Raport 2

Proiect:

SCAN-NEWS: Sistem inteligent pentru detecția și atenuarea răspândirii dezinformării și știrilor false în rețelele sociale

- Decembrie 2024 -

Director proiect: Conf. Habil. Dr. Ing. Elena-Simona APOSTOL
Membru: Ș.L. Dr. Ing. Mircea-Alexandru PREDESCU

Tranziția către era digitală a dus la o schimbare radicală a peisajului media, cu accent pe personalizarea conținutului și interacțiunea utilizatorilor. Această evoluție a diminuat importanța jurnalismului de investigație și a facilitat răspândirea rapidă a dezinformării. Cercetările actuale în acest domeniu se limitează adesea la analiza textului mesajelor, ignorând factorii sociali, culturali și politici care influențează receptarea și interpretarea acestora.

Raportul de față prezintă o analiză detaliată a activităților de cercetare și diseminare derulate în cadrul proiectului SCAN-NEWS, în intervalul August – Decembrie 2024.

1. Activități de cercetare

Proiectul SCAN-NEWS are ca prin focus analiza diferitelor tipuri de dezinformări prezente în rețelele sociale, precum:

- Știri false (fake news): reprezintă informații false prezentate ca fiind adevărate
- Dezinformare parțială: constă în informații adevărate, dar prezentate într-un mod înșelător
- Satire și umor confundate cu realitatea: sunt știri sau postări satirice sau umoristice care sunt interpretate greșit ca fiind adevărate
- Conspirații: constă în teorii nefondate care explică evenimente sau fenomene complexe prin intermediul unor explicații simple, dar false
- Misinformare: sunt informații incorecte, dar răspândite fără intenție de a înșela, ci din simplă neștiință sau eroare
- Etc.

Obiectivul principal al proiectului SCAN-NEWS constă în constă în două direcții majore: cercetarea și dezvoltarea de

- $\mathbf{[O_1]}$ noi modele de detecție de tip ansamblu care să considere în timpul antrenării pe lângă contextul cuvintelor extras cu transformere și informații despre polaritatea și tematica știrilor cât și graful de difuzie al știrii în rețeaua socială;
- $\mathbf{[O_2]}$ noi strategii de imunizare a rețelelor sociale care au ca scop principal oprirea răspândirii online a informațiilor false.

În cadrul direcției $\mathbf{[O_1]}$, ne propunem să să creăm noi tip de reprezentare a informațiilor pentru a îmbunătății performanțele modelelor de detecție a diferitelor tipuri de dezinformare din rețelele sociale. Aceste reprezentări vectoriale conțin atât informația postată de diferiții utilizatori (numită mai departe context local), cât și informații legate de reacțiile utilizatorilor la respectivele postări, subiectele discutate și polaritatea (numit de noi context global).

Diagrama GANTT a proiectului																				
WP	Luna	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
WP1	Activități de management																				
S1.1	Management administrativ				D1a				D1b						D1c						D1d
S1.2	Management tehnic									D1e						D1f					
WP2	Analiza cerințelor și soluțiilor similare existente																				
S2.1	Analiza critică a soluțiilor existente			D2a																	
S2.2	Cerințe funcționale și non-funcționale				D2b																
S2.5	Cerințe tehnologice																				
WP3	Colectarea și adnotarea datelor																				
S3.1	Analiza seturilor de date disponibile online					D3a															
S3.2	Colectarea de noi date din mediile sociale							D3b													
S3.3	Preprocesare și extragerea caracteristicilor								D3c	M3a											
WP4	Proiectarea și implementarea sistemului																				
S4.1	Proiectarea arhitecturii																				
S4.2	Implementarea algoritmilor de Învățare Automată													D4a							
S4.3	Scierea codului modulelor arhitecturii																D4b				
S4.4	Integrarea modulelor																	D4c	M4a		
S4.5	Proiectarea interfeței cu utilizatorii																	D4d			
WP5	Evaluare și validare																				
S5.1	Definirea scenariilor de test															D5a					
S5.2	Testare experimentală a modulelor și sistemului														D5b		D5c		D5d		M5a
S5.3	Optimizări ale sistemului																			D5e	
WP6	Diseminare și exploatare																				
S6.1	Diseminarea rezultatelor							D6a				D6b				D6c			D6d		
S6.2	Exploatare sistem											M6a									M6b

Figură 1. Diagrama GANTT

Conform calendarului de activități prezentat în propunerea de proiect și ilustrat în Figură 1. Diagrama GANTT, principalele activități de cercetare derulate în cadrul etapei 2 a proiectului SCAN-NEWS sunt următoarele:

- **(WP3) Colectarea și adnotarea datelor:** S-au continuat activitățile de colectare, stocare și adnotare a datelor obținute din platforme sociale. Pentru colectare, la acest moment, am folosit platforma X (Twitter). Au fost colectate atât date în limba engleză cât și în limba română, majoritatea fiind în engleză. În urma metodologiei definite în etapa anterioară, s-a început procesul de adnotare pentru adnotare. Am creat o pagină web ce poate fi accesată de adnotatori. Aceștea primesc postări aleatorii și sunt rugați să eticheteze postarea cu una din clasele disponibile. Fiecare postare, este adnotată de minim 3 adnotatori. Adițional, am mai colectate și alte data seturi publice deja adnotate. Aceste data seturi sunt fie în limba engleză, fie multilingve și sunt din platformele Twitter sau Facebook. O parte din aceste data seturi au și informații de rețea (ex., ID-ul utilizatorului care a postat original, cât și ID-urile utilizatorilor care au diseminat informația sau au reacționat la postul respectiv) pentru a se putea reconstitui graful de propagare.

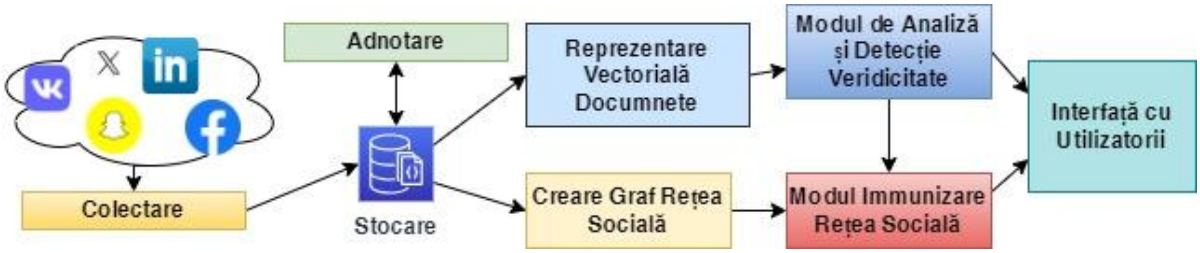
- **(WP4) Modelarea unei soluții noi de Analiză a sentimentelor**: S-au construi și implementat modele Deep Learning de tip ansamblu pentru 1) a extrage poliritatea mesajelor din mediul social ce conțin știri false, 2) a detecta poziția autorului ce a distribuit respectivul mesaj.

- **(WP4) Construirea de modele pentru analiza semantică latentă (*latent semantic analysis*) și extragerea subiectelor *(topic modeling)***

- **(WP4) Algoritm de detecție:** În cadrul aceste acțiuni s-au implementarea noi algorimit de Deep Learning pentru detecția diferitelor tipuri de manipulări din rețelele sociale.

- **(WP6) Diseminarea și exploatare**: Pentru această etapă, au fost publicate 3 articole, două de jurnal Q1 și un articol de conferință (mai multe detalii în *Secțiunea 2. Activități de publicare*)

Arhitectura soluției finale este prezentată în Figură 2. Arhitectura generală SCAN-NEWS. Până la finalizarea raportului 2, am realizat sarcini de proiectare, implementare și testare pentru toate modulelor. În continuarea vom prezenta detalii de implementare și rezultatele obținute pentru modulele implementate și testate în cadrul etapei 2. Ca mențiune, acest raport conține doar rezultatele obținute care sunt deja publicate.



Figură 2. Arhitectura generală SCAN-NEWS

Analiza și Detecția Veridicității Informațiilor postate în Mediile Sociale

În cadrul acestei secțiuni vom prezenta succint următoarele componente: *Sub-Modulul Analiza sentimentelor, Sub-Modulul Topic Modeling, Modulul de Analiză și Detecție, și Interfața cu Utilizatorul.*

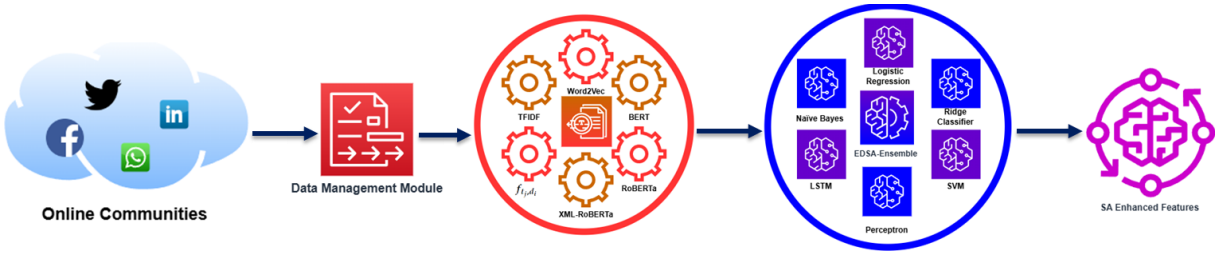
Sub-Modulul Analiza sentimentelor

Analiza sentimentelor poate fi un instrument puternic în îmbunătățirea detectării dezinformarii și a stirilor false din rețelele sociale, prin:

- Detectarea sentimentelor extreme: postările cu sentimente extrem de pozitive sau negative, mai ales atunci când nu sunt susținute de dovezi concrete, pot indica dezinformare sau propagandă.
- Identificarea sentimentului inconsecvent: Contradicțiile dintre sentimentul exprimat în text și faptele subiacente pot fi un indicator pentru conținut înșelător.

Pentru submodulul de Analiza a sentimentelor, ne-am propus să construim modele pentru extragerea polarității la nivel de conținut și a poziției autorului referitor la știrile/informațiile discutate de acesta.

Prin acest sub-modul dorim să realizăm și detecția poziției autorului față de subiect (-l.eng. stance detection). Prin această metodă se analizează stilurile de scris ale autorilor și poate fi utilizată pentru a detecta diferite tipuri de dezinformare, de exemplu, propagandă, conspirație. Pentru extragerea polarității, am propus o noua solutie de tip ansamblu, a carei diagrama de functionare este prezentată în figura de mai jos.



Solutia aplică diferite metode de analiză a sentimentelor pe seturile de date preprocesate, de exemplu, modele clasice de Machien Learning (ex. Naïve Bayes, Support Vector Machines) și modele Deep Learning (ex. BiLSTM, Transformers) și utilizează un sistem de vot de tip ansamblu pentru a determina clasa corectă pentru fiecare document (i.e., postare dată ca input).

Soluția propusă de noi a fost antrenată pe setul de date disponibil online Setiment140, ce conține 1.6 milioane de tweet-uri adnotate (pozitiv, neutru sau negativ). Pentru a verifica scalabilitatea soluției propuse de noi, am împărțit setul de date în 3 subseturi: C1, C2, si C3 (C3 contine intreg setul de date). Modul de împărțire este descris în tabelul de mai jos.

Dataset	Tweets	SCT features	SFE Features
C_1	20 000	30 651	31 018
C_2	500 000	307 890	311 240
C_3	1 600 233	684 492	691 646

Am folosit mai multe tehnici de preprocesare a textului, înainte de a da informația către algoritmi de Învățare Automată:

- Sentiment Clean Text (SCT) - textul este divizat în tokeni și semnele de punctuație sunt eliminate. Pentru a păstra polaritatea sentimentului, se păstrează majusculele, literele duplicate din cuvinte și stop words
- Sentiment Clean Text with Feature (SFE) - în plus față de divizarea textului în tokeni și eliminarea semnelor de punctuație, se aplică o tehnica de feature engineering: 1) contracțiile sunt extinse, de exemplu, "don't" devine "do not, 2) textul este îmbunătățit prin combinarea negației cu cuvintele care o urmează, de exemplu, "not good" devine "not_good".

Dataset	Algorithm	Term Weight	SCT Accuracy	SCT Precision	SCT Recall	SFE Accuracy	SFE Precision	SFE Recall
C_1	NB	f_{t_j,d_i}	0.761	0.706	0.841	0.754	0.694	0.852
	LR	$TFIDF$	0.756	0.703	0.766	0.752	0.690	0.752
	RC	f_{t_j,d_i}	0.767	0.710	0.852	0.753	0.690	0.861
	SVM	$TFIDF$	0.729	0.707	0.772	0.748	0.725	0.796
	LSTM	Word2Vec	0.930	0.931	0.930	0.923	0.922	0.923
	Perceptron	BERT	0.828	0.828	0.828	0.776	0.776	0.776
	Perceptron	RoBERTa	0.866	0.866	0.866	0.777	0.777	0.777
	Perceptron	XLM-RoBERTa	0.828	0.828	0.828	0.776	0.776	0.776
	EDSA Ensemble		**0.949**	**0.950**	**0.948**	**0.940**	**0.941**	**0.940**
C_2	NB	f_{t_j,d_i}	0.794	0.752	0.857	0.788	0.738	0.867
	LR	$TFIDF$	0.790	0.727	0.802	0.794	0.730	0.807
	RC	f_{t_j,d_i}	0.810	0.772	0.862	0.794	0.747	0.865
	SVM	$TFIDF$	0.778	0.799	0.742	0.781	0.763	0.815
	LSTM	Word2Vec	0.855	0.859	0.855	0.850	0.851	0.850
	Perceptron	BERT	0.856	0.856	0.856	0.800	0.800	0.800
	Perceptron	RoBERTa	0.887	0.887	0.887	0.809	0.809	0.809
	Perceptron	XLM-RoBERTa	0.863	0.863	0.863	0.807	0.807	0.807
	EDSA Ensemble		**0.901**	**0.900**	**0.902**	**0.877**	**0.878**	**0.877**
C_3	NB	f_{t_j,d_i}	0.796	0.749	0.863	0.786	0.731	0.873
	LR	$TFIDF$	0.800	0.739	0.810	0.804	0.742	0.815
	RC	f_{t_j,d_i}	0.814	0.774	0.865	0.794	0.745	0.866
	SVM	$TFIDF$	0.803	0.795	0.820	0.806	0.795	0.825
	LSTM	Word2Vec	0.824	0.827	0.824	0.835	0.835	0.835
	Perceptron	BERT	0.863	0.863	0.863	0.807	0.807	0.807
	Perceptron	RoBERTa	0.892	0.892	0.892	0.817	0.817	0.817
	Perceptron	XLM-RoBERTa	0.870	0.870	0.870	0.814	0.814	0.814
	EDSA Ensemble		**0.912**	**0.913**	**0.910**	**0.872**	**0.873**	**0.872**

Cele mai bune rezultate le obtinem cand folosim prima tehnica de preprocesare, SCT.
În tabelul de mai sus, se poate observa ca arhitectura de tip ansablu propusa (numită de noi EDSA), identifică corect sentimentul pentru fiecare postare, chiar dacă modelele individuale ale ansamblului propus nu sunt întotdeauna precise în predicțiile lor.

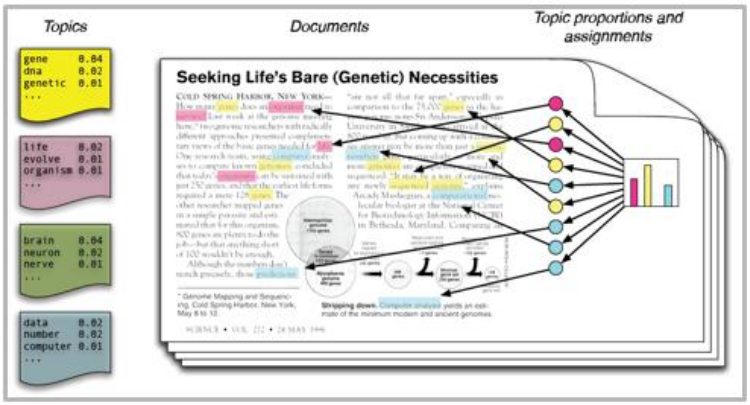
Ca și concluzie finală, soluția propusă de noi dă rezultate mai bune decat alte solutii curente cu care a fost comparată. Comparația este realizată folosind trei metrici: acuratețe, precizie și recall.

Model	Accuracy	Precision	Recall
RoBERTa-LSTM [65]	0.897	0.900	0.900
RoBERTa-GRU [66]	0.896	0.900	0.900
RoBERTa-BiLSTM [67]	0.896	0.900	0.900
Ensemble model (average) [67]	0.898	0.900	0.900
Ensemble model (majority) [67]	0.898	0.900	0.900
P2V + ELMo +BERT + FastText + AE + LSTM [63]	0.890	0.910	0.890
P2V + ELMo +BERT + FastText + AE + Transformer [63]	0.900	0.910	0.900
EDSA Ensemble SCT	**0.912**	**0.913**	**0.910**

Sub-Modulul Topic Modeling

Topic modeling este o tehnică folosită pentru a descoperi „temele” abstracte care apar într-o colecție de documente, texte sau orice altceva.
In cadrul proiectului, submodulul de Topic Modeling este folosit pentru analiza semantică latentă *(-l. Eng.* latent semantic analysis) și pentru extragerea subiectelor.

Extragerea subiectelor poate ajuta la identificarea grupurilor sau indivizilor care sunt frecvent vizați de discursul de incitare la ură, și poate ajuta modelul de Analiză și Detecșie să înțeleagă tendințele discursurilor de dezinformare.



Analiza semantică latentă rezolvă problema sinonimelor prin identificarea termenilor care statistic apar împreună și presupune folosirea unele structuri semantice latente în date care sunt parțial ascunse prin caracterul aleatoriu al alegerii cuvântului. Prin identificarea unor termeni sintactici diferiți, dar asemănători din punct de vedere semantic, folosind o structură numită spațiu „concept” ascuns, se descopera asemănarea dintre diferite articole, ceea ce ajuta la analiza asemănării dintre știrile reale și cele false și a veridicității știrilor.

Noi modele de detecție a știrilor false

In raportul anterior, am prezentat o serie de modele si rezultatele parțiale obținute de acestea in cazul detecției. Pentru aceasta etapa, am propus noi modele de tip Deep Learning.

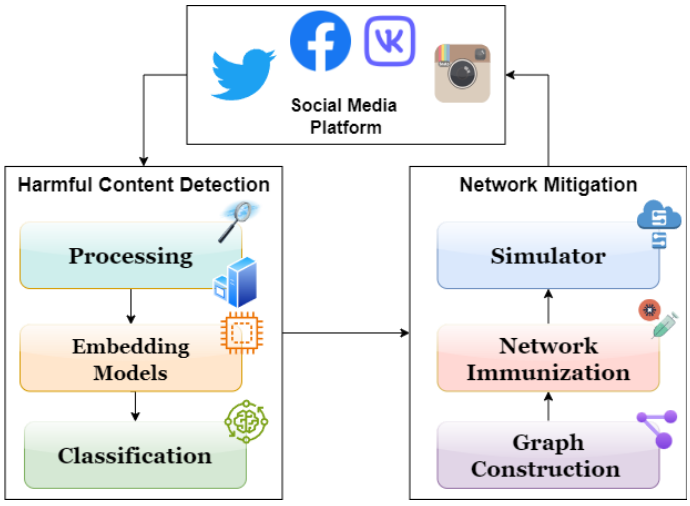


Diagrama de funcționare este prezentată în figura de mai sus.
Au fost folosite și testate mai multe modele de Învățare Automată.

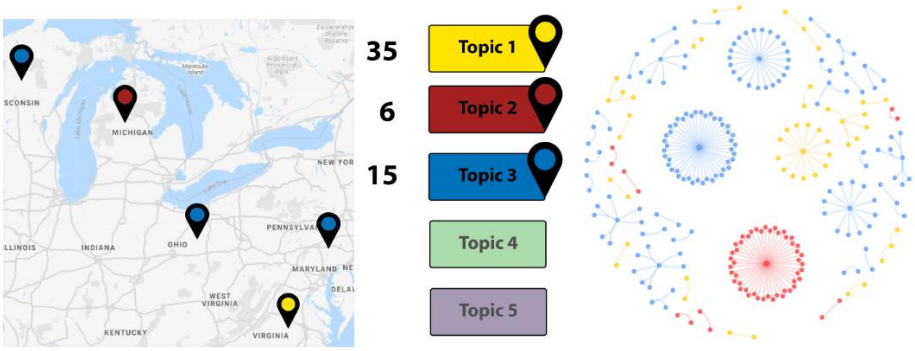
Model	Embedding	Accuracy	Precision	Recall	F1
BiLSTM-Dense	Word2Vec	0.9053	0.9228	**0.9333**	0.9280
BiLSTM-Dense	GloVe	**0.9122**	0.9406	0.9242	**0.9323**
BiLSTM-Dense	BERT	0.8270	0.8792	0.8883	0.8837
BiLSTM-Dense	RoBERTa	0.8255	0.9206	0.8026	0.8576
BiLSTM-CNN-Dense	Word2Vec	0.9107	0.9380	0.9220	0.9315
BiLSTM-CNN-Dense	GloVe	0.9050	0.9370	0.9143	0.9250
BiLSTM-CNN-Dense	BERT	0.8254	0.8780	0.8850	0.8885
BiLSTM-CNN-Dense	RoBERTa	0.8230	0.9180	0.8000	0.8550
BiGRU-GMP	Word2Vec	0.8955	0.9252	0.8952	0.9138
BiGRU-GMP	GloVe	0.8970	0.9012	0.9056	0.9145
BiGRU-GMP	BERT	0.8325	0.8359	0.9257	0.8785
BiGRU-GMP	RoBERTa	0.8671	0.8844	0.9166	0.9002
BiLSTM-GMP	Word2Vec	0.8981	0.9551	0.8860	0.9192
BiLSTM-GMP	GloVe	0.8671	0.8844	0.9166	0.9002
BiLSTM-GMP	BERT	0.8470	0.8792	0.8883	0.8837
BiLSTM-GMP	RoBERTa	0.8683	0.8789	0.9264	0.9020
BiGRU-CNN-GMP	Word2Vec	0.8922	0.8982	0.9021	0.9103
BiGRU-CNN-GMP	GloVe	0.8850	0.8950	0.9074	0.9100
BiGRU-CNN-GMP	BERT	0.8300	0.8370	0.9205	0.8750
BiGRU-CNN-GMP	RoBERTa	0.8650	0.8814	0.9150	0.8975
2BiLSTM	Word2Vec	0.8953	**0.9690**	0.8584	0.9148
2BiLSTM	GloVe	0.9053	0.9228	**0.9333**	0.9280
2BiLSTM	BERT	0.8505	0.8822	0.8904	0.8863
2BiLSTM	RoBERTa	0.8671	0.8873	0.9128	0.8999
2BiLSTM	RoBERTa	0.8671	0.8873	0.9128	0.8999
Logistic Regression [4]	TFIDF	N/A	0.91	0.90	0.90
Ensemble [31]	Multiple	N/A	0.76	0.83	0.793
BiGRU-Attention [31]	Glove	N/A	0.77	0.82	0.790

Cea mai bună performanță este obținută folosind modelul BiLSTM-GMP cu GloVe embedding. Această arhitectură înlocuiește primul strat ascuns Dens al arhitecturii BiLSTM cu un nivel de GlobalMaxPooling (GMP). GMP oferă o complexitate moderată și necesită un număr rezonabil de parametri. Acest lucru ajută la evitarea a două probleme comune, și anume overfitting (soluția ar avea performanțe slabe pe date necunoscute) și vanishing gradients (soluția ar avea dificultăți în a învăța din straturile mai profunde).

Interfața cu Utilizatorul: Conștientizarea dezinformării

In cadrul proiectului, am propus si o platforma pentru Conștientizarea dezinformării. Conștientizarea dezinformării este esențială din mai multe motive: protecția indivizilor, menținerea armonie sociale, consolidarea proceselor democratice, promovarea gândirii critice Prin informare și discernământ, putem combate colectiv răspândirea dezinformării și crea un mediu online mai pozitiv și incluziv.

Soluția propusă de noi permite utilizatorilor sa isi defineasca cuvintele cheie de interes, ex. Covid, Alegeri. Sistemul va cauta postari ce contin aceste cuvinte cheie, va specifica daca este de incredere sau nu. Daca postul este de tip Fake News va specifica clasa: propaganda, click-bait, dezinformare, etc. Sistemul va aplica tehnica de Topic modeling pe posturile colectare pentru a detecta subiectele acestor postari. Intr-o ultima etapa, se vor afisa utilizatorului nu doar postarile gasite cat si locatia geografica de unde ele au fost trimise in reteau sociala.



(a) Google map view (b) Subscriptions (c) Communities per topic

2. Activități de publicare

În cadrul proiectului SCAN-NEWS au fost publicate, cu afilierea AOȘR 6 articole. Trei dintre cele 6 articole este de jurnal și sunt cotat Q1 ISI conform Journal Citation Reports (JCR) pe 2023. Progresul actual depășește programul stabilit în propunerea de proiect pentru primul an (Figură 1. Diagrama GANTT). Dintre cele șase, trei articole sunt raportate în decembrie 2024, restul au fost raportate în prima perioadă a proiectului.

Lista publicațiilor pentru proiectul SCAN-NEWS este redată mai jos.

Raportate în Decembrie 2024

- Alexandru Petrescu, Ciprian-Octavian Truică, **Elena-Simona Apostol**, Adrian Paschke. *EDSA-Ensemble: an Event Detection Sentiment Analysis Ensemble Architecture*. IEEE Transactions on Affective Computing, August 2024, DOI: 10.1109/TAFFC.2024.3434355 (Jurnal cotat ISI Q1, F.I. = 9.6 conform JCR 2023)
- **Elena-Simona Apostol**, Ciprian-Octavian Truică, Adrian Paschke. *ContCommRTD: A Distributed Content-based Misinformation-aware Community Detection System for Real-Time Disaster Reporting*. Transactions on Knowledge and Data Engineering Journal, November 2024. DOI: 10.1109/TKDE.2024.3417232 (Jurnal cotat ISI Q1, F.I. = 8.9 conform JCR 2023)
- Ciprian-Octavian Truică, Ana-Teodora Constantinescu, **Elena-Simona Apostol**. *StopHC: A Harmful Content Detection and Mitigation Architecture for Social Media Platforms*. IEEE International Conference on Intelligent Computer Communication and Processing (ICCP 2024), Octombrie 2024 (Conferință ISI cotată Național în CORE)

Raportate în Iulie 2024

- **Elena-Simona Apostol**, Özgur Coban, Ciprian-Octavian Truică. *CONTAIN: A community-based algorithm for network immunization*. Engineering Science and Technology, an International Journal. Elsevier. 55:1-10(101728), ISSN 2215-0986, Iulie 2024. DOI: 10.1016/j.jestch.2024.101728 (Jurnal cotat ISI Q1, F.I. = 5.1 conform JCR 2023)
- **Elena-Simona Apostol**, Adrian-Cosmin Cojocaru, Ciprian-Octavian Truică. *Large-Scale Graphs Community Detection using Spark GraphFrames.* The 23rd International Symposium on Parallel and Distributed Computing (ISPDC 2024). Iulie 2024 (Conferință ISI cotată C în CORE)
- Alexandru Petrescu, Ciprian-Octavian Truică, **Elena-Simona Apostol**. *Language-based Mixture of Transformers for EXIST2024*. Conference and Labs of the Evaluation Forum (CLEF2024). Publicată în Lecture Notes in Computer Science (LNCS). Septembrie 2024 (Conferință ISI)

Data: 05.12.2024

Semnătură director,
Conf. Dr. Ing. Habil. Apostol Elena-Simona