FACULTATEA DE AUTOMATICĂ ȘI CALCULATOARE

Universitatea Națională de Știință și Tehnologie POLITEHNICA București

Academia Oamenilor de Știință din România

Raport 1

Proiect:

NetGuardAI: Sistem inteligent pentru detectarea și stoparea conținutului dăunător pe rețelele sociale

- Iulie 2025 -

Director proiect: Ș.L. Dr. Abil. Ing. Ciprian-Octavian TRUICĂ
Membri: As. Drd. Ing. Alexandru PETRESCU
Drd. Ing. Anamaria VIȘAN

Platformele online sunt sisteme complexe care influențează mediul comercial, social și politic, dezbătându-se subiecte importante din viața reală, de exemplu, emigrarea, sănătate, alegeri, schimbări climatice, etc. Aceste medii online permit utilizatorilor să fie anonimi și să aibă libertatea de exprimare. Pe lângă avantajele lor evidente, unii utilizatori abuzează de această libertate pentru a răspândi conținut dăunător, de exemplu, dezinformare, propagandă, teorii conspiraționiste sau discursuri abuzive, agresive și ofensive. Controlul online împotriva actorilor rău intenționați, care au acest comportament antisocial, este o sarcină complexă. Tehnicile automate de detectare și imunizare pot reduce eficient influența negativă a comportamentului antisocial folosit de acești actori malițioși. Obiectivele principale ale proiectului NetGuardAI sunt:

1. Îmbunătățirea eficienței sistemelor de detectare a conținutului dăunător în mediile online prin algoritmi de inteligență artificială și învățare automata;
2. Protejarea utilizatorilor de conținutul dăunător prin strategii de imunizare.

1. Activități de cercetare

Activitățile de cercetare pentru primul raport de cercetare al proiectului NetGuardAI conform diagramei GANTT (Figura 1) sunt:

1. Analiza cerințelor și soluțiilor similare
2. Începerea colectării și procesării de date
3. Implementarea modulelor de detectare a conținutului dăunător
4. Evaluare modelelor de învățare automată
5. Diseminarea rezultatelor

WP	An	2025									2026									
	Luna	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
WP1	Management proiect																				
T1.1	Management administrativ și tehnic				D1.1				D1.2							D1.3					D1.2
WP2	Analiza cerințelor și soluții similare																				
T2.1	Analiza critică a soluțiilor de detecție a conținutului dăunător								D2.1	M2.1											
T2.2	Analiza critică a soluțiilor de imunizare								D2.2												
T2.3	Cerințe funcționale și non-funcționale								D2.3	M2.2											
T2.4	Cerințe tehnologice								D2.3												
WP3	Colectarea și procesarea datelor																				
T3.1	Analiza seturilor de date disponibile online								D3.1												
T3.2	Colectarea noi date și metadate									D3.2											
T3.3	Stocarea seturilor de date colectate										D3.3										
T3.4	Preprocesare și extragerea caracteristicilor											D3.4	M3.1								
WP4	Implementarea sistemul NetGuardAI																				
T4.1	Antrenarea modelelor pentru analiza sentimentelor													D4.1							
T4.2	Antrenarea modelelor pentru extragerea subiectelor														D4.2						
T4.3	Antrenarea modele pentru detectarea conținutului dăunător															D4.3					
T4.4	Dezvoltarea algoritmilor de imunizare																D4.4				
T4.5	Proiectarea aplicației web/mobilă																	D4.5	M4.1		
WP5	Evaluare, validare și optimizări																				
T5.1	Definirea scenariilor de test																D5.1				
T5.2	Testare experimentală a sistemului																	D5.2			
T5.3	Optimizări ale sistemului																		D5.3	M5.1	
WP6	Diseminare și exploatare																				

Figură 1. Diagrama GANTT

În conformitate cu calendarul activităților din propunerea de proiect, reprezentat grafic în diagrama GANTT (Figura 1), principalele activități de cercetare desfășurate până în această etapă a proiectului NetGuardAI sunt următoarele:

- **WP2 – Analiza cerințelor și soluții similare:** Analiza literaturii de specialitate pentru a identifica soluții similare de detectare a conținutului dăunător și strategii de imunizare a rețelelor sociale. Până în momentul de față, s-a analizat o parte din literatura de specialitate și s-au identificat soluții existente, s-a început analiza funcțională și non-funcțională împreună cu cea

tehnică. Activitățile pentru acest pachet de lucru se vor continua până în luna 8 de implementare a proiectului.

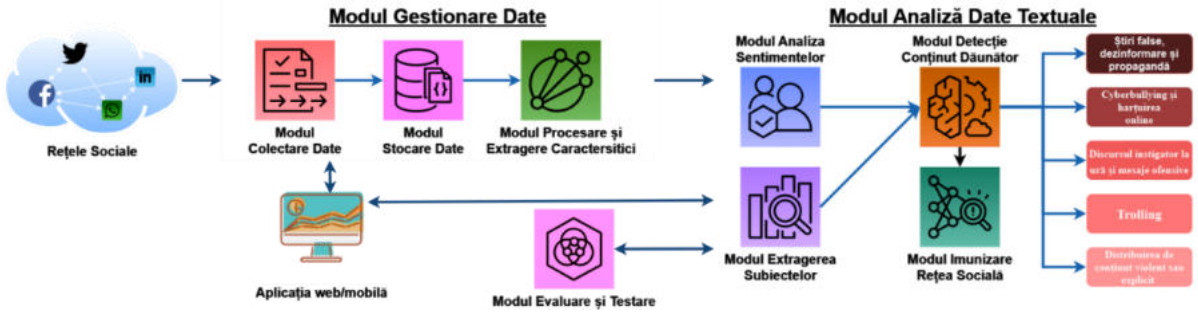
- **WP3 – Colectarea și procesarea datelor: a**naliza seturilor de date existente și colectarea unui nou set de date. În această perioadă s-au identificat și analizat mai multe seturi de date, în limba Română (FakeRom[1]), Engleză și Spaniolă (Exist 2025[2]). S-a început implementarea unui modul pentru procesarea datelor textuale. Activitățile pentru acest pachet de lucru se vor continua până în luna 8 de implementare a proiectului.

- **WP4 – Implementarea sistemului NetGuardAI:** implementarea modulelor de detecție a conținutului dăunător și implementarea strategiilor de imunizarea a rețelelor sociale. Am realizat dezvoltarea unor modelelor de analiză de sentimente și de detecție a conținutului dăunător bazate pe vectorizări de tip transformer (ex. BERT, RoBERTa, DeBERTa, etc.) si modele lingvistice mari (LLMs – Large Language Models) precum Llama2. Modelele dezvoltate i-au în considerare atât informația textuală (ex. mesajele postate de utilizator, sentimentele generate, etc.) cât și informații legate de context (ex. marcatori de discurs, perechi de categorii, polaritatea entităților, etc.)

- **WP5 – Evaluare, validare și optimizări:** Evaluarea modelelor de învățare automată și a strategiilor de imunizare a rețelelor sociale. Pentru modelele dezvoltate am implementat un modul pentru evaluare și validarea performanțelor, calculând metrici standard de evaluare a modelelor de inteligență artificială, ex. acuratețe, precizie, etc.

- **WP6 – Diseminare și exploatare**: Pregătirea și publicarea articolelor de jurnal și lucrărilor la conferințe internaționale. Crearea unui depozit accesibil online, sub licență deschisă, unde să se găsească codul sursă, modelele și setul de date, pentru a încuraja reproductibilitatea rezultatelor. Au fost publicate 4 articole (mai multe detalii în *Secțiunea 2. Activități de publicare*). Am creat un depozit pentru date și cod pe platforma Github[3].

Structura soluției finale este ilustrată în Figura 2, care prezintă arhitectura generală a NetGuardAI. Până la redactarea acestui raport, au fost realizate activități de proiectare, implementare și testare pentru componentele de detecție. În etapa următoare, intenționăm să demarăm dezvoltarea modulului destinat imunizării rețelelor sociale. În cele ce urmează, vom detalia aspectele legate de implementare și vom prezenta rezultatele obținute pentru modulele dezvoltate până în prezent. Menționăm că acest raport include exclusiv rezultatele deja publicate.



*Figură 2. Arhitectura generală **NetGuardAI***

[1] https://huggingface.co/datasets/mateiaass/FakeRom
[2] https://nlp.uned.es/exist2025/
[3] https://github.com/orgs/DS4AI-UPB/repositories

Analiza și detecția marcatorilor discursivi

Expresiile polilexicale pot transmite diferite tipuri de informații semantice și pragmatice, iar studiul acestora este esențial pentru generarea și procesarea limbajului. Printre aceste studii, există unele care vizează expresiile polilexicale ce funcționează ca marcatori discursivi [1]. Analiza marcatorilor discursivi joacă un rol crucial în înțelegerea structurii discursului, ceea ce o face relevantă pentru diverse domenii, inclusiv lingvistica și studiile computaționale, precum și pentru detectarea conținutului dăunător în mediile sociale. Cercetările din acest domeniu au condus la dezvoltarea mai multor abordări pentru identificarea, extragerea și clasificarea marcatorilor discursivi în seturi de date monolingve și multilingve. Aceste abordări se împart în două categorii principale: (i) bazate pe corpusuri și taxonomii funcționale și (ii) metode computaționale. Mai mult, în ultima vreme, eforturile s-au orientat către crearea unor instrumente interlingvistice, cum ar fi depozite de date interogabile de marcatori discursivi, pentru a facilita analiza multilingvă.

Marcatorii discursivi sunt un set de expresii lingvistice care reprezintă o parte inseparabilă a discursului și îndeplinesc funcții esențiale în înțelegerea discursului oral și scris. Marcatorii discursivi pot fi cuvinte simple sau expresii polilexicale formate din conjuncții, adverbe și locuțiuni prepoziționale. Aceștia indică o legătură între unități discursive, adică între enunțuri, secvențe mai lungi de text și chiar între text și contextul extralingvistic. Marcatorii discursivi îndeplinesc multiple funcții atât în monologuri, cât și în comunicarea interactivă, cum ar fi conversațiile și dialogurile. Rolurile lor includ, fără a se limita la acestea, stabilirea coerenței între propoziții și fraze, indicarea ezitării, facilitarea schimbului de replici, semnalarea schimbărilor de subiect, marcarea limitelor între intervenții, exprimarea rezervei, transmiterea atitudinii, gestionarea interacțiunii cu interlocutorii, solicitarea aprobării și indicarea tranzițiilor. Astfel, termenul de marcator discursiv este definit ca un element lingvistic care are ca funcție principală structurarea discursului, semnalarea relațiilor dintre enunțuri și ghidarea interpretării, mai degrabă decât contribuția la sensul propozițional. Acești marcatori contribuie la gestionarea coerenței, coeziunii și interacțiunii în comunicarea orală și scrisă.

Corpusul paralel pe care l-am colectat include date din 10 limbi, utilizând transcrieri publice ale discursurilor TED ca o extindere a corpusului paralel TED-EHL, găzduit în depozitul LINDAT/CLARIN-LT. Acest corpus multilingv constă în aliniamente lingvistice, având limba engleză ca limbă pivot, și cuprinde 1,3 milioane de propoziții. Selecția se bazează pe prezența expresiilor polilexicale (MWEs) care funcționează ca marcatori discursivi. Corpusul multilingv include enunțuri în engleză, lituaniană, bulgară, portugheză europeană, macedoneană, poloneză, română, ebraică, italiană și germană. Corpusurile paralele bilingve – engleză și o altă limbă – au conținut în medie peste 10.000 de enunțuri (vezi Tabelul 1), fiecare enunț fiind identificat în mod unic printr-o combinație de trei tipuri de ID-uri. Pentru a obține un corpus paralel coerent în 10 limbi, corpusurile bilingve au fost comparate automat, iar intersecția celor 10 corpusuri a fost identificată și apoi împărțită din nou în corpusuri bilingve, conținând exemplele în limba engleză și exemplele în cealaltă limbă.

Table 1: Setul de date multilingual

Language	**Aligned sentences with MWE**
English	43 600
Macedonian-English	2 846
German-English	15 852
Lithuanian-English	4 112
Bulgarian-English	19 209
European Portuguese-English	4 398
Polish-English	17 408
Romanian-English	18 946
Hebrew-English	23 566

A fost utilizată structura corpusului din Tabelul 2. Primele trei coloane conțin ID-uri, urmate de patru coloane referitoare la enunțul în limba engleză. Acestea includ expresia polilexicală (MWE), o descriere a markerului discursiv (DM), un context scurt în care acesta apare, o fereastră de context mai largă și o adnotare care indică dacă MWE-ul funcționează ca marker discursiv în text. Pentru limbile țintă, altele decât engleza, următoarele patru coloane prezintă aceleași informații corespunzătoare limbii respective. Adnotatorii trebuiau să evalueze dacă MWE-ul joacă rolul de marker discursiv sau nu și să completeze coloana 6 cu 1 (dacă are rol de marker discursiv) sau 0 (dacă are rol de cuvânt de conținut) pentru engleză, iar în coloana 9, corespunzător, pentru limba țintă. Această metodă de reprezentare evită multe dintre complicațiile sistemelor convenționale de adnotare a discursului, în special faptul că markerii discursivi diferiți din aceeași propoziție pot fi adnotați în moduri multiple, iar toate aceste adnotări trebuie condensate într-un format coerent și ușor de citit, care să păstreze suprapunerile între segmentele argumentative și relațiile încrucișate.

Table 2: Structura setului de date

Column	Description
id	Unique identifier
vid	Video unique identifier
lid	Line unique identifier
DM EN	Discourse marker in English
Sentence chunk EN	The sentence in English where the DM appears
Larger Textual Context EN	The full paragraph in English
DM Presence EN	The presence of a DM in English, i.e., 1 present, 0 otherwise
Sentence chunk TL	The sentence in the TL where the DM appears
Larger Textual Context TL	The full paragraph in the TL
DM Presence TL	The presence of a DM in the TL, i.e., 1 present, 0 otherwise DM
Target language	Discourse marker in the TL

Am utilizat un segment adnotat manual și validat din corpusul paralel în engleză, lituaniană, bulgară, și ulterior în italiană, și am antrenat două modele de învățare automată interlingvistice, bazate pe FastText și XLM-RoBERTa-Large, pentru a prezice existența marcatorilor discursivi în texte noi, nevăzute anterior. Pentru antrenarea modelului a fost utilizată o rată de învățare de 0.00001, pe durata a 3 epoci. Biblioteca k-train, construită pe baza bibliotecii transformer HuggingFace, a fost folosită pentru fine-tuning-ul modelului. Împărțirea setului de date pentru antrenare și testare a fost de 80%-20%. Fine-tuning-ul s-a realizat pe parcursul a 10 iterații. Rezultatele acestor experimente, prezentate în Tabelul 3, arată o performanță foarte bună pentru limba lituaniană cu ambele modele, și scoruri diferite pentru cele două modele aplicate pe datele în limba bulgară.

Table 3: Rezultate detecție marcatorii discursivi

Model	Accuracy	Precision		Recall	Specificity	F1-Score	MCC
FastText (EN)	0.4558	0.6515	0.1928	0.8467	0.2976	0.0507	
FastText (BG)	0.5764	0.6457	0.6457	0.4733	0.6457	0.1191	
FastText (LT)	0.9321	0.9369	0.9942	0.0548	0.9647	0.1285	
FastText (IT)	0.5700	0.7400	0.5100	0.6800	0.6000		
XLM-RoBERTa (EN)	0.9180	0.8900	0.7860	0.9130	0.9030	0.8080	
XLM-RoBERTa (BG)	0.8260	0.8260	0.8300	0.8220	0.8290	0.6520	
XLM-RoBERTa (LT)	0.8289	0.9899	0.8242	0.8904	0.8995	0.4393	
XLM-RoBERTa (IT)	0.6900	0.8000	0.6900	0.6900	0.7400	0.3700	

Pornind de la aceste rezultate, ne propunem să integrăm în modelele noastre de detectare a conținutului dăunător și modelele monolingvistice și multilingvistice de detectare a marcatorilor discursivi pentru a îmbunătății acuratețea.

Modele de detecție a conținutului dăunător folosind transformere

Pentru acest set de experimente [2], am utilizat setul de date de la conferință Exist 2025 compus din texte colectate de pe X (fostul Twitter) în Engleză și Spaniolă. Au fost explorate mai multe tehnici de preprocesare pentru a pregăti datele textuale în vederea analizei. Faza inițială, denumită **Curățare Ușoară** (Light Cleaning), a constat în eliminarea emoji-urilor, a adreselor URL și a mențiunilor, precum și în normalizarea spațiilor albe pentru a reduce zgomotul din date. Pornind de la aceasta, etapa de **Curățare Clasică** (Classic Cleaning) a extins procesul prin eliminarea caracterelor non-ASCII, a semnelor de punctuație și a cifrelor, precum și prin convertirea întregului text în litere mici pentru a asigura consistența. A treia metodă, **Curățare Agresivă** (Aggressive Cleaning), a inclus pașii din **Curățarea Clasică**, urmată de eliminarea cuvintelor de legătură/comune (stopwords) și aplicarea algoritmului de stemming, pentru a reduce cuvintele la forma lor de bază. Având în vedere natura bilingvă a datelor, a fost inclus un pas de augmentare, care constă în adăugarea versiunilor traduse ale tweeturilor. Această etapă asigură o reprezentare mai uniformă a conținutului în spaniolă și engleză în cadrul setului de date, îmbunătățind capacitatea de generalizare a modelului. Au fost explorate două soluții de traducere: prima a utilizat API-ul Google Translate, iar a doua a folosit modelul NLLB (No Language Left Behind) dezvoltat de Facebook. Ambele abordări au urmărit menținerea calității traducerii, gestionând în același timp în mod eficient limbajul specific domeniului.

Pentru experimentele noastre, am început prin evaluarea unei game de modele bazate pe arhitectura transformer. Abordarea noastră de clasificare a prioritizat eticheta care apărea cel mai frecvent per tweet; în caz de egalitate, tweetul a fost clasificat ca sexist. Am testat, de asemenea, impactul preprocesării ușoare și am utilizat metricele de acuratețe și scorul F1 pentru a determina modelele cu cele mai bune performanțe (Tabelul 4).

Table 4: Rezultate preprocesare

Model	Pre-processing	Accuracy	F1-No	F1-Yes
DeBERTa	Yes	0.80	0.82	0.77
DeBERTa	No	0.81	0.82	**0.79**
mDeBERTa-large	Yes	0.76	0.81	0.69
mDeBERTa-large	No	0.79	0.82	0.75
mDeBERTa-base	Yes	0.80	0.83	0.76
XLM-RoBERTa	Yes	0.80	0.83	0.76
XLM-RoBERTa	No	0.80	0.83	0.76
HateBERT	Yes	0.75	0.73	0.74
HateBERT	No	0.78	0.81	0.75
Detoxify	Yes	0.78	0.80	0.77
Detoxify	No	0.77	0.82	0.71
RoBERTa-hate-speech	Yes	0.76	0.80	0.68
RoBERTa-hate-speech	No	0.78	0.81	0.74

În a doua fază, ne-am concentrat pe evaluarea eficienței celor trei strategii de preprocesare folosind modelul mDeBERTa. Dintre metodele Curățare Ușoară, Curățare Clasică și Curățare Agresivă, abordarea Curățare Clasică a oferit cele mai bune rezultate (Tabelul 5). Pe baza acestor constatări inițiale, am selectat modelele mDeBERTa și XLM-RoBERTa-base pentru antrenare și evaluare suplimentară în cadrul celor trei sarcini, deoarece ambele au demonstrat performanțe solide.

Table 5: Rezultate performanță model

Model	Pre-processing	Accuracy	F1-No	F1-Yes
mDeBERTa	Light	0.81	0.82	0.80
mDeBERTa	Classic	0.82	0.83	0.80
mDeBERTa	Aggressive	0.80	0.80	0.80

Pornind de la aceste experimente destul de bune, ne-am decis să investigăm ce se întâmplă dacă folosim modele auto-adaptive de tranformere pentru detectarea conținutului dăunător.

Modele de detecție a conținutului dăunător folosind modele auto-adaptive de transformere

Pentru acest set de experimente [3,4], am utilizat o combinație de transformere exclusiv pentru limba engleză și modele multilingve, provenite de pe HuggingFace. Această selecție ne permite să construim modele care se auto adaptează la limba în care au fost scrise postările. Modulul folosește 3 tipuri de combinații, în funcție de ponderile obținute de cele mai bune modele în limba engleză și cele multilingve. Considerăm modelul dominant pe cel în limba engleză, în cazul în care limba inputului este engleza, iar în caz contrar pe cel multilingv. Când prezentăm rezultatele, combinațiile utilizate sunt:

- Half-Half,
- Dominant-75\%
- Dominant.

Sistemul trebuie să decidă dacă un tweet dat conține sau nu expresii sau comportamente sexiste. Setul de date este adnotat de mai mulți evaluatori, fiecare oferind propria etichetă. Pentru a unifica aceste etichete, facem o medie cu greutate egală, rezultând o singură etichetă: „DA” sau „NU”. Cele mai performante modele pentru această sarcină sunt cele care au fost fine-tunate pe date din Twitter (Tabelul 6). Așa cum am menționat în secțiunea anterioară, vom folosi o combinație între cel mai bun model bazat pe limba engleză și cel mai bun model multilingv.

Table 6: Rezultate performanță model auto-adaptive de transformere.

ModelName	Epoch	F1-Score	Loss	Train(s)	Eval(s)
twitter-roberta	**4**	**0.7789**	**0.4863**	**1 463**	**5**
twitter-xlm-roberta-base-sentiment-multilingual	*3*	*0.7665*	*0.4670*	*7 039*	*159*
twitter-xlm-roberta-base-sentiment	3	0.7482	0.4902	6 373	146
bert-toxic-comment-classification	4	0.7463	0.5211	6 969	117
distilbert-uncased-english	4	0.7406	0.5348	2 918	3
distilbert-base-multilingual-cased-sentiments	4	0.7379	0.5123	4 407	3
MiniLM-L12-H384	5	0.7338	0.5059	296	3
xlm-roberta	4	0.7327	0.5520	1 834	9
roberta-hate-speech-dynabench-r4	3	0.7126	0.5220	6 080	154

Eficiența abordării propuse de noi se reflectă în rezultate. Sistemul propus care utilizează modele auto-adaptive de transformere demonstrează robustețea prin eficiența ridicată dată de scorul F1.

2. Activități de publicare

În contextul proiectului NetGuardAI, au fost publicate 4 articole științifice cu afilierea AOȘR. Dintre acestea, un articol a fost publicat într-un jurnal cotat Q1 ISI conform clasificării WOS Journal Citation Reports (JCR) pentru anul 2024, având un factor de impact de 1.7. Progresul realizat până în prezent depășește planificarea pentru această perioadă conform diagramei GANTT (Figura 1).

Listă publicații:

[1] Elena-Simona Apostol, **Ciprian-Octavian Truică**, Mariana Damova, Purificação Silvano, Giedre Valunaite Oleškeviciene, Chaya Liebeskind, Dimitar Trajanov, Anna Baczkowska, Emma Angela Montecchiari, Christian Chiarcos. *Multiword Discourse Markers Across Languages: A Linguistic and Computational Perspective*, International Journal of Applied Linguistics, Wiley, ISSN 0802-6106, 2025. DOI: https://doi.org/10.1111/ijal.12755 (**Q1 Journal**)

[2] Maria-Diana Cotelin, Elena-Simona Apostol, **Ciprian-Octavian Truică**. *NetGuardAI at EXIST2025: Sexism Detection using mDeBERTa*, Working Notes of the Conference. and Labs of the Evaluation Forum (CLEF 2025), 2025.

[3] **Alexandru Petrescu**, **Ciprian-Octavian Truică**, Elena-Simona Apostol. *Language-based Mixture of Transformers for Sexism Identification in Social Networks*, Conference and Labs of the Evaluation Forum (CLEF 2025), 2025.

[4] **Alexandru Petrescu**, Elena-Simona Apostol, **Ciprian-Octavian Truică**. *Awakened at EXIST2025: Adaptive Mixture of Transformers*, Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025), 2025.

Data: 15.07.2025

Semnătură director,
Ș.L. Dr. Abil. Ing. Habil. Ciprian-Octavian TRUICĂ

Semnături membri,
As. Drd. Ing. Alexandru PETRESCU

Drd. Ing. Anamaria VIȘAN