

## Raport 2

Proiect:

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

- Decembrie 2025 -

**Director proiect:** Conf. 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 automată;
2. Protejarea utilizatorilor de conținutul dăunător prin strategii de imunizare.

## 1. Activități de cercetare

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

- Analiza cerințelor și soluțiilor similare (terminat)
- Colectarea și procesării de date (în curs)
- Implementarea modulelor de detectare a conținutului dăunător (în curs)
- Evaluare modelelor de învățare automată (în curs)
- 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												
T2.2	Analiza critică a soluțiilor de imunizare									D2.2	M2.1											
T2.3	Cerințe funcționale și non-funcționale									D2.3												
T2.4	Cerințe tehnologice									D2.3	M2.2											
WP3	Colectarea și procesarea datelor																					
T3.1	Analiza seturilor de date disponibile online									D3.1												
T3.2	Colectarea noi date și metadata										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																					

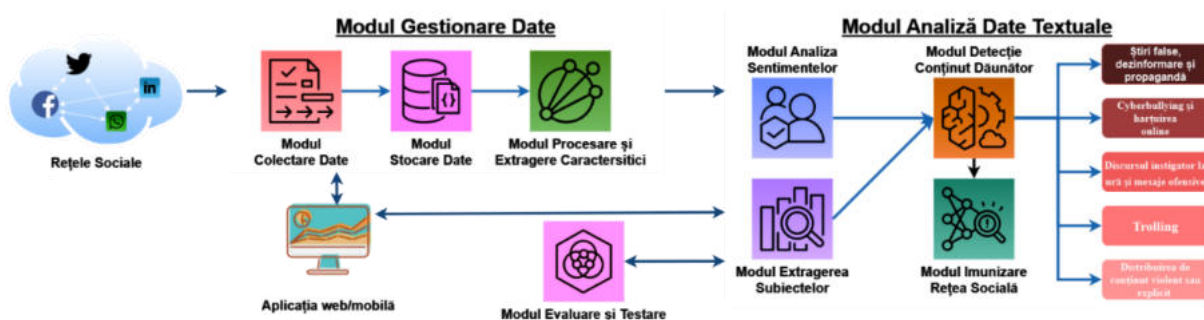
Figură 1. Diagrama GANTT

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 s-au încheiat, însă se va reveni la el în cazul în care este necesară reactualizarea statusului actual în cercetare pentru viitoare publicații.

- **WP3 – Colectarea și procesarea datelor:** analiza 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<sup>1</sup>), Engleză și Spaniolă (Exist 2025<sup>2</sup>), LIAR<sup>3</sup>, Hate Speech Offensive<sup>4</sup>, Online Sexism<sup>5</sup>, ISOT<sup>6</sup>, Twitter15 și Twitter16<sup>7</sup>. S-a început implementarea unui modul pentru procesarea datelor textuale.
- **WP4 – Implementarea sistemului NetGuardAI:** implementarea modulelor de detecție a conținutului dăunător și implementarea strategiilor de imunizare a rețelelor sociale. Am realizat dezvoltarea unor modele de analiză de sentimente și de detecție a conținutului dăunător bazate pe vectorizări de tip transformer (ex. BERT, RoBERTa, DeBERTa, etc.) și modele lingvistice mari (LLMs – Large Language Models) precum Llama2. Modelele dezvoltate iau 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 7 articole în total, dintre care 3 pentru această etapă (mai multe detalii în *Secțiunea 2. Activități de publicare*). Am creat un depozit pentru date și cod pe platforma Github<sup>8</sup>.

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 și analiza sentimentelor. În etapa următoare, intenționăm să demarăm dezvoltarea modului 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

<sup>1</sup> <https://huggingface.co/datasets/mateiaass/FakeRom>

<sup>2</sup> <https://nlp.uned.es/exist2025/>

<sup>3</sup> [https://www.cs.ucsb.edu/~william/data/liar\\_dataset.zip](https://www.cs.ucsb.edu/~william/data/liar_dataset.zip)

<sup>4</sup> [https://huggingface.co/datasets/tdavidson/hate\\_speech\\_offensive](https://huggingface.co/datasets/tdavidson/hate_speech_offensive)

<sup>5</sup> <https://github.com/rewire-online/edos>

<sup>6</sup> <https://onlineacademiccommunity.uvic.ca/isot/2022/11/27/fake-news-detection-datasets/>

<sup>7</sup> <https://www.dropbox.com/scl/fi/flgahafqckxtup2s9eez8/rumdetect2017.zip>

<sup>8</sup> <https://github.com/orgs/DS4AI-UPB/repositories>

## *Detectarea limbajului dăunător folosind modele lingvistice mari*

Am dezvoltat un model bazat pe LLaMA2, pe care l-am antrenat folosind tehnica LoRA (Low-Rank Adaptation), numit HarmLLaMA. HarmLLaMA ce permite antrenarea eficientă a modelelor mari cu costuri computaționale reduse. Obiectivul nostru a fost să evaluăm dacă un model pre-antrenat de dimensiuni mari poate fi specializat cu succes pentru detecția limbajului ofensator, a urii, sexismului și mesajelor agresive, întrecând performanțele soluțiilor din literatura curentă. În paralel, am investigat influența preprocesării asupra performanței modelului și impactul ajustării hiperparametrilor.

Pentru a fundamenta cercetarea, am construit un flux de lucru complet de procesare, care include:

- curățarea datelor (eliminarea tagurilor, linkurilor, emoji-urilor, caracterelor ne-ASCII etc.),
- tokenizarea textelor folosind tokenizator pe sub-cuvinte,
- antrenarea modelului cu LoRA, prin adăugarea unor matrici de rang redus pe straturile de atenție ale LLAMA2,
- integrarea unui cap de clasificare (classification head) liniar pentru clasificarea finală.

Am ales două seturi de date reale și relevante pentru problemă:

1. Hate Speech Offensive un set debalansat, cu peste 24.000 de postări X/Twitter și Facebook, etichetate ca „offensive”, „hate speech” sau „neither”;
2. Online Sexism un set de date bine balansat, cu aproximativ 20.000 de postări pe X/Twitter, etichetate ca „sexist” sau „non-sexist”.

Aceste seturi de date ne-au permis să evaluăm modelul în contexte diferite: atât clasificare binară, cât și multi-clasă, atât pe date curate cât și zgomotoase, și atât pe distribuții echilibrate, cât și puternic dezechilibrate.

În etapa experimentală, am analizat în detaliu fiecare element al fluxului de lucru. Preprocesarea s-a dovedit esențială în cazul setului de date Hate Speech Offensive, unde acuratețea HarmLLaMA a crescut semnificativ după curățarea datelor, în timp ce pentru setul de date Online Sexism impactul a fost minim. Am testat diferite numere de epoci și am stabilit că 2 epoci reprezintă un echilibru optim între performanță și evitarea supraînvățării (overfitting). Am investigat efectele mărimii batch-ului și am concluzionat că batch = 20 duce la un timp de convergență mai rapid și performanțe mai stabile. De asemenea, am realizat un proces sistematic de ajustare a hiperparametrilor, analizând combinații variate ale ratei de învățare și rata de învățare (learning rate - lr) și penalizarea ponderilor (weight decay - wd). Rezultatele au arătat că rata de învățare este mult mai decisivă pentru performanță decât penalizarea ponderilor, iar valorile optime identificate au fost: lr = 9e-5, wd = 1e-2.

Pentru o evaluare robustă, am aplicat validarea încrucișată (3-fold cross-validation), folosind stratificare pentru setul de date dezechilibrat. Acest pas ne-a permis să confirmăm generalizarea bună a modelului și stabilitatea predicțiilor pe subseturi variate ale datelor. Variabilitatea redusă între antrenări indică faptul că HarmLLaMA nu este sensibil la împărțirile setului de date și nu suferă de supraînvățare.

Unul dintre obiectivele noastre principale a fost compararea modelului cu alte tehnici moderne. Am antrenat și un model BERT-cased pentru setul de date Hate Speech Offensive și am folosit BERTweet, modelul specializat în procesarea postărilor de pe X/Twitter, pentru setul de date Online Sexism. Rezultatele au demonstrat că HarmLLaMA depășește constant atât modelele BERT, cât și alte soluții din literatură, în toate metricele: acuratețe, precizie, recall și F1-score. De exemplu, pe setul de date Hate Speech Offensive, HarmLLaMA atinge un scor F1 de 0.93 (Tabelul 1), peste BERT și peste alte modele hibride CNN/BERT. Pe setul de date Online Sexism, modelul nostru obține un F1 scor de 0.91 (Tabelul 2), depășind sisteme precum RoBERTaLarge, DeBERTa și diverse arhitecturi ansamblu.

Tabel 1. Rezultate obținute pe setul de date Hate Speech Offensive

Model	Accuracy	Precision	Recall	F1-Score
SVM [14]	N/A	0.91	0.90	0.91
BERT	0.90	0.89	0.90	0.90
BERT <sub>base</sub> +LSTM [26]	N/A	0.91	0.92	0.92
BERT <sub>base</sub> +CNN [26]	N/A	0.92	0.92	0.92
HARMLLAMA	0.93	0.92	0.93	0.93

Tabel 2. Rezultate obținute pe setul de date Online Sexism

Model	Accuracy	Precision	Recall	F1-Score
BERT [35]	0.88	0.86	0.85	0.85
Model-Ensemble [34]	N/A	N/A	N/A	0.85
RobertaLarge [28]	N/A	N/A	N/A	0.86
DeBERTa [20]	N/A	N/A	N/A	0.88
HARMLLAMA	0.91	0.90	0.90	0.91

Un alt aspect important sunt limitările modelului. Deși performanța este ridicată, modelul depinde în continuare de calitatea datelor de antrenament, iar biasurile din acestea pot afecta deciziile. Detectarea limbajului dăunător este adesea subiectivă, iar distincțiile dintre ofensiv și discurs legitim sunt uneori subtile chiar și pentru adnotatori umani. În plus, deși LoRA reduce costurile, antrenarea modelelor lingvistice mari rămâne dificil de aplicat în sisteme cu resurse limitate. Modelul are, de asemenea, limitări inerente de interpretabilitate, ceea ce poate ridica probleme în domenii sensibile, precum moderarea automată.

În concluzie, cercetarea noastră demonstrează că HarmLLaMA este un model extrem de competitiv pentru detecția limbajului dăunător, depășind soluțiile consacrate din literatură. Pentru viitor, planurile noastre includ extinderea modelului către mai multe seturi de date, testarea altor LLM-uri mai ușoare, precum Gemma, și integrarea HarmLLaMA în sisteme de detecție în timp real, alături de arhitecturi precum ContCommRTD sau StopHC.

În al doilea set de experimente am dezvoltat o arhitectură nouă pentru detectarea credibilității informațiilor [2], construită pe baza modelului ALBERT, pe care l-am adaptat prin fine-tuning pentru clasificarea știrilor ca adevărate sau false. Metodologic, am structurat soluția în trei componente esențiale. Mai întâi, am aplicat un modul riguros de preprocesare, prin care am standardizat textele prin normalizare, eliminarea duplicatelor, curățarea spațiilor, a adreselor URL și a valorilor lipsă. În continuare, am folosit modelul ALBERT ca bază, înghețând straturile de encoder și antrenând doar straturile superioare, astfel încât modelul să păstreze cunoștințele lingvistice pre-antrenate, dar să învețe pe folosind seturile de date sarcina de detecție. În plus, am introdus un modul Weight Learner, capabil să ajusteze dinamic hiperparametrii, în special rata de învățare, pentru a accelera înghețarea și a preveni supraînvățarea. Am folosit tokenizarea ALBERT și o lungime maximă de 256 tokeni, iar pentru evaluare am aplicat mecanisme clasice pentru a testa robustețea modelului precum determinarea celor mai buni parametrii (hyperparameter tuning) și validarea în cruce (K-fold cross-validation), .

Pentru validare, am utilizat două seturi de date: LIAR și ISOT. În cazul LIAR, după binarizarea etichetelor și eliminarea anomaliilor, modelul a întâmpinat dificultăți din cauza dezechilibrului de clase, motiv pentru care am folosit diferite metode de eșantionare. Acuratețea finală obținută a fost de 65%,

depășind alte modele precum LSTM, CNN sau chiar ALBERT pre-antrenat. Pe datasetul ISOT, mult mai mare și echilibrat, modelul a obținut o acuratețe de 98%, cu pierdere de antrenare și validare în scădere constantă. Experimentele au arătat că ratele mici spre medii de învățare conduc la cele mai bune performanțe, iar modulul de învățare adaptivă a hiperparametrilor a permis o convergență mai rapidă și o performanță mai stabilă. Comparativ cu alte modele din literatură, soluția noastră a obținut cele mai bune rezultate pe ambele seturi de date, demonstrând că un model compact precum ALBERT, combinat cu tehnici eficiente de fine-tuning, poate depăși arhitecturi mai complexe în detectarea știrilor false.

### *Model de ansamblu pentru analiza sentimentelor bazată pe aspecte*

Am propus și am dezvoltat o arhitectură ansamblu heterogenă [3] concepută pentru a aborda în mod eficient problema complexă a analizei sentimentelor bazate pe aspecte (ABSA). Motivația principală a pornit de la limitările modelelor actuale, care, deși funcționează bine în analiza sentimentelor la nivel de propoziție sau document, întâmpină dificultăți majore atunci când trebuie să identifice aspecte multiple, adesea cu polarități diferite, în interiorul aceluiași text. Am considerat că această problemă necesită o abordare modulară, scalabilă și mai ales diversificată din punct de vedere al reprezentărilor lingvistice.

Pentru a reduce complexitatea sarcinii, am împărțit ABSA în două etape: extragerea termenilor pentru aspect (ATE) și analiza sentimentelor pentru termeni (ATSA). În cadrul fiecărei etape, am construit un ansamblu format din șase modele care combină trei arhitecturi neuronale, Linear, BiLSTM și CNN-BiLSTM, cu două categorii de vectorizări de termeni derivate din modele Transformer, respectiv BERT și BART, atât în varianta lor pre-antrenată, cât și în versiuni antrenată special pentru sarcina analizată. Ideea centrală a fost să obțin o diversitate ridicată a predicțiilor, astfel încât ansamblul să beneficieze de complementaritatea arhitecturilor și să atingă o generalizare superioară. Pentru agregarea rezultatelor tuturor modelelor am folosit o strategie simplă, dar eficientă: alegerea clasei majoritare folosind argmax pe scorurile de clasificare.

Am testat arhitectura pe două dintre cele mai relevante seturi de date existente pentru ABSA: SemEval 2016 și MAMS (Multi-Aspect Multi-Sentiment). Acestea oferă atât situații cu un număr redus de aspecte per propoziție, cât și exemple cu densitate mare de opinii pentru categorii. Rezultatele au confirmat eficiența abordării: am obținut până la 99.96% acuratețe pentru ATE și 99.93% pentru ATSA (cum se vede în Tabele 5-7), depășind semnificativ modelele din literatura curentă, inclusiv variante avansate de BERT, RoBERTa, modele de tip graf sau metode bazate pe arhitecturi cu atenție. Această performanță a fost obținută cu doar 2 epoci de antrenare, ceea ce demonstrează că fine-tuningul bine structurat poate duce la rezultate excelente chiar și pe seturi de date de dimensiuni moderate.

*Tabel 5. Rezultate ATE pe setul de date SemEval*

ATE results comparison on the SemEval datasets (Note: highlighted cells show the best results).

Model	Dataset	F1-Score
CRF [48]	SE2014T4R	79.62
W+L+D [49]	SE2014T4R	84.31
W+L+D+B [49]	SE2014T4R	84.97
CMLA [50]	SE2014T4R	77.80
DE-CNN-CRF [14]	SE2016T5R	74.10
DE-CNN [14]	SE2016T5R	74.37
<b>ATESA-BERT - our solution</b>	SE2016T5R	<b>93.80</b>

Tabel 6. Rezultate ATSA pe setul de date SemEval

ATSA comparison on the SemEval dataset (Note: highlighted cells show the best results).

Model	Dataset	Accuracy
BERT-single [51]	SE2014T4R	93.3
BERT-pair-QA-M [51]	SE2014T4R	95.4
BERT-pair-NLI-M [51]	SE2014T4R	94.4
BERT-pair-QA-B [51]	SE2014T4R	95.6
BERT-pair-NLI-B [51]	SE2014T4R	95.1
BERT for ABSA [21]	SE2016T5R	89.8
SA-BERT [22]	SE2016T5R	92.02
SA-BERT-XGBoost [22]	SE2016T5R	93.71
DeBERTa [23]	SE2016T5R	89.46
<b>ATESA-BÆRT - our solution</b>	SE2016T5R	<b>99.84</b>

Tabel 7. Rezultate ATSA pe setul de date MAMS

ATSA results comparison on the MAMS dataset (Note: highlighted cells show the best results).

Model	Accuracy
HAGNN-BERT [52]	66.92
HAGNN-GloVe [52]	72.58
CapsNet-BERT [46]	83.39
CapsNet-BERT-DR [46]	82.97
RoBERTa-TMM [54]	85.64
TransEncAsp+SCAPT [55]	80.54
BERTAsp+SCAPT [55]	85.63
RGAT-BERT [56]	84.52
AGIAN-BERT [30]	82.02
<b>ATESA-BÆRT - our solution</b>	<b>99.93</b>

În concluzie, am demonstrat că un ansamblu heterogen, construit pe baza unor transformere ajustate prin antrenare și modele neuronale tradiționale, poate depăși obstacolele majore ale ABSA și poate oferi rezultate robuste, precise și generalizabile. Direcțiile viitoare includ extinderea modelului către mai multe domenii, optimizarea performanței pe texte zgomotoase și explorarea unor variante de modele lingvistice mari, care să reducă resursele de calcul fără a compromite acuratețea. Acest studiu va sta la baza viitoarelor cercetări în detectarea conținutului dăunător bazat pe aspecte.

## 2. Activități de publicare

În contextul proiectului NetGuardAI, au fost publicate 3 articole științifice cu afilierea AOȘR pentru acest raport. Dintre acestea, un articol a fost publicat într-un jurnal cotate Q1 ISI conform clasificării WOS Journal Citation Reports (JCR) pentru anul 2024, având un factor de impact de 7.6. Progresul realizat până în prezent depășește planificarea pentru această perioadă conform diagramei GANTT (Figura 1).

Listă publicații raport 2:

- [1] **Ciprian-Octavian Truică**, Elena-Simona Apostol, Alexandru-Gabriel Ilie, Adrian Paschke. *HarmLLaMA: Harmful Language Detection with Large Language Models*. International Conference on Intelligent Computer Communication and Processing (ICCP 2025), October 2025 (Rank National: Romania Conference) (BDI: **ISI, IEEE, Scopus**)
- [2] Daria-Elena Burghilea, **Ciprian-Octavian Truică**, Elena-Simona Apostol. *VERIT-ALBERT: A Finetuned LLM Approach for Verifying Information Credibility*. RoEduNet Conference Networking in Education and Research, September 2025. DOI: [10.1109/RoEduNet68395.2025.11208267](https://doi.org/10.1109/RoEduNet68395.2025.11208267) (BDI: **IEEE, Scopus**)
- [3] Elena-Simona Apostol, Alin-Georgian Pistică, **Ciprian-Octavian Truică**. *ATESA-BÆRT: A heterogeneous ensemble learning model for Aspect-Based Sentiment Analysis*. Knowledge-Based Systems, 326:1-13(113987), ISSN 0950-7051, Septembrie 2025. DOI: [10.1016/j.knosys.2025.113987](https://doi.org/10.1016/j.knosys.2025.113987) (**Q1 Journal ISI, IF=7.6**)

Data: 04.12.2025

Semnătură director,  
*Conf. Dr. Abil. Ing. Ciprian-Octavian TRUICĂ*

---

Semnături membri,  
*As. Drd. Ing. Alexandru PETRESCU*

---

*Drd. Ing. Anamaria VIȘAN*

---