A GÂNDI, A ROSTI, A SĂVÂRȘI URA (HATE THOUGHTS, HATE SPEECH, HATE ACTIONS)

Andrei VLĂDESCU*

Abstract

Hate speech communications, whether expressed through text, image or sound, convey a two-way message: one to the target and the social group to which it belongs, as a factor of their minimization and dehumanization, and the second to other supporters of extremist-aggressive ideologies, to reinforce the sense of belonging to the group and to a common vision.

An aggressive discourse is all the more dangerous as it is stated by a charismatic speaker, who addresses a public marked by frustration, in a social or historical context favorable to violence and who benefits from an influential mean of dissemination which is unique or the main source of information for the target audience.

Aggressive online speech is an eloquent example of how advanced technologies bring both opportunities and challenges regarding the need for a balance between fundamental rights and principles, as is the case between freedom of expression and the defense of human dignity. The connection of aggressive online discourse with the real world is difficult to demonstrate through systematic empirical evidence, being less preferred the study of the dynamics underlying the transfer of energy from words to overt discrimination, hostile behavior and violent.

The purpose of this study is to identify the social groups most vulnerable to aggressive public discourse, as a primary stage in constructing an assessment tool which, based on the vocabulary or the symbolisms used, could determine the affinity of an individual or a group of persons for radical ideologies close to actionable hate.

Keywords: *aggressive speech, discrimination, vulnerable social group, aggressive public discourse, intolerance.*

*Doctorand în Ştiinţele Comunicării al Şcolii Naţionale de Studii Politice şi Administrative, andreivladescu@yahoo.com.

Introducere

Conceptul de *hate speech* este folosit de multe ori ca termen generic, în care se amestecă ameninţări concrete la adresa unor persoane sau grupuri sociale cu situaţii în care oamenii pur şi simplu se eliberează de frustrări acumulate ca urmare a unor neîmpliniri sau eşecuri individuale. Agresivitatea discursivă este element al unui sistem complex, din care mai fac parte libertatea de exprimare, drepturile individuale şi de grup, dar şi concepte precum demnitate, libertate şi egalitate, motiv pentru care definirea sa este dificilă şi deseori contestată.

O idee din categoria *hate speech*, indiferent dacă este exprimată prin text, imagine sau sunete, transmite mai multe mesaje: unul este direcţionat către individul vizat (şi, prin extrapolare, către grupul social în care avesta este inclus), ca parte a minimalizării şi dezumanizării acestuia, iar al doilea către ceilalţi promotori ai unor ideologii similare, pentru ca aceştia să nu se simtă izolaţi, să fie conştienţi că nu sunt singuri şi astfel să se consolideze sentimentul de apartenenţă la un grup (Waldron, 2012).

Altfel spus, pe lângă dezumanizarea victimelor, discursul agresiv are rolul de a augmenta identitatea de grup a surselor acestuia, prin crearea unei relaţii bazate pe antagonismul „noi" versus „ei" (Gelber 2011; Heinze 2017; Butler 1997). Waldron (2012) susţine că unul dintre efectele *hate speech* este crearea unei ameninţări faţă de pacea socială, „un fel de otravă cu acţiune lentă, care se acumulează ici şi colo, cuvânt cu cuvânt, astfel încât, în cele din urmă, devine mai greu şi mai puţin firesc chiar pentru membrii buni ai societăţii să îşi joace rolul în menţinerea acestui bun public".

Potrivit lui Susan Benesch, director executiv al proiectului Dangereous Speech (https://dangerousspeech.org), conceptul de *hate speech* este o categorie prea vastă ca să permită sesizarea acelor acte discursive care ar putea fi utilizate ca indicatori de avertizare timpurie pentru violenţă offline. Conform acesteia, „discursul agresiv este comun multor societăţi, inclusiv celor unde genocidul are o probabilitate mică de a se întâmpla; pe de altă parte, deşi discursul agresiv nu creşte în mod considerabil riscul de violenţă în masă, poate provoca daune emoţionale şi psihice grave" (Benesch, 2012). Din această perspectivă,

autoarea a propus cinci criterii care ar trebui îndeplinite pentru evaluarea gradului de pericol al comunicărilor cu conținut agresiv (Benesch, 2012):

- un vorbitor charismatic, cu grad ridicat de influență asupra audienței;
- un public marcat de frustrări pe care vorbitorul le poate specula;
- un act discursiv perceput clar ca apel la violență;
- un context social sau istoric favorabil violenței, indiferent de motiv sau pretext, pe fondul unor competiții pentru resurse, a unei lipse a eforturilor de soluționare a nemulțumirilor sau a unor episoade violente anterioare;
- un mijloc de diseminare influent prin sine, prin simplul fapt că este unica sau principală sursă de informare pentru audiența relevantă.

Detectarea unei comunicări de tip *hate speech* este o sarcină dificilă, pentru că, în primul rând, există dezacorduri în ceea ce privește modul în care este definit conceptul, iar pe de alta pentru că un anumit conținut comunicațional poate fi considerat agresiv pentru unii și nonagresiv pentru alții. De exemplu, potrivit unor autori, *hate speech* reprezintă acea comunicare scrisă sau verbală care „discreditează în orice mod o persoană sau un grup bazat pe caracteristici precum rasă, culoare, etnie, sex, orientare sexuală, naționalitate sau religie" (Warner și Hirshberg, 2012) sau se poate considera că este *hate speech* dacă vizează „doar grupuri sociale defavorizate și le aduce prejudicii atât direct, cât și indirect" (Waldron, 2012).

Comparativ, Parekh (2012) este extrem de precis atunci când propune trei caracteristici definitorii ale discursului agresiv, o comunicare urmând a fi considerată *hate speech* dacă:

- „este îndreptată către o persoană specificată sau ușor de identificat sau către un grup de indivizi bazat pe o caracteristică arbitrară și irelevantă normativ";
- „stigmatizează grupul-țintă, atribuindu-i implicit sau explicit calități considerate pe larg ca extrem de nedorite";
- „grupul-țintă este privit ca o prezență nedorită și ca obiect legitim de ostilitate".

Din perspectiva unor autori, conceptul de *hate speech* se referă de cele mai multe ori la expresii care susţin incitarea la daune, prejudicii sau vătămare bazată pe identificarea ţintei agresivităţii cu un anume grup social, iar în unele situaţii, conceptul este extrapolat la expresii care facilitează apariţia şi întreţinerea unui climat favorabil prejudecăţilor şi intoleranţei care conduc la discriminare, ostilitate şi violenţă (Gagliardone et al., 2015).

Alte definiţii interpretează discursul de tip *hate speech* ca „limbaj folosit pentru a exprima ura faţă de un grup sau este destinat să fie peiorativ, să umilească sau să insulte membrii grupului" (Davidson et al., 2017), „atac deliberat îndreptat către un anumit grup de oameni şi motivat de aspecte ale identităţii grupului" (de Gilbert et al. 2018) sau „limbaj care atacă sau diminuează, care incită la violenţă sau ură împotriva grupurilor, bazată pe caracteristici specifice, cum ar fi aspectul fizic, religia, descendenţă, originea naţională sau etnică, orientarea sexuală, identitate de gen sau altele, şi poate aborda diferite stiluri lingvistice, chiar şi în forme subtile sau când se foloseşte umorul". (Fortuna et al., 2018)

Din această ultimă perspectivă, ne punem o întrebare legitimă: oare putem considera că „Noaptea ca hoţii" intră sau nu în categoria *hate speech*,? Altfel spus, prezenţa conţinutului mundan în sine ar trebui sau nu să echivaleze cu *hate speech*? Sau putem considera că lumescul nu vizează neapărat stigmatizarea unui individ şi poate fi utilizat în scopuri stilistice sau sublinierea unei idei, iar discursul agresiv poate denigra sau ameninţa o persoană sau un grup de oameni fără să recurgă la utilizarea vreunui termen sau sintagmă dificil de reprodus. (Malmasi şi Zampieri, 2017)

Ce este atât de special la discursul agresiv online

Plasat în punctul de tangenţă al mai multor tipuri de tensiuni, discursul agresiv online constituie un exemplu grăitor al modului în care tehnologiile avansate aduc deopotrivă oportunităţi şi provocări în ceea ce priveşte necesitatea unui echilibru între drepturile fundamentale şi principii, cum este cazul celui între libertatea de exprimare şi apărarea demnităţii umane.

Caracterul discursului agresiv online și conexiunea sa cu lumea reală sunt teme larg dezbătute și dificil de demonstrat prin dovezi empirice sistematice, fiind preferată varianta punerii accentului pe modul de soluționare, pe studierea efectului în detrimentul cauzei și mai puțin a dinamicii care stă la baza transferului energiei dinspre cuvinte spre discriminarea manifestă, comportamente ostile și chiar apariția unor acțiuni violente. Discursul agresiv poate rămâne în mediul virtual pe o perioadă nedeterminată, în diferite formate și pe mai multe platforme, putând fi astfel replicat într-o manieră repetată și care îi atribuie cun caracter practic itinerant.

Vom aminti aici un adevăr care are deja valențe de truism, anume că oamenii sunt mai agresivi și chiar brutal de sinceri, mergând pâna la impolitețe, atunci când comunică online, deoarece sunt perfect conștienți că sunt anonimi (cu un minim de efort intelectual și cu minime cunoștințe tehnologice) și deci pot acționa după bunul plac, fără consecințe imediate, fie acestea sancțiuni sociale, expunere la un atac fizic sau față de aplicarea legii. De altfel, în articolul "*The Online Disinhibition Effect*" (2004), psihologul John Suller a menționat un set de șase factori care ar putea explica modificarea semnificativă a comportamentului oamenilor atunci când se manifestă online:

- anonimatul disociativ („acțiunile mele nu pot fi atribuite persoanei mele");
- invizibilitatea („nimeni nu poate spune cum arăt și nici nu îmi poate judeca tonul");
- asincronia („acțiunile mele nu se produc în timp real");
- introiecția („nu îi văd pe acești oameni, trebuie să ghicesc cine sunt și care este intenția lor");
- imaginația disociativă („aceasta nu este lumea reală, aceștia nu sunt oameni reali");
- minimizarea autorității („nu există autoritate aici, deci pot acționa liber").

Potrivit HateBase, aplicație web care stochează cazuri de *hate speech* online la nivel internațional (http://www.hatebase.org), deși cele mai multe cazuri de agresivitate din mediul virtual au ca ținte persoane alese în funcție de etnie și naționalitate, incitarea la ură focalizată pe religie și categorii sociale este, de asemenea, în creștere.

În acest punct, se cuvine să arătăm că dificultatea identificării agresivităţii discursive este accentuată de faptul că cele mai importante platforme de socializare tratează uşor diferit ideea de *hate speech*:

- Facebook: „Definim *hate speech* un atac direct asupra oamenilor, pe baza a ceea ce noi numim caracteristici protejate – rasă, etnie, origine naţională, apartenenţă religioasă, orientare sexuală, castă, sex, gen, identitatea de gen, boală sau dizabilitatea gravă. De asemenea, oferim unele protecţii pentru statutul de imigraţie. Definim atacul ca fiind acel discurs violent sau dezumanizant, declaraţiile vizând inferioritatea sau apelurile la excludere sau segregare”[1];
- Twitter: „Comportamentul marcat de ură: Nu puteţi promova violenţa împotriva sau atacul direct sau ameninţările la adresa altor persoane pe motive de rasă, etnie, origine naţională, orientare sexuală, gen, identitate de gen, apartenenţă religioasă, vârstă, dizabilitate sau boli grave”.[2]

Prezenţa *hate speech* în România

Pentru identificarea grupurilor sociale menţionate drept **cele mai vulnerabile** faţă de discriminare şi intoleranţă în România studiul de faţă a analizat comparativ un set de rapoarte referitoare la acest aspect, realizate în perioada 1999-2018: An Overview on Hate Crime ond Hate Speech (Centre for Legal Resources Romania şi Europe Monitoring and Reporting Hate Speech/EMORE - 2017), „Raportul anual cu privire la discursul intolerant şi instigator la ură din România” (ActiveWatch – 2015, 2016, 2017, 2018), „Raport asupra rasismului şi intoleranţei în România” (Comisia Europeană împotriva Rasismului şi Intoleranţei/ECRI – 1999, 2002, 2006, 2014, 2019), “Racism and Related Discriminatory Practices in Romania” (European Network Against Racism/ENAR – 2011, 2012) şi “No Into Hate” (Institutul Român pentru Evaluare şi Strategie/IRES - 2018)

---

[1] Sursa: https://www.facebook.com/communitystandards/objectionable_content.
[2] Sursa: https://help.twitter.com/en/rules-and-policies/hateful-conductpolicy.

În urma extragerii grupurilor sociale menționate în fiecare raport ca fiind vulnerabile față de discriminare/rasim și, pe cale de consecință, față de atitudini agresive, inclusiv de tip discursiv, au fost obținute următoarele rezultate:

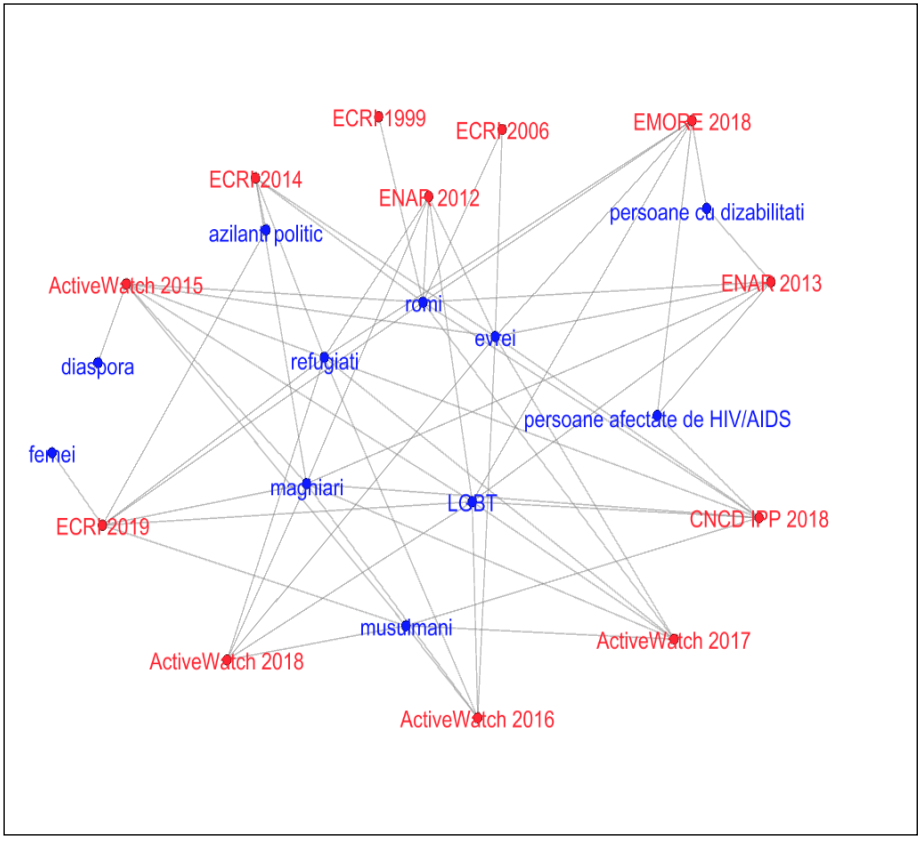

Figura 1: Centralizarea grupurilor sociale vulnerabile față de discriminare și intoleranță

Motivul pentru care aceste categorii sociale figurează constant între „preferințele" autorilor de opinii agresive ar putea fi faptul că sunt „altfel", comparativ cu majoritatea, fie din perspectivă etnică/ religioasă, fie din cea sexuală, fiind astfel dificil pentru un român „autentic" (creștin ortodox, heterosexual, naționalist) să se raporteze de o manieră tolerantă la acestea.

O altă posibilă explicaţie se bazează pe rezultatele sondajului *No Into Hate*, realizat în 2018 de Institutul pentru Politici Publice la solicitarea Consiliului Naţional pentru Combaterea Discriminării, care arată că în ţara noastră există un „nivel ridicat de neîncredere în homosexuali (74%), romi (72%), imigranţi (69%), musulmani (68%), persoane cu HIV/SIDA (58%), persoane de altă religie (58%), maghiari (53%) şi evrei (46%)". (IRES – 2018) În acelaşi timp, studiul denotă existenţa în România a unei intoleranţe semnificative faţă de homosexuali (59% nu ar accepta să le fie rude, iar 52% nu i-ar accepta ca prieteni), imigranţi (39%, respectiv 30%), musulmani (39%, respectiv 28%), romii (29%, respectiv 18%), maghiarii (24%, respectiv 15%) şi evrei (21%, respectiv 14%).

De asemenea, peste 2 treimi dintre români (72%) nu agreează căsătoria între membri ai comunităţii LGBT, în timp ce 60% nu sunt de acord cu parteneriatul civil, peste „60% din români îi consideră pe musulmani ca fiind un pericol potenţial, iar 52% cred că imigranţii ar trebui opriţi la graniţele Europei" (IRES – 2018). Nu în ultimul rând, atrage atenţia faptul că naţionalismul este considerat de aproape jumătate dintre români drept o mişcare necesară pentru România, iar neîncrederea în străinii care critică România este afirmată de 50% dintre cei chestionaţi.

Încrederea în celălalt influenţează determinant modul în care românii se raportează la alte etnii, semnificativ în acest sens fiind un sondaj realizat de Compania de Cercetare Sociologică şi Branding pentru Asociaţia Pro Democraţia în anul 2011 (Stereotipuri la adresa romilor. Perceptia privind etnia romilor. Rolul presei. Identificarea de alternative de îmbunătăţire), potrivit căruia nivelul de încredere în romi (19%), evrei (30%) şi maghiari (33%) era mult inferior celui cu care erau investiţi români (65%) şi germanii (52%).

Un alt aspect de interes a rezultat din Rapoartele anuale cu privire la discursul intolerant şi instigator la ură identificat în spaţiul public din România, realizate de organizaţia nonguvernamentală ActiveWatch în perioada 2014-2018, anume acela că cel mai frecvent discursul agresiv public a fost direcţionat către **comunitatea LGBT**[3]

[3] Acronim pentru Lesbian, Gay, Bisexual, Transexual, recent devenit LGBTT - Lesbian, Gay, Bisexual, Transexual, Transgender.

(32% din cazuri), urmată de **maghiari** (23%), **romi**, **evrei**, **musulmani şi populaţia de culoare** (fiecare cu 8%), **refugiaţi** (7%), la distanţă mare şi practic marginale ca ţinte ale discursurilor agresive fiind femeile şi diaspora / cetăţeni români relocaţi temporar în alte state membre UE în căutare de locuri de muncă (fiecare cu 3%).

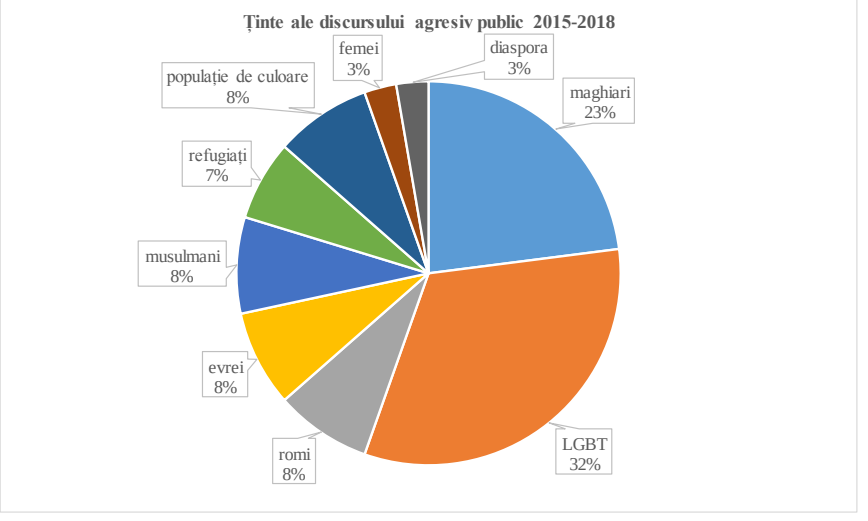

Figura 2: Ţinte ale discursului agresiv public (2015-2018)

Din Rapoartele ActiveWatch rezultă un detaliu semnificativ, referitor la autorii afirmaţiilor agresive, care pot fi încadraţi astfel:

- persoane implicate în viaţa politică (în special când participă la campania electorală);
- membrii unor organizaţii extremiste (chiar dacă în acest moment în România nu există un actor politic de extremă dreapta care să genereze un impact semnificativ în societate);
- unii reprezentanţi ai mass media;
- o serie de persoane publice din diferite domenii ale vieţii sociale din România;
- elemente „ultras" (persoane din componenţa galeriilor care susţin diferite echipe sportive);

- foşti sau actuali reprezentanţi ai autorităţilor / instituţiilor naţionale (atrage atenţia faptul că declaraţiile formulate de oficialităţi au un potenţial ridicat de legitimare a discriminării şi, pe cale de consecinţă, de stimulare a agresivităţii, inclusiv în mediul virtual, aceasta apărând întrucâtva ca justificată oficial);
- autori necunoscuţi (sau, mai degrabă, neidentificabili, aceştia fiind parte a unor grupuri informale care au exprimat colectiv diferite opinii agresive).

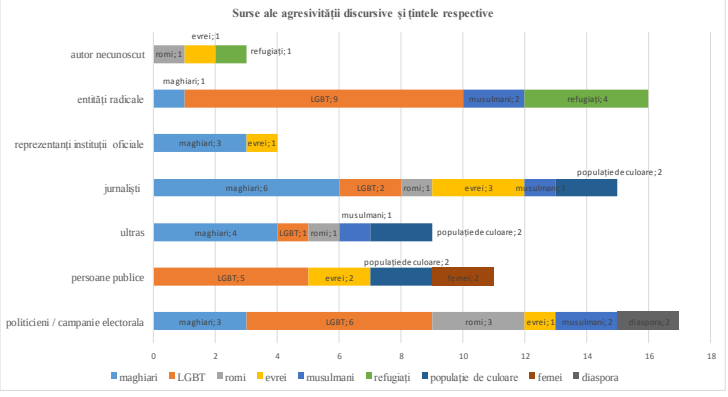

Figura 3: Surse ale agresivităţii discursive şi ţintele respective

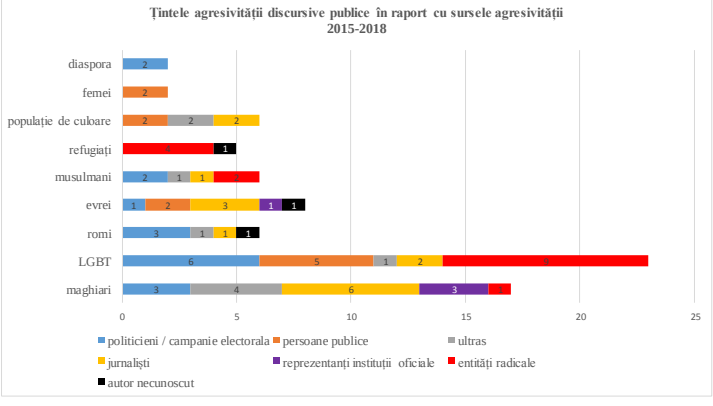

Figura 4: Ţintele agresivităţii discursive în raport cu sursele agresivităţii (2015-2018)

Corespondențele între sursele discursului sgresiv public și țintele vizate de acestea sunt reprezentate în graficul alăturat, din care rezultă, pe de o parte, că există surse preocupate de mai multe ținte, iar pe de alta, că unele ținte sunt în „atenția" mai multor surse.

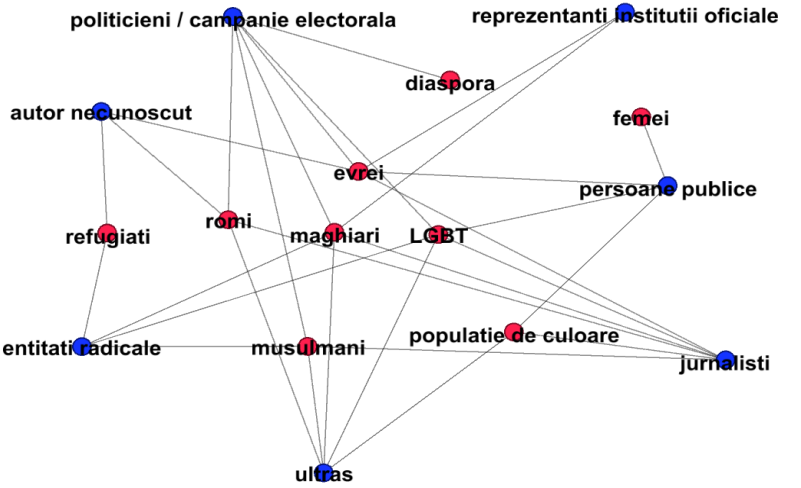

Figura 5: Sursele discursului agresiv public din România și țintele vizate de acestea

Concluzii

Atrage atenția faptul că grupurile sociale ale comunităților evreiască, maghiară și LGBT nu sunt doar cele mai vizate ținte (sub aspect cantitativ), dar și țintele de care sunt interesate mai multe categorii de agresori, respectiv 5 surse. În plan secund se plasează grupurile sociale ale comunităţilor romă și musulmană, vizate de cîte 4 categorii de surse.

Din această perspectivă, o cercetare viitoare ar trebuie să se axeze pe studierea modului în care sunt referite fiecare dintre țintele principale, motivațiilor care sunt invocate de autorii agresiunilor discursive și a modului în care evoluția intensității *hate speech* este sau nu influențată de evoluțiile socio-economice din țara noastră.

Referințe:

1. ActiveWatch. „Raport anual privind discursul instigator la ură din România 2014—2015”. Preluat în 01.11.2019, din: http://activewatch.ro/ro/antidiscriminare/publicatii/s-a-lansat-primul-raport-anual-privind-discursul-instigator-la-ura-din-romania-2014-2015
2. ActiveWatch. „Raportul anual privind discursul instigator la ură din România, 2015—2016”. Preluat în 01.11.2019, din ActiveWatch website: http://activewatch.ro/ro/antidiscriminare/publicatii/s-a-lansat-raportul-anual-privind-discursul-instigator-la-ura-din-romania-2015-2016
3. ActiveWatch. „Raportul anual privind discursul instigator la ură din România, 2017”. Preluat în 01.11.2019, din: http://activewatch.ro/ro/antidiscriminare/publicatii/raportul-anual-privind-discursul-instigator-la-ura-din-romania-2017-lansare
4. ActiveWatch. „Raportul anual cu privire la discursul intolerant și instigator la ură, 2018”. Preluat în 01.11.2019, din: http://activewatch.ro/ro/antidiscriminare/publicatii/raportul-anual-cu-privire-la-discursul-intolerant-si-instigator-la-ura
5. Benesch, S., (2012). “Dangerous Speech: A Proposal to Prevent Group Violence”. *World Policy Institute*, 12 January. Disponibil la adresa https://worldpolicy.org/wp-content/uploads/2016/01/Dangerous-Speech-Guidelines-Benesch-January-2012.pdf, accesat la 01.11.2019.
6. Butler, J., (1997). “Excitable Speech: A Politics of the Performative”. London: Routledge.
7. Centre for Legal Resources Romania și Europe Monitoring and Reporting Hate Speech. An Overview on Hate Crime ond Hate Speech (2017), preluat în 01.11.2019 din http://www.crj.ro/wp-content/uploads/2018/04/eMore-country-profile-Romania-website.pdf
8. Davidson T, Warmsley D, Macy MW, Weber I., (2017). “Automated Hate speech Detection and the Problem of Offensive Language”. ICWSM;
9. de Gilbert O., Perez N, Garc’ia-Pablos A, Cuadros M. (2018), “Hate speech Dataset from a White Supremacy Forum”. In: 2nd Workshop on Abusive Language Online@EMNLP.
10. European Commission against Racism and Intolerance. (2019). “Country Report on Romania”. Preluat în 01.11.2019 din

https://rm.coe.int/first-report-on-romania-romanian-translation-/16808b5b8c
11. European Commission against Racism and Intolerance. (2014). "Country Report on Romania". Preluat în 01.11.2019 din https://rm.coe.int/second-report-on-romania-romanian-translation-/16808b5b8f
12. European Commission against Racism and Intolerance. (2006). "Country Report on Romania". Preluat în 01.11.2019 din https://rm.coe.int/third-report-on-romania-romanian-translation-/16808b5b92
13. European Commission against Racism and Intolerance. (2002). "Country Report on Romania". Preluat în 01.11.2019 din https://rm.coe.int/fifth-report-on-romania-romanian-translation-/168094c9e7
14. European Commission against Racism and Intolerance. (1999). "Country Report on Romania". Preluat în 01.11.2019 din https://rm.coe.int/fourth-report-on-romania-romanian-translation-/168094d770
15. Fortuna P., Nunes S. (2018). "A Survey on Automatic Detection of Hate speech in Text". *ACM Comput Surv*. 2018; 51(4):85:1–85:30.
16. Gagliardone, I., Gal, D., Alves, T., Martinez, G., (2015). „Countering Online Hate speech". UNESCO Series on Internet Freedoms. Paris: UNESCO Publishing. Disponibil la adresa http://unesdoc. unesco.org/images/0023/002332/233231e.pdf, accesat la 01.11.2019
17. Gelber, K., (2011). "Speech Matters: Getting Free Speech Right". Brisbane, Australia: University of Queensland Press.
18. Heinze, E., (2017). "Hate speech and Democratic Citizenship". Oxford: Oxford University Press.
19. Malmasi S., Zampieri M., (2017). "Detecting Hate speech in Social Media", RANLP 2017 – Recent Advances in Natural Language Processing Meet Deep Learning
20. Parekh, B. (2012). "Is There a Case For Banning Hate speech?" In M. Herz & P. Molnar (Eds.) *The Content and Context of Hate speech: Rethinking Regulation and Responses* (pp. 37-56). Cambridge: Cambridge University Press.
21. Suler, J., (2004). "The Online Disinhibition Effect", în *CyberPsychology & Behavior* 7(3):321-6, iulie 2004, disponibil la adresa https://www.researchgate.net/publication/ 8451443_The_ Online_Disinhibition_Effect, accesat la 01.11.2019

22. Waldron, J. (2012). "The Harm in Hate speech. Cambridge", Ma: Harvard University Press.
23. Warner, W., Hirschberg, J., (2012). "Detecting hate speech on the world wide web". In: Proceedings of the SecondWorkshop on Language in Social Media. pp. 19–26. Association for Computational Linguistics.
24. European Network Against Racism. (2011). "Racism and Related Discriminatory Practices in Romania". Preluat în 01.11.2019 din: https://www.enar-eu.org/IMG/pdf/22._romania.pdf
25. European Network Against Racism. (2012). "Racism and Related Discriminatory Practices in Romania". Preluat în 01.11.2019 din: https://www.enar-eu.org/IMG/pdf/romania.pdf
26. Institutul Român pentru Evaluare și Strategie. (2018). „Sondaj de opinie la nivel național privind nivelul discriminării în România și percepțiile actuale asupra infracţiunilor motivate de ură – NoIntoHate 2018". Preluat în 01.11.2019 din https://main.components.ro/uploads/1d3a0bf8b95391b825aa56853282d5da/2019/02/Sondaj_de_opinie_NoIntoHate_2018.pdf