# Marcarea sentimentelor în text: emoţiile autorului şi opinia cititorului

Victoria BOBICEV, Liviu CARCEA
*Technical University of Moldova*
victoria.bobicev@ia.utm.md, liviu.carcea@ia.utm.md

***Rezumat* — Analiza automată a sentimentelor este o direcţie de cercetare foarte activă în ultimii ani. Însă sentimentele sunt subiective şi greu de analizat automat. Textele sunt marcate manual şi apoi utilizate ca sursa pentru învăţarea automată. Însă marcarea manuală la fel are multe nuanţe legate de subiectivitate. În articol se discută două tipuri de marcare manuală şi acordul între adnotatori în ambele cazuri. În final se compară rezultatele analizei automate în baza textelor marcate manual.**

***Cuvinte cheie* — analiza sentimentelor, marcarea sentimentelor, interfaţa pentru marcare, acord între adnotatori, analiza automată a sentimentelor.**

## I. INTRODUCERE

În ultimii decenii internetul a crescut considerabil. Volumul textelor postate online creşte zilnic cît de la surse oficiale de informaţie atît şi de la utilizatori. Informaţia postată de utilizatori în calitate de comentarii, discuţii în bloguri şi forumuri, evaluări ale produselor, filmelor sau muzicii a atras atenţie sporită a cercetătorilor din sociologie, psihologie, marketing şi altele.

Analiza automată a sentimentelor este o direcţie de cercetare foarte activă în ultimii ani. Însă sentimentele sunt subiective şi greu de analizat automat. Textele sunt marcate manual şi apoi utilizate ca sursa pentru învăţarea automată. Însă marcarea manuală la fel are multe nuanţe legate de subiectivitate.

În articol se discută două tipuri de marcare manuală şi acordul între adnotatori în ambele cazuri. În final se compară rezultatele analizei automate în baza textelor marcate manual.

## II. LUCRĂRI PRECEDENTE

În multe cazuri este nevoie de marcare manuală a sentimentului mesajelor online. Cu scopul de a obţine o marcare relevantă şi cît mai obiectivă marcarea sentimentelor în texte se efectuiază de cîteva persoane independent una de alta şi apoi marcările se compară. Ca regulă, marcarea coincide pentru unele texte iar pentru altele nu. Astfel, este nevoie de estimat acordul între adnotatori. Pe parcursul anilor au fost propuse mai multe metrici de acord între adnotatori pentru evaluarea acordului.

Procentajul de acord este cea mai uşoară şi mai simplă măsură. Măsura este deseori criticată din cauza că nu face diferenţa între categorii, dar oferă totuşi o aproximare de bază a acordului între adnotatori.

Însă pe parcurs au fost propuse alte măsuri de acord, mai avansate. Unele din măsurile acestea sunt Cohen's kappa ($\kappa$), Fleiss' kappa ($K$), Krippendorf's alpha ($\alpha$) [6]. Aceşti coeficienţi sunt numiţi coeficienţi ce măsoară valoarea acordului ce depăşeşte acurd întîmplător în caz dacă adnotatorii aveau să marceze textele la întîmplare. Formula generală pentru aceşti coeficienţi este:

$$k = (Ao - Ae)/(1 - Ae) \qquad (1)$$

unde $Ao$ este acordul observat şi $Ae$ este acordul întâmplător în cazul în care adnotatorii aleg etichetele aleatoriu.

În cazul lui Cohen k, acordul aşteptat $Ae$ se calculează bazându-se pe faptul că alocarea aleatorie a categoriilor la elementele este guvernată de distribuţii anterioare care sunt unice pentru fiecare adnotator şi care sunt obţinuţi din distribuţia lor reală.

Alpha ($\alpha$) lui Krippendorf este un coeficient de acord pe baza ipotezei conform căreia acordul aşteptat este calculat prin analizarea distribuirii globale a marcărilor fără a lua în considerare adnotatorii care au produs aceste marcări. Această măsură poate fi utilizată în cazul marcărilor multiple şi permite lipse de valori.

Adnotarea subiectivităţii poate fi centrată pe percepţia cititorului [4] sau pe dispoziţia autorul unui text [5]. În lucrarea curentă, ne-am propus să comparăm aceste două tipuri de marcări.

Fleiss k este generalizarea la mai mult de doi adnotatori; acordul aşteptat se calculează pe baza presupunerii că alocarea aleatorie a categoriilor la articole, de orice adnotator, este guvernată de distribuirea categoriilor articolelor în lumea reală.

Analiza sentimentelor de ştiri este diferită de cea a altor tipuri de text. [7] au subliniat că există mai multe probleme specifice legate de acest tip de texte; una dintre ele fiind problema separarării ştirilor bune sau rele şi sentimentelor pozitive şi negative. Annotatorii tind să interpreteze greşit intenţia autorului şi să marcheze interpretarea lor ca fiind adevăratul sentiment al textului. În [5] a fost descris experimentul de adnotare a titlurilor de ştiri. Acordului între adnotatori chiar şi pentru aceste fragmente scurte de text, a fost relativ scăzut, sub 50%. Numai la a doua etapă de re-adnotare a aceloraşi citate, au reuşit să ajungă la un acord de 80%, creând un set de instrucţiuni de marcare detaliate cu exemple şi explicaţii multiple.

[3] au adaptat metodologia lui [5] în experimentul de adnotare şi adnotările făcute de primii doi autori ai lucrării

au ajuns la un acord de 76%.

[1] a afirmat, de asemenea, că în transcrierile de ştiri sau în articolele de ştiri sentimentul ataşat la o declaraţie poate fi mult mai puţin evident. Adnotarea a fost făcută folosind Amazon Mechanical Turk (https://www.mturk.com/) şi fiecare unitate de adnotare a obţinut trei adnotări independente. Lucrarea nu a raportat un acord de adnotare, deşi au menţionat că aproximativ 6% din exemple au fost aruncate din cauza dezacordului adnotatorilor.

În lucrările care descriu diferite tipuri de adnotări manuale [2], [5] au fost enumerate mai multe motive ale acordului slab intre adnotatori. Ambiguitatea unităţilor de adnotare şi subiectivitatea adnotatorilor au fost una dintre principalele probleme întâmpinate în acest proces. Un acord substanţial intre adnotări poate fi atins numai după mai multe iteraţii de adnotări de către aceiaşi adnotatori şi cu instrucţiuni de adnotare îmbunătăţite iterativ.

În [4], fiecare titlu de ştiri din corpus a fost adnotat de şase annotatori care nu primiseră nicio instrucţiune suplimentară şi, prin urmare, şi-au adnotat propriile sentimente provocate de text. Apoi au calculat adnotările medii ale titlurilor.

În [2], masteranzii au adnotat mesaje pe forum de sănătate cu şase emoţii de bază, iar calculul Fleiss Kappa [6] a fost relativ scăzut: 0,26. Ei au explicat acest dezacord prin variabilitatea dintre oameni şi specificul textelor corporale. Cu toate acestea, au continuat cu experimente de clasificare automată a sentimentelor care dau rezultate rezonabile (cel mai bun rezultat F = 0,65).

## III. MARCAREA SENTIMENTELOR ÎN FORUM

Primul experiment de marcare manuală a sentimentelor care descriem în acest articol este marcarea sentimentelor în forumuri dedicate problemelor sănătăţii. Problemele personale legate de sănătate sunt unele care preocupă majoritatea populaţiei. Analiza sistematică a 19 studii din 1999-2009 a evidenţiat mai multe motive pentru utilizarea forumurilor medicale cum ar fi căutarea informaţiilor pentru a afla despre aspectele psihologice, fizice şi sociale ale tratamentelor disponibile, evaluări ale tratamentelor alternative sau în căutarea unui sprijin emoţional prin comunicare anonimă, acces imediat şi constant la persoane cu aceleaşi probleme.

Am analizat sentimentele exprimate de participanţii la forumul medical de fertilizare in vitro (ivf.ca). Acest forum reuneşte femeile care utilizează tratamente de FIV cu speranţa de a concepe. Pentru analiza empirică, am selectat 1321 de posturi care au acoperit 80 de tematici legate de FIV (de exemplu, “Peste 40 de ani şi însărcinate“). Pe cînd în majoritatea experimentelor cu sentimente se utilizează clasificarea în sentimente pozitive şi negative, în cazul nostru polaritatea aceasta era nepotrivită şi am avit nevoie de alte etichete de sentimente.

Analiza preventivă a conţinutului discuţiilor pe forum a reafirmat faptul că majoritatea postărilor prezentau schimbul de experienţe personale, furnizarea de informaţii sau sfaturi, expresii de recunoştinţă / prietenie, chat, cereri de informaţii şi expresii de susţinere [8].

Astfel, începând cu mai multe sentimente posibile, am clasificat în final textele în *încurajare, gratitudine, confuzie, fapte + încurajare* şi *fapte*. Fiecare post în discuţiile selectate pentru marcare a fost marcat de două persoane. Nici una din persoanele acestea nu avea informaţie despre deciziile altor adnotatori. Textele în care adnotatorii nu au fost de acord cu o etichetă de clasă au fost consideraţi *ambigue*. Aceasta marcare este descrisă în [9] în mai multe detalii. Adnotatorii au lucrat cu textele date de două ori: prima etapă fiind analiza textelor din punct de vedere a tuturor sentimentelor posibile şi a doua etapă ataşarea unei etichete din cele stabilite la finalul primei etape. Respectiv, acordul obţinut este comparativ bun (*Fleiss Kappa* = 0.73). Astfel de accord este considerat înalt [6]. Statistica rezultatelor marcării este prezentată în tabelul 1.

TABEL I. PROCENTAJUL TEXTELOR AMBIGUE ÎN MARCAREA SENTIMENTELOR ÎN POSTĂRILE UTILIZATORILOR

| Categorie | Numarul de posturi | Procentajul |
|---|---|---|
| *Incurajare* | 206 | 27% |
| *Gratitudine* | 88 | 12% |
| *Confuzie* | 48 | 6% |
| *Fapte + Incurajare* | 73 | 10% |
| *Fapte* | 187 | 25% |
| ***Ambigue*** | **150** | **20%** |
| ***Total*** | 752 | 100% |

În [10] este descrisă continuarea experimentului de marcare a sentimentelor pe acelaşi forum cu acelaşi set de sentimente. O interfaţă web a fost creată pentru adnotatori care oferea acelaşi set de sentimente: *încurajare, gratitudine, confuzie* şi *fapte.*

Însă adnotatorilor a fost dată şi posibilitatea de a adăuga orice altă emoţie care ei observă în textele propuse spre analiză.

Ca rezultat, adnotatorii au adăugat o mulţime de sentimente ca: susţinere, îngrijorare, incertitudine, compasiune, speranţă, optimism, dispreţ, îngrijorare, tristeţe, bucurie, fericire, dezamăgire, frustrare.

În cazul forumurilor textele participanţilor conţin exprimările sentimentelor sale în abundenţă. Scopul autorilor textelor este de a împărtăşi emoţiile şi sentimentele sale. Astfel, începînd marcarea am optat pentru modelul bazat pe dispoziţia autorul textului şi am cerut adnotatorilor să analizeze sentimentul exprimat în text de autorul lui. Acest model este firesc pentru marcarea textului generat de utilizatori (User Generated Content). În majoritatea cazurilor utilizatorii nu ascund opiniile şi sentimentele sale ci le exprimă liber şi activ.

Textele marcate manual au fost utilizate în experimentele de calsificare automată. În primul pas au fost stabilite etichetele de sentimente prin alegerea sentimentului ataşat de cel puţin doi adnotatori. Apoi au fost efectuate experimentele de clasificare automată pe baza etichetelor date. Rezultatul obţinut a fost satisfăcător, F-measure în jur de 0,65.

## IV. MARCAREA SENTIMENTELOR ÎN NOUTĂȚI

Absolut altă situaţie este în cazul noutăţilor postate online de ziare şi agenţii profesionale. Cum a fost deja menţionat, în cazul acesta autorii textului îl prezintă în forma maximal neutră însă aşteptînd să trezească reacţie

emoţională în cititor.

A doilea experiment a fost dedicat marcării sentimentelor în noutăţi scris în limba rusă şi ucrainieană [12]. Textele pentru procesare au fost extrase de pe două surse de ştiri: ucrainene (https://tsn.ua/) şi ruseşti (http://censor.net.ua/). Iniţial corpusul ucrainean conţinea 5817 texte de ştiri, iar rus conţinea 10194 texte de ştiri. O mare parte a textelor a fost neutră deoarece textele conţineau ştiri, nu comentariile utilizatorilor. Acestea au fost difuzate în forma comparativ neutră. Chiar dacă atitudinea autorului a fost prezentă într-un text, ea a fost exprimată implicit, fără cuvinte ce exprimă direct sentimente. Astfel, am selectat textele care conţineau cel putin un cuvînt din lexiconul afectiv. Cu scopul acesta au fost utilizate lexicoanele descrise în [11]. După filtrarea au fost lăsate 2018 de texte ruse şi 2133 ucrainene pentru marcarea manuală.

Adnotarea a fost realizată de studenţii Institutului Politehnic Kharkiv prin interfaţa online. În cazul dat au fost propuse doar trei variante de etichete: pozitiv, negativ, neutru.

Aproximativ 40 de studenţi au participat la adnotări ale fiecărei părţi, ruşi şi ucraineni generând în 7248 adnotări pentru texte ruse şi 6733 adnotări pentru limba ucraineană. Întrebarea la care au răspuns adnotatorii a fost: Ce sentimente evocă acest text? De asemenea, aceştia aveau posibilitatea de a adăuga comentarii dacă considerău necesar. În final, fiecare text a fost adnotat de 2 până la 5 elevi folosind trei etichete: pozitiv, negativ, neutru. Numărul mediu de adnotatori pe text a fost de 3,59 pentru limba rusă şi de 3,2 pentru textele ucrainene. Acordul inter-adnotator calculat pe aceste texte a fost extrem de scăzut: Fleiss Kappa = 0,14 pentru textele ucrainene şi Fleiss Kappa = 0,24 pentru rusă.

TABEL II. PROCENTAJUL TEXTELOR AMBIGUE ÎN MARCAREA SENTIMENTELOR ÎN NOUTĂŢI

| Categorie | Procentajul de texte ambigue ucrainene | Procentajul de texte ambigue ruse |
|---|---|---|
| *Society* | 14% (63 din 436) | 16% (62 din 384) |
| *Politics* | 21% (103 din 487) | 13% (65 din 510) |
| *Incidents* | 12% (58 din 486) | 13% (30 din 239) |
| *Sport* | 14% (58 din 426) | 29% (92 din 413) |
| *Economics* | 31% (68 din 218) | 18% (43 din 245) |
| ***Total*** | 19% (***350 din 1849***) | 15% (***292 din 1994***) |

În cazul textelor ucraineni, doi adnotatori au selectat o etichetă şi un adnotator a selectat altă etichetă pentru 2/3 dintre texte. Chiar dacă în acest caz putem folosi eticheta selectată de doi adnotatori ca fiind cea corectă, distribuţia voturilor este de fapt identică cu distribuţia "din întâmplare", iar Fleiss Kappa pentru o astfel de adnotare este egală cu 0. În afară de aceasta, o mare parte din texte a fost marcată cu toate trei etichete de trei persoane. Aceste texte au fost absolut ambigue şi nu era posibil de definit eticheta lor. Statisticile privind rezultatele marcării textelor de ştiri sunt prezentate în tabelul II.

Textele absolut ambigue au fost eliminate din seturile utilizate în experimentele de clasificare automată iar textele care au fost marcate cu aceiaşi etichetă de două persoane au fost păstrate în set cu eticheta dată.

Experimentele de clasificare automată în trei clase : pozitiv, negativ şi neutru au demonsrat un rezultat comparativ bun : F-measure în jur de 0,73.

## V. PROBLEMELE MARCĂRII

În această secţiune, analizăm şi comparăm aceste două experimente şi rezultatele lor.

Datele comparative pentru aceste două experimente sunt prezentate în tabelul III.

TABEL III. COMPARAREA EXPERIMENTELOR DE MARCARE A SENTIMENTELOR

| Categorie | Primul experiment | Al doilea experiment |
|---|---|---|
| *Tipul textelor* | postările forumului despre problemele sănătăţii | textele noutăţilor de pe siturile de ştiri |
| *tipul marcării* | centrată pe dispoziţia autorului | centrată pe percepţia cititorului |
| *intrebarea către adnotatori* | ce sentimente exprimă autorul textului | ce sentiment evocă în cititor textul dat |
| *înţelegerea între adnotatori* | *Fleiss Kappa* = 0.73 pentru doi adnotatori şi *Fleiss Kappa* = 0.46 pentru trei | *Fleiss Kappa* = 0.14 pentru texte ucraineni şi *Fleiss Kappa* = 0.24 pentru ruse |
| *rezultatele experimentelor de clasificare automată* | F-measure în jur de 0,65 | F-measure în jur de 0,73 |

După cum observăm din tabelul III înţelegerea între adnotatori în cazul postărilor utilizatorilor este mai bună decît în cazul noutăţilor. Lucrul acesta poate fi explicat cu cîteva cause.

În cazul noutăţilor adnotatorii marcau sentimentele sale evocate de textul noutăţii şi acestea puteau să fie absolute opuse. Cum a fost menţionat şi în alte cercetări, oamenii reacţionează foarte diferit la aceiaşi informaţie. De exemplu, un anuunţ sportiv raportează despre un meci între două echipe de fotbal şi rezultatul lui în care prima echipă a cîştigat. Daca adnotatorul este fanatul echipei date, acesta marchează textul ca pozitiv. În caz opus, dacă adnotatorul este fanatul echipei care a pierdut acesta marchează textul ca negativ şi dacă adnotatorul nu este fanatul fotbalului şi nici nu cunoaşte echipele menţionate, acesta marchează textul ca neutru; textul într-adevăr fiind doar constataţia faptului fără orice discuţie pe margina lui. La fel se întîmpla şi cu alte tipuri de noutăţi; adnotatorii marcau sentimentele sale care erau diferite în dependenţă de poziţiile sale faţă de faptele descrise în text. De exemplu, în politică ca şi în sport fiecare are părerea sa ce este corect şi ce nu şi reacţionează destul de emoţionant la noutăţile politice.

În cazul marcării posturilor adnotatorilor scopul de bază a adnotatorilor a fost definit altfel. Lor li se cerea de indicat sentimentele autorilor textelor care erau exprimate destul de deschis, ca de exemplu : “Mă gândesc că, dacă sunt aşa stresat, oricum, nu este un moment bun“ sau “Ma bucur pentru tine! Arăţi mai relaxat“. Astfel, pentru adnotatori este mai simplu de indicat acelaşi sentiment şi de ajuns la o înţelegere mai bună. Putem constata că adnotatorii

identificau sentimentele autorului în baza textului scris destul de uniform. Una din problemele întîmpinate este faptul că textile date sunt despre problemele sănătăţii şi în cazuri cînd autorul textului descria problemele adnotatorii au dificultăţi cu marcare. Dintr-un punct de vedere acestea sunt descrieri cu aspect negativ dar din alt punct de vedere autorul textului poate reda informaţia dată în forma neutră, doar ca o descriere pentru alţii însoţînd cu informaţii despre tratament sau medicamente utilizate.

Însă, trebuie de spus că experimentele de detectare automată a etichetei au demonstrat că rezultatele mai bune au fost obţinute pentru textele de noutăţi decît pentru postările forumurilor. Aici la fel pot fi propuse cîteva explicaţii posibile.

În primul rînd, trebuie de reţinut că în cazul forumurilor a fost efectuată clasificarea în patru clase, ce complică sarcina şi duce la rezultate mai slabe decît în cazul clasificării doar în sentiment pozitiv şi negativ.

A doua cauză poate fi mărimea textelor. În cazul noutăţilor textile au fost comparative mici (100 – 150 cuvinte) şi, ce este esenţial, conţineau informaţie doar pe un subiect. Postările forumurilor au fost destul de lungi (150 – 300 cuvinte), mai lungi decît sunt postările obişnuite ale utilizatorilor şi în cazul postărilor mai lungi conţineau cîteva părţi cu tonalitatea diferită. Un post putea să se înceapă cu cuvinte de mulţumire interlocutorului precedent pentru susţinere, apoi să conţină descrierea stării autorului care putea să aibă tonalitatea negativă şi să termine cu enunţuri de speranţă şi urări de bine pentru alţi inerlocutori. Astfel, textul în sine deja putea să fie ambiguu şi să creeze dificultăţi în procesul clasificării.

Astfel, pentru fiecare sarcină de analiză a sentimentelor trebuie de luat în seamă mai multe aspecte : mărimea textelor, uniformitatea lor, tematica lor, numărul şi caracterul sentimentelor. Un aspect important este scopul cercetării care determină ce anume trebuie să fie evidenţiat în text analizat. Aici sunt posibile variante multiple : prezenţa sau absenţa subiectivităţii în text ; sentimentul autorului textului ; cauza sentimentului dat sau ţinta lui ; percepţia şi sentimentele cititorului şi elementele textului care le evocă.

## VI. CONCLUZIE

În lucrarea dată am comparat şi am analizat două experimente de marcare manuală a sentimentelor în textele online. În cadrul fiecărui experiment s-a realizat marcarea manual a sentimentelor, calculul acordului între adnotatori, selectarea etichetelor pentru textile date şi apoi experimentele de detectare automată a sentimentelor în baza marcării manuale.

Au fost evidenţiate şi discutate problemele şi dificultăţile marcării manuale şi comparate rezultatele clasificării automate.

## REFERINŢE


[1] Joseph G. Ellis, Brendan Jou, and Shih Fu Chang. 2014. Why We Watch the News: A Dataset for Exploring Sentiment in Broadcast Video News. In the Proceedings of the 16th International Conference on Multimodal Interaction (ICMI 2014), pages 104-111.

[2] Soumia Melzi, Amine Abdaoui, Jerome Azé, Sandra Bringay, Pascal Poncelet, and Florence Galtier. 2014. Patient's rationale: Patient Knowledge re-trieval from health forums. In the Proceedings of the Sixth International Conference on eHealth, Telemedicine, and Social Medicine, pages 140-145.

[3] Patrik F. Bakken, Terje A. Bratlie, Cristina Marco and Jon Atle Gulla. 2016. Political News Sentiment Analysis for Under-resourced Languages. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 2989–2996.

[4] Strapparava, C. and R. Mihalcea. Semeval-2007 task 14: Affective text. Proceedings of the 2008 ACM symposium on Applied computing, 2008.

[5] Balahur, A. and R. Steinberger. Rethinking Sentiment Analysis in the News: from Theory to Practice and back. Proceedings of the 1st Workshop on Opinion Mining and Sentiment Analysis, 2009.

[6] Ron Artstein, Massimo Poesio. 2008 Inter-coder agreement for computational linguistics. Computational Linguistics Journal Volume 34 Issue 4, pages 555-596. doi: 10.1162/coli.07-034-R2

[7] Alexandra Balahur, Ralf Steinberger, Mijail A. Kabadjov, Vanni Zavarella, Erik Van der Goot, Matina Halkia, Bruno Pouliquen, and Jenya Belyaeva. 2013. Sentiment Analysis in the News. In the Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'2010), pages 2216-2220.

[8] Malik S. and N. Coulson. Coping with infertili-ty online: an examination of self-help mechanisms in an online infertility support group. Patient Educ Couns, vol. 81, no. 2, pp. 315–318, 2010

[9] Marina Sokolova and Victoria Bobicev. 2013. What Sentiments Can Be Found in Medical Forums? In Galia Angelova, Kalina Bontcheva, Ruslan Mitkov: Recent Advances in Natural Language Processing, RANLP 2013, Bulgaria.

[10] Victoria Bobicev and Marina Sokolova. Inter-Annotator Agreement in Sentiment Analysis: Machine Learning Perspective. Recent Advances in Natural Language Processing, RANLP 2017, Bulgaria, 2017.

[11] Olessia Koltsova, Svetlana Alexeeva, and Sergei Koltcov. 2016. An Opinion Word Lexicon and a Training Dataset for Russian Sentiment Analysis of Social Media. In the Proceedings of the Interna-tional Conference “Dialogue 2016”, pages 277-287.

[12] Olga Kanishcheva and Victoria Bobicev. Good News vs. Bad News: What are they talking about? Recent Advances in Natural Language Processing, RANLP 2017, Bulgaria.