „ciot", în care fiecare nod este o regulă de decizie, bazată pe o singură caracteristică sau atribut (Jijo și Abdulazeez 2021, 20-28). Clasificatorul Decision Stump se numește „ciot", deoarece constă dintr-un singur nivel, spre deosebire de arbori de decizie mai complecși cu mai multe nivele (Sammut 2017).

4.5 ZeroR – Clasificatorul ZeroR este un algoritm simplu, de bază pentru clasificare, care prezice întotdeauna cea mai frecventă clasă din setul de date de antrenament (Devasena și alții 2011). Se numește „ZeroR", deoarece nu folosește nicio caracteristică de intrare pentru a face predicții, ci se bazează doar pe distribuția de clasă a datelor de antrenament. Clasificatorul ZeroR este adesea folosit ca model de bază pentru a compara performanța altor clasificatoare mai complexe.

4.6 SGD – Clasificatorul Stochastic Gradient Descent este un algoritm pentru antrenarea clasificatoarelor liniare și a modelelor de regresie în învățarea automată (Chollet 2017, 48-50). Este deosebit de util pentru seturile de date mari, deoarece actualizează parametrii modelului, folosind loturi mici de date la un moment dat, mai degrabă decât întregul set de date, ceea ce poate duce la o convergență mai rapidă

și la cerințe mai mici de memorie. Clasificatorul SGD este utilizat în mod obișnuit pentru sarcini, precum clasificarea textului, clasificarea imaginilor și procesarea limbajului natural.

4.7 CVPS – Selectarea parametrilor CV (CVPS) se referă la procesul de selectare a celor mai buni hiperparametri pentru un model de învățare automată, folosind validarea încrucișată (Varma și Simon 2006, 1-8). Această tehnică implică împărțirea datelor de antrenament în k pliuri, antrenarea modelului cu k-1 pliuri și validarea acestuia cu pliul rămas. Acest proces se repetă pentru fiecare pliu, iar performanța medie este utilizată pentru a selecta cei mai buni hiperparametri.

4.8 RFC – Clasificatorul filtrat randomizat (RFC) este un algoritm de învățare automată care combină conceptele de selecție și clasificare a caracteristicilor. Este conceput pentru a selecta un subset de caracteristici relevante din datele de intrare, înainte de a antrena un model de clasificare (Alam și alții 2021). Prin selectarea unui subset de caracteristici relevante, algoritmul poate îmbunătăți eficiența și acuratețea modelului de clasificare (Alam și alții 2021). În plus, prin antrenarea mai multor modele pe diferite subseturi de date, algoritmul poate oferi predicții mai solide și poate reduce riscul de supraadaptare.

4.9 LMT – Arborele modelului logistic (LMT) combină arbori de decizie cu regresia logistică pentru a construi un model de clasificare. A fost dezvoltat pentru a aborda limitările arborilor de decizie standard, care pot suferi de supraadaptare și nu pot captura interacțiuni complexe între caracteristicile de intrare (Chen și alții 2017).

4.10 LWL – Învățarea ponderată local (LWL) este un algoritm de învățare supravegheat care utilizează o abordare neparametrică pentru a afla relația de bază dintre caracteristicile de intrare și variabilele de ieșire (Tuyen și alții 2021). Acest lucru permite LWL să capteze relații complexe, neliniare în date, evitând în același timp supraadaptarea.

4.11 CvC – Clasificarea prin clusterizare este o metodă de învățare semisupravegheată care utilizează algoritmi de grupare pentru a crea etichete pentru datele neetichetate (Bergsma și alții 2013). Metoda funcționează prin gruparea, mai întâi, a datelor etichetate în diferite grupuri, în funcție de caracteristicile lor. Apoi, punctele de date neetichetate sunt atribuite acelorași grupuri ca și punctele de date etichetate. În cele din urmă, cea mai comună etichetă din cadrul fiecărui cluster este atribuită punctelor de date neetichetate din acel cluster.

4.12 WIHW – Weighted Instances Handler Wrapper este o tehnică de învățare automată care ajustează distribuția de clasă a unui set de date prin atribuirea de ponderi fiecărei instanțe, pe baza etichetei sale de clasă (Khosravi, Khozani și Mao 2021). Wrapperul WIH funcționează prin potrivirea unui clasificator la setul de date original și apoi prin modificarea setului de date prin atribuirea de ponderi fiecărei

instanțe, pe baza clasei sale. Instanțele care sunt clasificate greșit au o pondere mai mare, în timp ce instanțelor care sunt clasificate corect li se acordă o pondere mai mică. Acest proces se repetă până când performanța clasificatorului pe setul de date modificat converge.

4.13 Ridor – Acest model este un algoritm de clasificare, bazat pe arbore de decizie care utilizează conceptul de inducție bazată pe reguli pentru a îmbunătăți performanța clasificării (Lakmali și Haddela 2017). Funcționează prin construirea unui arbore de decizie, în care fiecare nod reprezintă un test pe un atribut, iar fiecare ramură reprezintă rezultatul testului. Modelul Ridor diferă de arbori de decizie standard prin faptul că folosește un set de reguli pentru a determina când să se oprească partiționarea datelor în subgrupuri suplimentare. Aceste reguli includ un număr minim de instanțe pe frunză și un număr maxim de reguli care trebuie utilizate (Jijo și Abdulazeez 2021, 20-28).

4.14 MLP – Algoritmul Multi-Layer Perceptron (MLP) este un tip de rețea neuronală feed-forward care este utilizat, în mod obișnuit, în sarcinile de învățare supravegheată, cum ar fi clasificarea și regresia, și a fost propus de Rosenblatt în 1950 (Ozbay și Alatas 2020). Este format din mai multe straturi de noduri sau neuroni, fiecare neuron, dintr-un strat conectat la toți neuronii din stratul anterior. Stratul de intrare primește datele de intrare, iar stratul de ieșire produce rezultatul final sau predicția. Straturile ascunse dintre straturile de intrare și de ieșire efectuează transformări neliniare ale datelor de intrare pentru a extrage caracteristici semnificative (Botalb și alții 2018, 1-18).

4.15 OLM – Modelul de învățare ordinal (OLM) este un tip de algoritm de învățare supravegheată, utilizat pentru problemele de regresie ordinală, propuse de Ben-David și colab. (Ben-David, Sterling și Pao 1989, 45-49). Într-o problemă de regresie ordinală, variabila țintă are o ordonare naturală, cum ar fi o evaluare de la 1 la 5, mai degrabă decât să fie nominală sau binară.

4.16 SimpleCart – Simple CART (Arbori de clasificare și regresie) a fost propus pentru prima dată de Leo Breiman în 1984 (Breiman, Friedman și alții 2017) și este un algoritm de arbore de decizie care partiționează recursiv datele în subseturi, pe baza valorilor caracteristicilor de intrare, pentru a minimiza impuritatea variabilei țintă (Loh 2011, 14-23).

4.17 ASC – Clasificatorul cu atribute selectate (ASC) este un algoritm de învățare supravegheat care combină selecția caracteristicilor cu un algoritm de clasificare pentru a îmbunătăți acuratețea acesteia și pentru a reduce complexitatea de calcul al modelului (Gnanambal și alții 2018, 3640-3644). ASC funcționează prin selectarea, mai întâi, a unui subset al celor mai relevante caracteristici din datele de intrare, folosind o metodă de selecție a caracteristicilor, cum ar fi câștigul de informații, raportul de câștig sau testul chi-pătrat.

4.18 J48 – Acest algoritm este, în mod regulat, modelul preferat pentru aplicațiile de clasificare. J48 este un algoritm de arbore de decizie și o implementare a algoritmului C4.5 (Bhargava și alții 2013). Funcționează prin partiționarea recursivă a datelor în subseturi, pe baza valorilor caracteristicilor de intrare, pentru a minimiza entropia sau câștigul de informații al variabilei țintă.

4.19 SMO – Sequential Minimal Optimization (SMO) este un algoritm utilizat, în principal, pentru a consolida antrenamentul mașinilor de suport vectorial (SVM), pentru sarcini de clasificare binară (Ozbay și Alatas 2020), și a fost introdus, inițial, în 1998 de către Platt (Platt 1998).

4.20 Bagging – este prescurtarea de la Bootstrap Aggregating și este o tehnică de învățare de ansamblu pentru îmbunătățirea stabilității și acurateței modelelor de învățare automată. Funcționează prin antrenarea mai multor instanțe ale aceluiași algoritm pe diferite subseturi de date de antrenament și apoi combinând predicțiile acestora printr-un mecanism de vot sau de mediere (Breiman 1996, 123-140).

4.21 Arborele de decizie – este un tip de algoritm de învățare supravegheată, utilizat atât pentru sarcinile de clasificare, cât și pentru cele de regresie. Este un model neparametric care împarte recursiv datele în subseturi, pe baza valorilor caracteristicilor de intrare, pentru a prezice valoarea variabilei țintă (Jijo și Abdulazeez 2021, 20-28).

4.22 IBk – Algoritmul „IBK" (Instance-Based K-Nearest Neighbor) este un algoritm de învățare automată, utilizat pentru sarcini de clasificare și regresie, care aparține familiei de algoritmi de învățare leneșă, unde modelul este antrenat prin stocarea întregului set de date de antrenament și prin realizarea de predicții, pe baza asemănării dintre noile date de intrare și instanțele de antrenament stocate (Moayedi și alții 2019).

4.23 KLR – Kernel Logistic Regression (KLR) este un algoritm de învățare supravegheat, utilizat pentru sarcini de clasificare. Este o extensie a algoritmului tradițional de regresie logistică, ce utilizează o funcție de nucleu pentru a transforma datele de intrare într-un spațiu dimensional mai înalt, permițând modelarea relațiilor neliniare dintre caracteristici (Zhu și Hastie 2005, 185-205).

4.24 Compararea performanței: Tabelul 1 arată performanța diferitelor modele, folosind setul de date ISOT Fake News. Algoritmul arborelui de decizie a funcționat mai bine decât toți ceilalți algoritmi, conform tuturor metricilor de evaluare, cu excepția regresiei. Algoritmul JRip a fost al doilea, ca performanță, la acurateță și precizie. Performanța algoritmilor IA supravegheați descriși a fost comparată de către o echipă de cercetare de la Departamentul de Inginerie software al Universității Firat din Elazig, Turcia (Ozbay și Alatas 2020).

TABELUL 1 Performanța pretinsă a algoritmilor IA supravegheați descriși în secțiunea 4, instruiți și evaluați, folosind setul de date ISOT Fake News, conform F.A. Ozbay și B. Alatas (Ozbay și Alatas 2020). Cele mai mari scoruri sunt subliniate.

Model	Acuratețe	Precizie	Regresie	Scorul F1
BayesNet	0,586	0,587	0,586	0,586
JRip	0,607	0,611	0,588	0,599
OneR	0,559	0,567	0,560	0,547
Decision Stump	0,564	0,574	0,564	0,549
ZeroR	0,501	0,501	**1.000**	0,667
SGD	0,589	0,590	0,583	0,586
CVPS	0,501	0,501	**1.000**	0,667
RFC	0,526	0,525	0,534	0,530
LMT	0,607	0,604	0,616	0,610
LWL	0,570	0,573	0,570	0,566
CvC	0,553	0,556	0,526	0,541
WIHW	0,501	0,501	**1.000**	0,667
Ridor	0,557	0,563	0,558	0,549
MLP	0,565	0,565	0,571	0,568
OLM	0,516	0,540	0,516	0,430
SimpleCart	0,604	0,607	0,586	0,597
ASC	0,588	0,598	0,534	0,564
J48	0,558	0,558	0,563	0,560
SMO	0,534	0,536	0,489	0,512
Bagging	0,598	0,603	0,576	0,589
Arbore de decizie	**0,968**	**0,963**	0,973	**0,968**
IBk	0,551	0,551	0,551	0,550
KLR	0,606	0,605	0,614	0,609

După cum se evidențiază în tabel, algoritmul bazat pe Arbori de Decizie oferă o performanță deosebită față de celelalte modele clasificate, atunci când se confruntă cu sarcini de clasificare binară de știri false. Rezultatul este confirmat și de alți autori, care au obținut un scor de precizie de peste 95% (Lyu și Lo 2020). Ideea de bază din spatele unui arbore de decizie este de a crea un model asemănător unui arbore care să reprezinte un set de decizii și posibilele consecințe ale acestora. Motivul pentru care modelul arborelui de decizie a funcționat mai bine decât celelalte modele supravegheate pentru identificarea de știri false poate fi că algoritmul este capabil să gestioneze în mod eficient datele cu dimensiuni mari și rare. Arborii de decizie pot gestiona datele categorice într-un mod precis, ceea ce este obișnuit în sarcinile NLP, unde cuvintele sau expresiile sunt adesea folosite drept caracteristici. În cazul datelor text în care cuvintele pot avea conotații diferite în propoziții diferite, algoritmii arborelui de decizie pot să nu fie cea mai bună alegere. Acest lucru se datorează faptului că arborii de decizie nu țin cont de contextul în care apar cuvintele și nu au un mecanism încorporat pentru manipularea cuvintelor cu sensuri multiple (Jijo și Abdulazeez 2021, 20-28).

5. Modele de învățare profundă bazate pe transformatori, utilizate în detecția știrilor false

Învățarea profundă (Deep Learning) este o altă metodă populară IA pentru clasificarea știrilor false. Modelele de învățare profundă și cele bazate pe transformatori pot contribui la detectarea știrilor false, permițând mașinilor să învețe modele și caracteristici complexe din cantități mari de date text (Young și alții 2018,

55-75). Aceste modele pot gestiona complexitatea și variabilitatea limbajului și pot identifica indiciile și modelele lingvistice subtile care disting știrile false de știrile reale (Chollet 2017). Principala diferență dintre învățarea supravegheată și învățarea profundă este tipul de modele utilizate. Algoritmii de învățare supravegheată pot fi aplicați la o gamă largă de probleme și tipuri de date, dar pot avea probleme cu date foarte complexe. Modelele de învățare profundă sunt concepute pentru a gestiona date complexe, cum ar fi imagini sau text, dar necesită cantități mari de date de antrenament și resurse de calcul.

5.1 XLNet – numit model eXtreme Learning NETwork, este un model NLP de actualitate, pre-antrenat, care a fost introdus în 2019 și despre care se pretinde că a atins valori mari în sarcini binare, care implică seturi de date echilibrate de știri false (Gautam, Venktesh și Masud 2021). Modelul se bazează pe arhitectura Transformer, care este un tip de rețea neuronală, potrivită pentru sarcinile de procesare a limbajului natural (NLP). Ceea ce diferențiază XLNet de alte modele pre-antrenate este utilizarea unei metode autoregresive care permite modelarea contextului bidirecțională, ceea ce ajută modelul să înțeleagă mai bine contextul și relațiile dintre cuvinte într-o propoziție. Această abordare permite XLNet să obțină reale performanțe pentru o gamă largă de sarcini în limbaj natural, inclusiv modelarea limbajului, răspunsul la întrebări și analiza sentimentelor (Gundapu și Mamidi 2021). Să luăm, de exemplu, un text precum „Președintele a făcut un anunț că noua politică ar fi în beneficiul tuturor americanilor, dar experții au criticat planul ca fiind dăunător economiei". Acesta conține mai multe indicii lingvistice, care sunt asociate cu știrile false, inclusiv utilizarea unui limbaj pozitiv („beneficiul tuturor americanilor"), urmat de un limbaj negativ („a criticat planul ca fiind dăunător"). Un model autoregresiv, precum XLNet, surprinde aceste modele subtile, luând în considerare contextul bidirecțional al fiecărui cuvânt din propoziție, permițându-i să identifice relațiile dintre cuvinte și expresii care indică știrile false.

5.2 BERT și ALBERT – Alte modele axate pe transformatori s-au bazat pe modelul BERT (Bidirectional Encoder Representations from Transformers) de la Google, (Devlin și alții 2019) care este un model de deep learning preantrenat, care se pretinde că a atins performanțe pentru o gamă largă de sarcini de procesare a limbajului natural, inclusiv răspunsuri la întrebări, analiza sentimentelor și traducerea limbajului. BERT este un model bazat pe transformatori care este antrenat pe un corp mare de date text, permițându-i să învețe reprezentări bogate ale limbajului, care pot fi reglate fin pentru sarcini specifice. BERT sau variante, precum ALBERT (A Lite BERT model) (Gundapu și Mamidi 2021), sunt considerate a fi extrem de eficiente în sarcini, precum înțelegerea limbajului natural și clasificarea textului, sarcini similare cu clasificarea binară a știrilor false.

5.3 RoBERTa – Modelul Robustly Optimized BERT Approach (RoBERTa) a fost introdus în 2019 (Liu și alții 2019) și se bazează pe arhitectura BERT, dar a fost antrenat pe un corpus de date mult mai mare decât BERT, cu o durată de

antrenament extinsă și cu tehnici de antrenament îmbunătățite. Acest lucru îi permite lui RoBERTa să surprindă mai bine relațiile și modelele complexe în textul în limbaj natural, rezultând o performanță îmbunătățită la o gamă largă de sarcini NLP, inclusiv clasificarea știrilor false. RoBERTa este reglată fin pentru clasificarea entităților și s-a susținut că are valori superioare, atunci când este aplicată pe seturi de date de știri false și reale (Liu și alții 2019).

5.4 FakeBERT – Unul dintre modelele anterioare, bazate pe BERT, și reglat fin pentru sarcinile de detectare a știrilor false a fost numit FakeBERT (Kaliyar, Goswami și Narang 2021c) și folosește o tehnică de creștere a datelor, numită back-translation. Aceasta implică traducerea articolelor de știri reale într-o altă limbă și apoi traducerea lor înapoi în limba originală, folosind un sistem de traducere automată. Acest proces ajută la generarea de date suplimentare de antrenament și la creșterea diversității acestora, ceea ce poate îmbunătăți acuratețea modelului și capacitatea de a detecta variații subtile în text. Traducerea inversă poate fi utilă pentru detectarea știrilor false prin generarea de date sintetice pentru modelele de antrenament. Aceste date sintetice pot fi folosite pentru a mări seturile de date reale ale articolelor de știri etichetate, ajutând la îmbunătățirea performanței modelelor NLP, antrenate pentru detectarea știrilor false.

5.5 DeepFakE și EchoFakeD – Alți autori au folosit un model Deep Neural Network (DNN) și l-au ajustat pentru sarcinile de clasificare a știrilor false. Modele, precum DeepFakE (Kaliyar, Goswami și Narang 2021a, 1015-1037) sau EchoFakeD (Kaliyar, Goswami și Narang 2021b, 8597–8613) care au fost instruite pe seturi de date de știri false, precum BuzzFeed și PolitiFact, sunt creditate cu scoruri de precizie între 88% și 98%. Un DNN este un tip de rețea neuronală artificială care este compusă din mai multe straturi de noduri interconectate sau neuroni. Aceste straturi permit rețelei să extragă și să învețe caracteristici din ce în ce mai complexe din datele de intrare, permițându-i să facă predicții sau clasificări mai precise. Modelele DNN constau dintr-un strat de intrare, unul sau mai multe straturi ascunse și un strat de ieșire. Straturile ascunse conțin majoritatea neuronilor și sunt responsabile de procesarea și transformarea datelor de intrare. Fiecare neuron dintr-un model DNN primește input de la mai mulți neuroni din stratul anterior și utilizează o funcție de activare pentru a transforma intrarea, înainte de a o transmite la stratul următor (Kaliyar, Goswami și Narang 2021c, 11765–11788). Prin antrenamentul pe un set de date specific unui domeniu (cum ar fi știrile false), DNN-urile pot învăța să identifice modele și caracteristici care sunt specifice domeniului sau limbii respective, îmbunătățindu-și acuratețea și eficacitatea pentru detectarea diferitelor clase.

5.6 LSTM-RRN și BiLSTM-RNN – Unii cercetători (Bahadad, Saxena și Kamal 2019) au experimentat cu arhitecturi bazate pe rețele neuronale recurente cu memorie pe termen scurt (LSTM-RRN) sau pe rețele neuronale recurente cu memorie bidirecțională și pe termen scurt (BiLSTM-RNN), cu un anumit grad de succes după reglarea fină, în ciuda faptului că un model LSTM-RNN este un tip de arhitectură

de rețea neuronală profundă, care este concepută pentru a procesa date secvențiale, cum ar fi datele din seria temporală sau textul în limbaj natural. Stratul LSTM permite rețelei să rețină informațiile din intrările anterioare pe o perioadă lungă de timp, făcându-l potrivit pentru sarcini precum detectarea știrilor false care necesită înțelegerea contextului sau a istoricului datelor. Anterior, LSTM-RNN s-a susținut că sunt eficiente pentru sarcini precum modelarea limbii, analiza sentimentelor și traducerea automată, dar au fost folosite doar recent chiar și pentru clasificarea știrilor false (Bahadad, Saxena și Kamal 2019, 74-82).

5.7 CNN – Modelele utilizate includ CNN-ul clasic (Luan și Lin 2019) sau rețelele neuronale convolute, care reprezintă un anumit tip de arhitectură de rețea neuronală profundă, folosită în mod obișnuit pentru sarcini de recunoaștere a imaginilor și video. Inovația cheie a CNN este utilizarea straturilor convolute, care aplică un set de filtre imaginii de intrare, permițând rețelei să învețe caracteristici și modele importante la diferite scări spațiale. Pe lângă straturile convolute, un CNN tipic include și straturi de grupare, care reduc dimensiunea spațială a hărților de caracteristici și straturi complet conectate, care realizează clasificarea finală. CNN-urile sunt capabile să identifice modele și caracteristici în datele text prin convoluția unui set de filtre peste secvența de text de intrare. Acest proces permite rețelei să capteze dependențele locale dintre cuvintele și expresiile adiacente din text, ceea ce este important pentru detectarea indiciilor lingvistice subtile care disting articolele de știri false de articolele de știri reale (Luan și Lin 2019).
Performanța pretinsă a modelelor prezentate mai sus a fost comparată, folosind aceleași metrici ca și tehnicile de învățare supravegheată, cu rezultatele prezentate în Tabelul 2.

Model	Acuratețe	Precizie	Regresie	Scorul F1
ROBERTa	**0,9996**	**0,9997**	**0,9994**	**0,9996**
LSTM-RRN	0,9697	0,97	0,97	0,97
BiLSTM - RRN	0,9875	0,97	0,97	0,97
ALBERT	0,9780	0,9781	0,9781	0,9780
FakeBERT	0,9874	0,99	0,99	0,99
DeepFakE	0,8864	0,8210	0,8460	0,8404
EchoFakeD	0,9230	0,9047	0,8636	0,8837
BERT	0,9813	0,9813	0,9813	0,9813
XLNet	0,9785	0,9787	0,9789	0,9785
CNN	0,9698	0,9698	0,9698	0,9698

TABELUL 2 Performanța revendicată a modelelor de deep learning și bazate pe transformatori pentru sarcinile automate de detectare a știrilor false pe seturile de date de știri false ISOT, BuzzFeed și PolitiFact. Cele mai mari scoruri sunt subliniate.

5.8 Discuție: Conform metricilor selectați, modelul RoBERTa are cele mai bune rezultate dintre toate arhitecturile comparate, cu mențiunea că, pentru astfel de sarcini de clasificare binară, modelele de învățare automată par să atingă, per ansamblu, metrici mai ridicați. Modelul RoBERTa obține o precizie ridicată în detectarea știrilor false prin valorificarea capacităților sale puternice de reprezentare a limbajului și a capacității de a capta eficient relațiile semantice dintre cuvinte și expresii. De exemplu, dacă luăm în considerare următorul titlu: „Oamenii de știință descoperă un nou tratament pentru cancer care funcționează în 100% din cazuri".

Un cititor uman poate fi imediat sceptic cu privire la acest titlu, deoarece pare prea frumos pentru a fi adevărat. Cu toate acestea, un model de învățare automată care este antrenat pe pachete de cuvinte sau reprezentări simple de încorporare a cuvintelor ar putea să nu poată surprinde nuanțele limbajului și poate clasifica incorect acest articol ca fiind real. În schimb, modelul RoBERTa este capabil să analizeze întregul context al titlului și să identifice indiciile subtile care sugerează că articolul este fals, cum ar fi utilizarea unui limbaj hiperbolic și lipsa dovezilor științifice care să susțină afirmația.

6. Tehnici de învățare nesupravegheate, utilizate în detectarea știrilor false

Învățarea nesupravegheată este o metodă IA care poate fi utilizată pentru clasificarea știrilor false atunci când datele etichetate (știri false sau adevărate, notate ca atare de un operator uman) nu sunt disponibile. Tehnicile de învățare nesupravegheate pentru detectarea știrilor false nu necesită date etichetate pentru a antrena modelul, ci se bazează, în schimb, pe identificarea tiparelor și a relațiilor din date pentru a clasifica noile instanțe ca știri reale sau false (Gangireddy și alții 2020, 75-83). O abordare comună nesupravegheată este gruparea, în care articolele de știri similare sunt grupate pe baza conținutului și modelelor lingvistice. Acest lucru poate ajuta la identificarea grupurilor de articole de știri care sunt similare ca stil și conținut, ceea ce poate ajuta la distingerea știrilor reale de cele false (Celebi și Aydin 2018, 164-170). O altă abordare nesupravegheată este modelarea subiectelor, care identifică subiecte și teme într-un corpus de text (Li și alții 2021). Modelarea subiectelor poate ajuta la identificarea subiectelor comune în articolele de știri false, cum ar fi teoriile conspirației, titlurile clickbait și senzaționalul. Detectarea anomaliilor este o altă tehnică nesupravegheată, în care modelul învață să identifice cazurile care se abat semnificativ de la normă (Celebi și Aydin 2018, 23-28). Acest lucru poate fi util în detectarea articolelor de știri false care conțin modele de limbaj sau sintaxă neobișnuită.
Astfel de metode s-au atins atunci când au fost instruite și testate pe setul de date PolitiFact (PolitiFact 2017) scoruri de acuratețe între 0,81 și 0,82 (Gangireddy și alții 2020).

6.1 Discuție: Tehnicile nesupravegheate se pot dovedi cruciale în identificarea campaniilor de dezinformare în curs de desfășurare pe rețelele sociale sau pe platforme online similare și pot fi combinate cu tehnici supravegheate pentru a îmbunătăți performanța. De exemplu, un model antrenat pe un set de date mic etichetat poate fi utilizat pentru a identifica acele cazuri de știri false, care pot fi apoi folosite pentru a antrena un model mai mare nesupravegheat pentru a îmbunătăți detectarea la scară mai mare. Tehnicile de învățare nesupravegheată pot fi eficiente în detectarea știrilor false, în special atunci când sunt utilizate în combinație cu tehnici supravegheate și cu analize umane pentru a verifica rezultatele (Celebi și Aydin 2018).

7. Sisteme bazate pe reguli pentru detectarea știrilor false

Sistemele bazate pe reguli sunt o altă metodă IA care poate fi folosită pentru clasificarea știrilor false. Sistemele bazate pe reguli implică definirea unui set de reguli sau euristici, care pot fi utilizate pentru a identifica articole de știri false pe baza unor caracteristici specifice, cum ar fi utilizarea unui limbaj încărcat emoțional sau prezența erorilor logice. Sistemele bazate pe reguli pot fi mai puțin precise decât metodele de învățare supervizată sau profundă, dar pot fi eficiente, atunci când sarcina este relativ simplă și datele etichetate nu sunt disponibile (Yuliani și alții 2019).

Aceste metode IA sunt utile pentru clasificarea știrilor false, deoarece sunt eficiente în identificarea tiparelor și structurilor din date. De exemplu, un model bazat pe reguli, folosit pentru a detecta propagarea știrilor false arabe în timpul Covid-19 a obținut o acuratețe de 79,7% (Alotaibi și Alhammad 2022), care a fost antrenat pe un set de date de 5.015.111 de tweeturi și este un succes destul de mare, ținând cont de limitările care decurg din dificultatea procesării limbii arabe.

7.1 Discuție: Sistemele bazate pe reguli par să fie mai puțin eficiente decât alte tehnici în identificarea unor forme mai nuanțate sau complexe de știri false care nu se încadrează perfect în categorii predefinite și ar trebui utilizate în combinație cu modele care folosesc date etichetate pentru a obține o mai mare acuratețe într-un model de management al dezinformării.

8. Modele hibride utilizate pentru detectarea știrilor false

Modelele hibride sunt o combinație de două sau mai multe metode IA. De exemplu, un sistem bazat pe reguli poate fi combinat cu o metodă de învățare supravegheată sau nesupravegheată pentru a îmbunătăți performanța sarcinii de clasificare. Modelele hibride pot fi mai precise și mai eficiente decât modelele cu o singură metodă, deoarece pot valorifica punctele forte ale mai multor metode.

8.1 CNN-RNN hibrid – Un model hibrid CNN-RNN a fost propus de Nasir et al. (Nasir, Khan și Varlamis 2021), cu succes limitat (precizie de 0,5). O astfel de arhitectură ar trebui să combine punctele forte ale CNN-urilor și ale RNN-urilor pentru a capta atât contextul local, cât și global al datelor de intrare, modelând în același timp dependențele temporale și contextul secvenței de intrare. Componenta CNN extrage caracteristici de nivel înalt din datele de intrare, în timp ce componenta RNN modelează dependențele temporale ale secvenței. Starea ascunsă finală a componentei RNN este apoi utilizată pentru clasificare (Nasir, Khan și Varlamis 2021).

8.2 CSI – Un alt model hibrid, numit CaptureScoreIntegrate (CSI) (Ruchansky, Seo și Liu 2017), care a folosit seturi de date, colectate de pe Twitter și Weibo, a obținut un succes promițător. Modelul este compus din 3 părți: Capture (include capturarea

conținutului și caracteristicilor articolului de știri, folosind un RRN), Score (care calculează scorul de credibilitate pentru sursa articolului) și Integrate (care clasifică rezultatele).

8.3 SVM-RNN-BI-GT – Un alt studiu a propus un model hibrid în care SVM și RNN cu GRU-uri bidirecționale sunt încorporate în valorificarea conținutului de știri și a comentariilor utilizatorilor în știrile false (Albahar 2021) pe un set de date PolitiFact (PolitiFact 2017).

8.4 HAN – Un alt model hibrid, propus pentru detectarea știrilor false, este HAN (Hierarchical Attention Network) (Albahar 2021). Acesta are o structură ierarhică proprie care reflectă structura ierarhică a știrilor prezentate în setul de date și are două niveluri de mecanisme de atenție, aplicate la cuvânt și propoziție-nivel, permițându-i să se ocupe diferențial de conținut mai mult și mai puțin important, atunci când construiește reprezentarea textului fals/adevărat (Yang și alții 2016, 1480-1489).

8.5 HSA-BLSTM – Hierarchical Social Attention – Memoria bidirecțională pe termen scurt a fost testată pe seturi de date, colectate de pe Twitter și Weibo (Albahar 2021), după ce a fost utilizată, inițial, pentru a detecta zvonuri prin valorificarea reprezentărilor ierarhice la diferite niveluri și contexte sociale (Guo și alții 2018, 943-951).

8.6 TCNN-URG – Rețeaua neuronală convoluțională de transfer – Generatorul de răspuns al utilizatorului (TCNN-URG) este, de obicei, folosit pentru a îmbunătăți calitatea răspunsurilor generate de chatboturile din rețelele sociale sau de asistenții virtuali (Qian și alții 2018, 3834-3840). Este compus dintr-un CNN și un autoencoder variațional condiționat și a fost, de asemenea, testat pentru detectarea automată a textului de știri false pe un set de date de știri false PolitiFact (Albahar 2021).

8.7 Rezultatele pentru modelele hibride, revendicate de echipele de cercetare menționate mai sus, sunt prezentate în Tabelul 3, folosind același sistem metric ca în tabelele 1 și 2, pentru a oferi o comparație între tipuri similare de abordări ale detecției de știri false într-o clasificare binară.

TABELUL 3 Performanța hibridă pentru sarcinile automate de detectare a știrilor false pe seturile de date de știri false ISOT, PolitiFact, Twitter și Weibo. Cele mai mari scoruri sunt subliniate.

Model	Acuratețe	Precizie	Regresie	Scorul F1
CNN-RNN hibrid	0,5	0,5	0,5	0,5
CSI-Twitter	0,892	0,9	0,8	0,894
CSI-Weibo	**0,953**	**0,953**	0,953	**0,954**
SVM-RNN-BI-GT	0,912	0,910	**0,961**	0,932
Han	0,837	0,824	0,941	0,810
TCNN–URG	0,712	0,711	0,860	0,860
HSA-BLSTM	0,846	0,894	0,868	0,881

8.8 Discuție: După cum se observă în Tabelul 3, în funcție de acuratețe, precizie și Scorul F1, cel mai performant model este modelul CSI, antrenat și evaluat pe un set de date Weibo (un site de microblogging chinezesc, similar cu Twitter) (Ruchansky, Seo și Liu 2017). Modelul CSI este separat în 3 părți, ceea ce permite CSI să producă o predicție pentru utilizatori și articole în mod independent, combinând informațiile pentru clasificare. Experimentele au fost efectuate de echipa de cercetare pe două seturi de date, obținute în lumea reală (Weibo și Twitter), care au demonstrat acuratețea modelului CSI în clasificarea articolelor de știri false.

Deoarece detectarea știrilor false este o sarcină complexă, care necesită analiza diferitelor caracteristici, cum ar fi informații lingvistice, temporale și legate de utilizator, modelele hibride pot depăși limitările unei singure abordări și pot obține rezultate mai bune, chiar dacă la data curentă sunt în urma altor modele în valorile lor. Pentru a aplica o astfel de cercetare într-un sistem de operare real, utilizat zilnic, ar trebui să se țină cont de faptul că modelele de învățare profundă, cum ar fi BERT sau CNN, pot surprinde tipare complexe în text, dar este posibil să nu ia în considerare caracteristicile temporale sau legate de utilizator, care sunt importante în detectarea știrilor false (Kaur, Boparai și Singh 2019, 2388-2392). Sistemele similare, bazate pe reguli, pot folosi cunoștințele de domeniu pentru a detecta modele suspecte în știri, dar s-ar putea să nu se generalizeze bine la exemple necunoscute, făcându-le astfel mai puțin precise decât abordările de învățare supravegheată, care au fost instruite anterior pe o mare varietate de exemple. Deoarece detectarea știrilor false este o problemă în continuă evoluție, necesitând noi tehnici sau modele pentru a detecta tipurile emergente de știri false, viitoarele modele hibride pot fi flexibile și adaptabile, permițând integrarea de noi modele sau caracteristici, pe măsură ce acestea devin disponibile, conducând la detectarea mai precisă și robustă a știrilor false (Thaher și alții 2021).

9. Concluzie

După cum se arată în articol, diferite metode de învățare automată IA returnează o gamă largă de valori, atunci când se confruntă cu o sarcină de clasificare binară care implică știri reale și false. Acest lucru arată că există încă suficient loc de îmbunătățire în acest domeniu. După compararea valorilor prezentate în tabelele 1, 2 și 3 și luând în considerare scorurile de acuratețe ale modelelor nesupravegheate și bazate pe reguli, modelul ML ajustat „RoBERTa" pretinde că atinge cele mai bune valori pentru setul de date de știri false ISOT (99,96% scor de acuratețe, 99,97% precizie, 99,94% regresie și 99,96% F-score). Modelul a fost dezvoltat de Facebook AI, se bazează pe modelul BERT, dar este antrenat pe un corp mai mare de date text și include tehnici suplimentare de pre-instruire pentru a-și îmbunătăți acuratețea. Performanța lui RoBERTa în ceea ce privește știrile false o face un candidat puternic pentru includerea într-un sistem integrat de management al dezinformării. În plus, RoBERTa poate fi ajustată pentru domenii specifice, cum

ar fi politica sau sănătatea, ceea ce este important în managementul dezinformării, unde subiectul poate fi specializat.

Trebuie să subliniem că RoBERTa este unul dintre modelele care are o comunitate mare și activă de utilizatori și dezvoltatori, ceea ce înseamnă că este bine susținută și actualizată frecvent cu noi funcții și îmbunătățiri. Acest lucru face mai ușor să se integreze într-un sistem mai mare și să rămână la curent cu cele mai recente cercetări în domeniul NLP. Deoarece greutățile preantrenate și modelele asociate ale RoBERTa sunt disponibile gratuit, este accesibilă unei game mai largi de utilizatori, indiferent de resursele sau de expertiza lor tehnică, oferindu-i o combinație de performanță ridicată, flexibilitate și accesibilitate pentru a fi inclusă într-un sistem software complementar.

Referințe

Alam, M.T., S. Ubaid, S.S. Sohail, M. Nadeem, S. Hussain și J. Siddiqui. 2021. "Comparative Analysis of Machine Learning based filtering techniques using MovieLens." *Procedia Computer Science 194* 2010-2017.

Albahar, Marwan. 2021. "A hybrid model for fake news detection: Leveraging news content and user comments in fake news." doi:https://doi.org/10.1049/ise2.12021.

Alotaibi, Fatimah L. și Muna M. Alhammad. 2022. "Using a Rule-based Model to Detect Arabic Fake News Propagation during Covid-19." *International Journal of Advanced Computer Science and Applications.* doi:10.14569/IJACSA.2022.0130114.

Bahadad, Pritika, Preeti Saxena și Raj Kamal. 2019. *Procedia Computer Science 165* 65: 74–82. doi:https://doi.org/10.1016/j.procs.2020.01.072.

Ben-David, A., L. Sterling și Y.H. Pao. 1989. "Learning and classification of monotonic ordinal concepts." *Computational Intelligence* 5 (1): 45-49. doi:https://doi.org/10.1111/j.1467-8640.1989.tb00314.x.

Bergsma, S., M. Dredze, B. Van Durme, T. Wilson și D. Yarowsky. 2013. "Broadly improving user classification via communication-based name and location clustering on twitter." *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: human language technologies.*

Bhargava, N., G. Sharma, R. Bhargava și M. Mathuria. 2013. "Decision tree analysis on j48 algorithm for data mining." *Proceedings of international journal of advanced research in computer science and software engineering.*

Botalb, A., M. Moinuddin, U.M. Al-Saggaf și S.S. Ali. 2018. "Contrasting convolutional neural network (CNN) with multi-layer perceptron (MLP) for big data analysis." *International conference on intelligent and advanced system (ICIAS).*

Breiman, L. 1996. "Bagging predictors." *Machine Learning* 24 (2): 123-140. doi:10.1023/A:1018054314350.

Breiman, L., J.H. Friedman, R. A. Olshen și C. Stone. 2017. *Classification and regression trees.* New York: Routledge. doi:https://doi.org/10.1201/9781315139470.

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

Chantar, H., M. Mafarja, H. Alsawalqah, A.A. Heidari, I. Aljarah și H. Faris. 2020. "Feature selection using binary grey wolf optimizer with elite-based crossover for Arabic text classification." *Neural Comput. Appl 32* 12201–12220. doi:https://doi.org/10.1007/s00521-019-04368-6.

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.* http://dx.doi.org/ 10.1016/j.catena.2016.11.032.

Chollet, François. 2017. *Deep Learning with Python.* New York: Manning.

Cohen, William W. 1995. "Fast effective rule induction." *Machine learning proceedings, 12th anual conference.* Morgan Kaufmann. 115-123.

Devasena, C.L., T. Sumathi, V.V. Gomathi și M.Hemalatha. 2011. "Effectiveness evaluation of rule based classifiers for the classification of iris data set." *Bonfring International Journal of Man Machine Interface 1.*

Devlin, J., M.W. Chang, K. Lee și K. Toutanova. 2019. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." doi:https://doi.org/10.48550/arXiv.1810.04805.

Gangireddy, Siva Charan Reddy, P. Deepak, Cheng Long și Tanmoy Chakraborty. 2020. "Unsupervised Fake News Detection: A Graph-based Approach." *Proceedings of the 31st ACM Conference on Hypertext and Social Media (HT '20)* 75-83. doi:https://doi.org/10.1145/3372923.3404783.

Gautam, Akansha, V. Venktesh și Sarah Masud. 2021. "Fake news detection system using xlnet model with topic distributions: Constraint@ aaai2021 shared task." *Combating Online Hostile Posts in Regional Languages during Emergency Situation: First International Workshop, CONSTRAINT 2021, Collocated with AAAI 2021, Virtual Event.*

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* 3640-3644.

Gundapu, Sunil și Radhika Mamidi. 2021. "Transformer based Automatic COVID-19 Fake News Detection System." *International Institute of Information Technology.*

Guo, H., J. Cao, Y. Zhang, J. Guo și J. Li. 2018. "Rumor Detection with Hierarchical Social Attention Network." *Proceedings of the 27th ACM international conference on information and knowledge management.* doi:https://doi.org/10.1145/3269206.3271709.

Holte, Robert C. 1993. "Very simple classification rules perform well on most commonly used data sets." *Machine learning 11.* 63-90.

Jijo, B.T. și A.M. Abdulazeez. 2021. "Classification based on decision tree algorithm for machine learning." *Journal of Applied Science and Technology Trends (JASTT)* 20-28.

Kaliyar, Rohit Kumar, Anurag Goswami,și Pratik Narang. 2021a. "DeepFakE: improving fake news detection using tensor decomposition-based deep neural network." *Journal of Supercomputing* 77 (2): 1015-1037. doi:10.1007/s11227-020-03294-y.

—. 2021b. "EchoFakeD: improving fake news detection in social media." *Neural Computing and Applications* 33: 8597–8613. doi:https://doi.org/10.1007/s00521-020-05611-1(0123456789().,-volV)(0123456789(). ,- volV).

—. 2021c. "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.

Kaur, Prabhjot, Rajdavinder Singh Boparai și Dilbag Singh. 2019. "Hybrid Text Classification Method for Fake News Detection." *International Journal of Engineering and Advanced Technology (IJEAT)* 8 (5): 2388-2392.

Khosravi, Khabat, Zohreh Sheikh Khozani și Luca Mao. 2021. "A comparison between advanced hybrid machine learning algorithms and empirical equations applied to abutment scour depth prediction." *Journal of Hydrology.* doi:https://doi.org/10.1016/j.jhydrol.2021.126100.

Lakmali, K.B.N. și P.S. Haddela. 2017. "Effectiveness of rule-based classifiers in Sinhala text categorization." *National Information Technology Conference (NITC).* Colombo, Sri Lanka. doi:10.1109/NITC.2017.8285655.

Langley, Pat, Iba Wayne și Kevin Thompson. 1992. "An analysis of Bayesian classifiers." *Proceedings of the Tenth National Conference of Artificial Intelligence.* California. 223-228.

Li, Dun, Haimei Guo, Zhenfei Wang și Zhiyun Zheng. 2021. "Unsupervised Fake News Detection Based on Autoencoder." *Access.* doi:https://doi.org/10.1109/ACCESS.2021.3058809.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer și Veselin Stoyanov. 2019. "RoBERTa: A Robustly Optimized BERT Pretraining Approach." *ArXiv.* doi:https://doi.org/10.48550/arXiv.1907.11692.

Loh, Wei-Yin. 2011. "Classification and regression trees." *WIREs Data Mining Knowl Discov*. doi:10.1002/widm.8.

Luan, Yuandong 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.

Lyu, Shikun și Dan Chia-Tien Lo. 2020. "Fake News Detection by Decision Tree." *SoutheastCon.* doi:https://doi.org/10.1109/SoutheastCon44009.2020.9249688.

Moayedi, H., D. Tien Bui, B. Kalantar și L. Kok Foong. 2019. "Machine-Learning-Based Classification Approaches toward Recognizing Slope Stability Failure." *Applied Sciences* 9 (21). doi:https://doi.org/10.3390/app9214638.

Nasir, J.A., O.S. Khan și I. Varlamis. 2021. "Fake news detection: A hybrid CNN-RNN based deep learning approach." *International Journal of Information Management Data Insights.* doi:10.1016/j.jjimei.2020.100007.

Ozbay, Feyza Altunbey și Bilal Alatas. 2020. "Fake news detection within online social media using supervised artificial intelligence algorithms." doi:https://doi.org/10.1016/j.physa.2019.123174.

Platt, John. 1998. *Sequential minimal optimization: A fast algorithm for training support vector machines.* Technical Report MSR-TR-98-14, Microsoft Research. PolitiFact. 2017. https://www.politifact.com/.

Qian, F., C. Gong, K. Sharma și Y. Liu. 2018. ”Neural User Response Generator: Fake News Detection with Collective User Intelligence.” *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence.*

Ruchansky, Natali, Sungyong Seo și Yan Liu. 2017. ”CSI: A Hybrid Deep Model for Fake News Detection.” *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (CIKM ‘17).* doi:https://doi.org/10.1145/3132847.3132877.

Sammut, C., Webb, G.I. (eds). 2017. ”Decision Stump. Encyclopedia of Machine Learning.” În *Encyclopedia of Machine Learning*, de C., Webb, G.I. (eds) Sammut, 262–263. Boston, MA.: Springer. doi:10.1007/978-0-387-30164-8_202.

Thaher, T., M. Saheb, H. Turabieh și H. Chantar. 2021. ”Intelligent Detection of False Information in Arabic Tweets Utilizing Hybrid Harris Hawks Based Feature Selection and Machine Learning Models.” *Symmetry 13* 556. doi:https://doi.org/10.3390/sym13040556.

Tuyen, T.T., A. Jaafari, H.P.H. Yen, T. Nguyen-Thoi, T. Van Phong, H.D. Nguyen și B.T. Pham. 2021. ”Mapping forest fire susceptibility using spatially explicit ensemble models based on the locally weighted learning algorithm.” *Ecological Informatics.* doi:https://doi.org/10.1016/j.ecoinf.2021.101292.

University of Victoria. 2017. ”ISOT Fake News dataset.” https://www.uvic.ca/ecs/ece/isot/datasets/fake-news/index.php.

Varma, Sudhir și Richard Simon. 2006. ”Bias in error estimation when using cross-validation for model selection.” *BMC bioinformatics 7.1.*

Yang, Z., D. Yang, C. Dyer, X. He, A. Smola și E. Hovy. 2016. ”Hierarchical attention networks for document classification.” *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies.*

Young, T., D. Hazarika, S. Poria și E. Cambria. 2018. ”Recent trends in deep learning based natural language processing.” *ieee Computational intelligenCe magazine* 13 (3): 55-75. doi:10.1109/MCI.2018.2840738.

Yuliani, S.Y., M.F.B. Abdollah, S. Sahib și Y.S. Wijaya. 2019. ”A framework for hoax news detection and analyzer used rule-based methods.” *International Journal of Advanced Computer Science and Applications.*

Zhu, J. și T. Hastie. 2005. ”Kernel logistic regression and the import vector machine.” *Journal of Computational and Graphical Statistics* 14 (1): 185-205.