

## NATURAL LANGUAGE COMPLEXITY. SENTIMENT ANALYSIS

**Daniela GÎFU, Scientific Researcher III, "Alexandru Ioan Cuza" University of Iași  
ACS Andrei SCUTELNICU, Institute of Computer Science, Iași-Romanian Academy,  
"Alexandru Ioan Cuza" University of Iași**

*Abstract. The paper brings in discussion one of the most current issues in the research area of natural language processing, sentiment analysis. A sentiment analysis means to have the necessary competence to identify the nature of speaker's opinions about entities or/and specific needs (journalists, competitors, public opinion, services, products, etc.), knowing that the decision-making process is affected by the received message. We are witnessing an explosion of "sentiments" about matters of public interest subjects, available, especially, on social media channels, including forums, blogs and other online or offline manifestation forms. In other words, we have the most offering possibility to monitor in real time commentator's sentiment and act accordingly. We present in this paper aspects about natural language complexity from emotional perspective and some language resources that can be used in sentiment analysis systems. This research supports the direct beneficiaries (marketing managers, PR firms, campaign managers, politicians, investors, online buyers), but, also, specialists in the field of natural language processing, linguists, psychologists, etc.*

*Keywords: sentiment, analysis levels of sentiment, language resources, sentiment lexicon.*

## 1. Introducere

Interesul pentru procesarea limbajului natural (PLN) și, implicit, pentru extragerea de opinii dintr-un text, de cele mai multe ori subiective, a crescut vizibil. Numeroasele lucrări care propun metode noi de extragere a cuvintelor/ expresiilor (des)calificatoare trezesc interesul liderilor de opinie, oamenilor de afaceri, bancherilor, actorilor politici etc., conștienți de beneficiile constituirii unor seturi de date privitoare la opiniile receptorilor. Focalizarea acestei lucrări asupra complexității limbajului natural și relațiilor logice ale acestuia este deosebit de importantă dacă se privește această abordare prin prisma translatării unei descrieri în limbaj natural într-o descriere riguroasă și neambiguă.

Opțiunea pentru o asemenea temă – analiza sentimentelor (AS), în terminologie americană *opinion mining*, întâlnite în textele vehiculate prin diverse canale media online (forumurile diverselor publicații, social-media, blogurile etc.) - vine din nevoia clarificării comportamentului descriptiv al consumatorului, afectat de multitudinea de mesaje promoționale, indiferent de natura și scopul lor. De pildă, atunci când o persoană dorește să cumpere un produs, de obicei va începe să caute comentarii și opinii scrise de alte persoane în mediul virtual, care au trecut prin această experiență (diversificarea ofertei, calitatea produsului etc.). La ora actuală, analiza sentimentelor, credem noi, este de mare actualitate și, totodată, interes în zona de cercetare în prelucrarea limbajului natural.

*Premisa prezentului studiu este aceea că estimarea orientării emoționale prezente în documentul-text ajută în construirea unei baze de date cu informații referitoare la subiecte, servicii, produse etc. de mare interes public, care poate servi implementării unui instrument de procesare a limbajului natural, util prezicerii nevoilor potențialului consumator.*

Lucrarea este structurată în cinci capitole. După o scurtă introducere cu privire la tema propusă, în capitolul doi facem o scurtă incursiune asupra preocupărilor anterioare care fac obiectul multor lucrări axate pe analiza sentimentelor. În capitolul trei sunt descrise tehnicile de analiză a sentimentelor, urmând ca în capitolul 4 să menționăm aspecte de modelare a percepției compoziționale, cu accent pe metoda propusă. Ultimul capitol evidențiază concluziile prezentului studiu, menționând și unul dintre proiectele de cercetare viitoare pe care Grupul de Cercetare în Tehnologii ale Limbajului Natural al Facultății de Informatică de la Universitatea “Alexandru Ioan Cuza” din Iași (NLP-Group@UAIC-FII) îl are în vedere.

## 2. Preocupări anterioare

Analiza sentimentelor a făcut obiectul unor cercetări întreprinse în 2001 de Das, Chen (Das & Chen, 2001) și Tong (2001) cu privire la opiniile exprimate pe piața de vânzare. De reținut, rămâne clasificarea gradului de pozitivitate al unui text la nivel de document, propoziție sau caracteristici, constând în cuvinte ce exprimă emoții (ex: *furios*, *supărat*, *fericit*). O abordare a determinării sentimentelor într-un text, spre exemplu, electoral (Gîfu, 2011) a constat în clasificarea cuvintelor cu încărcătură emoțională în două clase: pozitive și negative. Mai mult, există abordări care iau în considerare și clasa neutru (valoarea 0), asociind cuvintelor câte o valoare de la -5 la +5, scală pe care noi am redus-o de la -3 la +3, considerând-o suficient de fină pentru discursul public (Gîfu and Cristea, 2012) și chiar mai mult, de la -1 la +1 (metoda prezentei lucrări).

În termeni generali, AS constă în extragerea de opinii dintr-un text. Întâlnită și ca analiza subiectivității<sup>1</sup> într-o lucrare a lui K. Dave et. al (2003: 519-528), AS ar consta în „procesarea rezultatelor căutării pentru un anumit articol, generând o listă de attribute ale produsului (calitate, caracteristici etc.) și agregarea opiniilor pentru fiecare dintre ele (sărăcăcios, combinat, bun)”. Mai mult, AS a fost interpretată ca incluzând diverse tipuri de analiză și evaluare (Liu, 2006).

De la cercetarea subiectivității, AS s-a axat și pe cercetarea obiectivității într-un text, la final rezultând o clasificare a textelor în două clase: obiectiv și subiectiv, de cele mai multe ori mult mai dificil de întreprins decât clasificarea după polaritate (Mihalcea et. al., 2007).

## 3. Tehnici de analiză a sentimentelor

Există o explozie uriașă azi de „sentimente”, disponibile de la social media, inclusiv Twitter, Facebook, forumuri, bloguri etc. Analiza sentimentelor oferă organizațiilor posibilitatea de a monitoriza opiniile cu privire la produsele/serviciile și reputația acestora (așa-numitul feedback)<sup>2</sup>, de pe diferite site-uri de social media în timp real și să acționeze în consecință.

O primă clasificare a textelor analizate (propoziții/ fraze) ar putea ține seama de criteriul subiectivității. Vorbim de două clase principale: obiective, care conțin informații concrete, și subiective, care conțin opinii explicite, credințe și opinii cu privire la entitățile specifice. Atenția noastră se îndreaptă, mai mult, pe analiza textelor subiective. Cu toate acestea, și

<sup>1</sup> Au mai existat termeni ca *extragerea de recenzii sau calcul al afecțiunii*. (Ioana Ardeleanu, *Extragerea de opinii din texte*, lucrare de licență coord. de prof.univ.dr. Dan Cristea, 2013).

<sup>2</sup> Există pachete statistice majore, cum ar fi SAS și SPSS, care includ module specializate de analiza sentimentelor.

textele obiective pot fi interesant de analizat, atunci când descriu opiniile lor de tip „stock picking”.

Spre exemplu atunci când opinăm cu privire la o pensiune “X”, unde am fost cazați în vacanța de iarnă: *Pensiunea X dispune de camere spațioase, curate, bine încălzite*, unde ești întâmpinat de un personal amabil etc.

În acest text, aplicațiile cu privire la analiza sentimentelor trebuie să fie în măsură să ofere un scor pentru întreaga reexaminare (descriere), precum și scoruri pentru fiecare aspect individual (fiecare element descriptiv apreciativ). Vorbim, în fapt, de analiza globală și analiză secvențială.

Amintim câțiva autori reprezentativi cu preocupări și comentarii importante în privința analizei sentimentelor (Liu, 2010, 2012; Pang and Lee, 2008).

Ne vom concentra, în cele ce urmează, pe următoarele aspecte:

- Analiza sentimentelor la nivel de document;
- Analiza sentimentelor la nivel de propoziție/frază;
- Aspecte bazate pe analiza de sentiment;
- Analize de sentiment comparativă;
- Achiziția lexiconului - sentiment.

### 3.1. Analiza sentimentelor la nivel de document

Este cea mai simplă formă de analiză a sentimentului și se presupune că documentul conține o opinie cu privire la un singur obiect principal exprimat de autorul mesajului. În literatura de specialitate, am regăsit două abordări pentru analiza sentimentelor la nivelul documentului: *supervizată* și *nesupervizată*.

a) Abordarea *supervizată* presupune că există un set finit de clase. Documentul trebuie să fie clasificat, datele de antrenare fiind atribuite fiecărei clase. Cel mai simplu caz este atunci când există două clase: *pozitiv* și *negativ*. Desigur, poate fi adăugată și o clasă *neutru* sau poate fi luată în considerare o scală numerică, la care să fie raportat documentul (de exemplu: *SentiWordNet*<sup>3</sup>). Având în vedere datele de antrenare, sistemul învață un model de clasificare, folosind un algoritm de clasificare, cum ar fi SVM (*Support Vector Machines*)<sup>4</sup> sau KNN (K-nearest neighbors)<sup>5</sup>. Această clasificare este apoi folosită pentru a eticheta documente noi în diversele lor clase sentiment. Autori ca Pang, Lee and Vaithyanathan (2002) au arătat că precizia bună este atinsă chiar și atunci când fiecare document este reprezentat ca un grup de cuvinte (*bag of words*). Reprezentări mai avansate utilizează POS (partea de vorbire), lexicoane sentiment și structurile de parsare (segmentare la nivel de unitate lexicală<sup>6</sup>).

<sup>3</sup> SentiWordNet [Esuli and Sebastiani, 2006(a)] este o resursă lexicală pentru analiza sentimentelor. SentiWordNet atribuie fiecărui *synset* (mulțimi de cuvinte sinonime între ele care definesc un concept lexical) din WordNet [Esuli and Sebastiani, 2006(b)] trei scoruri-sentiment numerice: pozitivitate, negativitate și obiectivitate.

<sup>4</sup> SVM (Support Vector Machine) mapează vectorul de intrare într-un spațiu cu mai multe dimensiuni unde se construiește un hiperplan liniar de separație. Aceste clasificatoare se pot reprezenta prin funcții kernel (nucleu) ce definesc similarități între perechi de date.

<sup>5</sup> În cadrul metodei KNN (K-nearest neighbour), un exemplu necunoscut este clasificat prin votul majoritar al vecinilor lui, fiind atribuit clasei din care are majoritatea vecinilor din cadrul celor k vecini cei mai apropiați (vot majoritar). Pentru k=1 obiectul va fi atribuit clasei vecinului celui mai apropiat.

<sup>6</sup> unități lexicale pot fi: morfem, cuvânt, clauză, propoziție, frază, document.

b) Abordarea *nesupervizată* în document la nivel de analiză a sentimentelor se bazează pe determinarea orientării semantice (OS) din fraze specifice din cadrul documentului. Dacă media OS dintre aceste fraze este de peste un prag predefinit, documentul este clasificat ca fiind pozitiv. În caz contrar, se consideră negativ. Există două abordări principale de selecție a frazelor: un set de modele de POS-uri predefinite pot fi utilizate pentru a selecta aceste fraze (Turney, 2002) sau un lexicon de cuvinte-sentiment și expresii (Taboada et. al, 2011).

### 3.2. Analiza sentimentelor la nivel de propoziție/frază

Un singur document poate conține mai multe opinii chiar și cu privire la aceeași entitate. Când vrem să avem o imagine mai clară a opiniilor exprimate cu privire la o entitate (lucru, ființă, organizație, localitate etc.) trebuie să trecem la nivelul propoziției/ frazei.

Presupunerea inițială este aceea că avem știință de identitatea entității care apare în text. Mai mult, presupunem că există o singură opinie în fiecare frază. Această presupunere poate fi realizată prin divizarea frazei în propoziții (fragment de text care conține și un verb predicativ) în care fiecare propoziție conține doar o opinie. Înainte de a analiza polaritatea la nivel de frază trebuie să determinăm dacă propozițiile sunt subiective sau obiective. Numai propozițiile subiective vor fi apoi analizate. Majoritatea metodelor folosesc metode supravegheate pentru a clasifica propozițiile în două clase (Yu and Hatzivassiloglou, 2003).

Amintim abordarea bazată pe reduceri minimale, propusă în Pang și Lee (2004). Principala premisă este că propozițiile vecine ar trebui să aibă aceeași clasificare subiectivă. Putem apoi clasifica aceste fraze în clase pozitive sau negative. Totodată s-a arătat că este recomandabil să se adopte strategii diferite pentru acest tip de analize (Narayanan, Liu and Choudhary, 2009). La nivel semantic, pot interveni interogațiile, sarcasmul, metafora, umorul, elemente dificil de detectat, mai ales în anumite contexte (spre exemplu în contextul electoral).

### 3.3. Aspecte bazate pe analiza de sentiment

Abordările anterioare funcționează bine atunci când fie întregul document, fie fiecare fragment de text analizat se referă la o singură entitate. Cu toate acestea, în multe cazuri, oamenii vorbesc despre entități care au mai multe trăsături (attribute) și, desigur, opiniile diferă. Acest lucru se întâmplă de multe ori în comentarii cu privire la produse, servicii etc. sau în forumuri de discuții dedicate diverselor categorii de entități (cum ar fi actori politici, mașini, aparate foto, smartphone-uri și chiar produse farmaceutice).

Spre exemplu: *Becali a ajutat mult săracii, construindu-le case, dându-le bani, dar nimeni nu a știut exact cum a făcut atâția bani decât în momentul în care a fost arestat din cauza unor terenuri cumpărate pe nimic.*

Clasificarea acestui enunț, pozitiv (*Becali a ajutat mult săracii, construindu-le case, dându-le bani*) și negativ (*dar nimeni nu a știut exact cum a făcut atâția bani decât în momentul în care a fost arestat din cauza unor terenuri cumpărate pe nimic*) cu privire la acest personaj ar fi posibilă numai dacă am analiza propozițiile pe rând.

Analiza bazată pe încărcătura emoțională este problema de cercetare care se concentrează pe recunoașterea tuturor formelor de expresii-sentiment într-un document dat și aspectele la care se referă. Abordarea clasică, utilizată din ce în ce mai mult de departamentul de relații publice din marile companii/ societăți comerciale, constă în identificarea naturii emoționale a comentariilor cu privire la calitatea produselor, serviciilor, imaginea brandului etc. În limbaj natural, extragerea tuturor structurilor substantivale (NPs) și apoi păstrarea doar

a acelora a căror frecvență este peste un prag învățat experimental (Hu and Liu, 2004). De asemenea, analiza bazată pe încărcătura emoțională se concentrează pe recunoașterea tuturor formelor de expresii-sentiment într-un document dat, cât și pe entitățile la care se referă. Cu alte cuvinte, identificarea aspectelor evaluative care sunt menționate în mod explicit în propoziții. Există, totuși, multe opinii care nu sunt menționate în mod explicit în propoziții, putând fi deduse din expresiile-sentiment menționate implicit.

Spre exemplu: *care se poate deduce din fragmentul...*

### 3.4. *Analize de sentiment comparativă*

În multe cazuri, utilizatorii nu oferă o opinie directă despre un produs, preferând în schimb păreri comparabile, cum ar fi:

*Dacia Logan* arată mult mai bine decât *Dacia Solenza*.

În acest caz, scopul sistemului de analiză a sentimentelor este de a identifica propoziții care conțin opinii comparative, precum și pentru a extrage entitatea preferată din acestea. Autori ca Jindal and Liu (2006) descriu această metodă analitică. Utilizarea unui număr relativ mic de cuvinte, ca adjective adverbiale comparative (*mai mult, mai puțin, ușoare* etc.), adjective și adverbe superlative (*mai, cel puțin, cele mai bune*), clauze adiționale (*favoare, mare, prefera, decât, superioară, inferior, numărul unu, împotriva*), poate acoperi 98% din totalul opiniilor comparative.

Pentru aceste cuvinte/ sintagme fiind folosite foarte des, dar cu precizie scăzută, poate fi folosit un clasificator<sup>7</sup> pentru a filtra frazele care nu conțin opinii comparative. Un algoritm simplu pentru a identifica entitățile preferate pe baza tipului de comparații utilizate și prezența negației prezintă Ding, Liu and Zhang (2009).

### 3.5. *Achiziția lexiconului - sentiment*

Până acum am putut observa că lexiconul este resursa cea mai importantă pentru majoritatea tehnicilor de analiză a sentimentelor.

Există trei opțiuni pentru achiziționarea lexiconului-sentiment:

a) *abordări manuale*, în care cercetătorii construiesc un lexicon manual, constând dintr-un set de cuvinte selectate din dicționare explicative, ulterior, extins prin utilizarea de resurse lexicale deja existente (îmbogățirea cu sinonime, antonime). Am amintit deja de WordNet. Un algoritm la îndemână este cel propus de Kamps, J., Marx, M., Mokken, R.J. and de Rijke, M. (2004).

b) *abordări bazate pe corpus*, în care un set de cuvinte/ sintagme extrase dintr-un corpus relativ restrâns este extins prin utilizarea unui corpus mare de documente dintr-un singur domeniu.

Principalul dezavantaj al oricărui algoritm pe bază de dicționar (a) este acela că lexiconul dobândit este prea general și, prin urmare, nu surprinde particularitățile specifice unui domeniu specific (Dragut et al., 2010; Peng și Park, 2011).

Dacă dorim să creăm un lexicon-sentiment specific unui anumit domeniu trebuie să utilizăm un algoritm bazat pe un corpus. O lucrare clasică în acest domeniu (Hatzivassiloglou and McKeown, 1997) evidențiază conceptul de coerența sentimentelor, care permite o identificare a adjectivelor cu polaritate complexă (seed-adjectives). Altfel spus, un set de

<sup>7</sup> Spre exemplu, clasificatorul Naïve Bayes, o metodă statistică de clasificare și recunoaștere a formelor, unde fiecare document este privit ca o colecție de cuvinte, iar ordinea cuvintelor este considerată irelevantă.

conectori lingvistici (*și, sau, nici, fie, sau*) a fost folosit pentru a găsi adjectivele care sunt conectate la adjective cu polaritate cunoscută.

Spre exemplu: *bărbat puternic și armonios*.

Dacă știm că *puternic* este un cuvânt pozitiv, putem presupune că prin utilizarea conectorului *și* cuvântul *armonios* este pozitiv.

#### 4. Modelarea percepției compoziționale

Există o necesitate pentru o mai bună modelare a percepției compoziționale. La nivel de frază, aceasta înseamnă calcul mai precis al sentimentului general cu privire la subiectul (ceea ce aici numim entitate) comentat, cuvintele/sintagmele-sentiment transformând, în fapt, structura semantică a frazei. De menționat faptul că o entitate, pe parcursul textului, poate să apară sub forma unor structuri anaforice<sup>8</sup>.

De exemplu: **Mihai** învață pe brânci pentru examen. **El** vrea să fie bursier.

Este evident că pronumele **el** se referă la **Mihai**, ambele propoziții făcând parte din același context lingvistic.

Când într-un document se fac referiri la mai multe persoane, este esențial să se identifice textele relevante pentru fiecare entitate în parte. În acest moment, precizia în identificarea acestor texte relevante este departe de a fi satisfăcătoare. Deși există unele abordări care folosesc metode de clasificare pentru a identifica spre exemplu ironia, acestea nu sunt încă integrate în sisteme autonome de analiză a sentimentelor. O altă mare provocare sunt textele cu greșeli gramaticale, cu punctuație lipsă sau problematică, care abundă în argou etc.

Abordările actuale cu privire la analiza sentimentelor se opresc de obicei la extragerea sentimentelor din declarațiile subiective. Astfel de afirmații apar frecvent în articole de știri. Rămâne însă problema contextului pe care, deocamdată, nu a fost surprinsă algoritmic.

##### 4.1. Descrierea metodei

În prezenta lucrare, aplicația este capabilă să depisteze și să explice aprecieri calitative asupra unor entități (companii, produse, persoane etc.). În realizarea softului s-au parcurs următoarele etape (figura 1):

- *construirea unei antologii de entități, categorii și valori*, fiind necesară pentru obținerea unui rezultat cât mai corect și complet;
- *preprocesarea textului*, etapă care se referă la adnotarea, împărțirea unui text în entități, în cazul de față cuvinte, simboluri sau alte elemente importante numite *token*-uri. Fiecare token a fost adnotat automat la partea de vorbire (Simionescu, 2011);
- *extragerea grupurilor nominale de interes (text chunking)*<sup>9</sup>, cu alte cuvinte împărțirea unui text în secvențe de cuvinte corelate sintactic;
- *recuperarea legăturilor de natură anaforică*, moment în care ne propunem să nu pierdem nici o referire la o anumite entitate. Extragerea legăturilor de natură anaforică reprezintă găsirea referinței unui cuvânt/ unei expresii cu privire la o entitate din textul analizat;

<sup>8</sup> Anafora este procesul de utilizare a unei entități lingvistice cum ar fi un pronume, cu scopul de a referi la o altă entitate. A se vedea (Kamp & Reyle, 1993).

<sup>9</sup> Noun Phrase chunking (NP-chunk) se referă la extragerea grupurilor/expresiilor substantivale dintr-o propoziție.

Spre exemplu: „*Oamenii* încep să iasă în grupuri. Nimeni n-are să-*i* oprească.”

În propoziția de mai sus, *-i*-ul se referă la *Oamenii*. Pentru extragerea legăturii de natură anaforică, în aplicația noastră am folosit unealta informatică RARE (*Robust Anaphora Resolution Engine*)<sup>10</sup>, implementată de Eugen Ignat (2011).

- *extragerea entităților*, etapă în care s-a utilizat modulul NER (*Name Entity Recognition*) în vederea extragerii automate a entităților. În fapt, pot fi recunoscute entități precum nume de persoane, organizații, localități etc. Se primește la intrare (*input*) un fișier de tip .txt<sup>11</sup> și la ieșire (*output*) obținem un fișier de tip .txt ce conține doar entitățile menționate în textul analizat.

De exemplu: „Vodafone România oferă cea mai bună conectivitate pentru serviciile de date dintre toate rețelele mobile GSM/ UMTS/ CDMA din România”.

Din acest text va rezulta un fișier care conține următoarele entități: Vodafone, România, Vodafone România, GSM, UMTS, CDMA. În cazul unei entități care apare în textul preprocesat de mai multe ori, fișierul de la ieșire va reține acea entitate doar o singură dată.

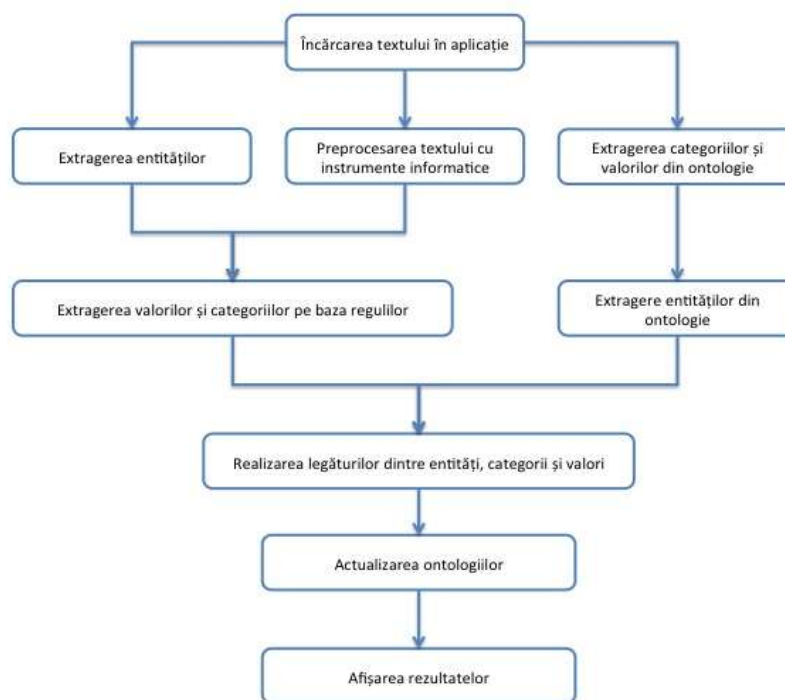


Figura 1 – Arhitectura programului informatic

- *recunoașterea categoriilor, valorilor și relaționarea cu entitățile*. Având la dispoziție fișierele rezultate în urma parcurgerii etapelor anterioare, se vor extrage automat categoriile, valorile și relaționarea cu entitățile cu ajutorul unui set de reguli<sup>12</sup>.

Rămânând la exemplul anterior, în fișier textul va apărea sub forma:

<entity type=„company”>Vodafone România</entity>

<sup>10</sup> Scopul instrumentului RARE este de a extrage lanțuri coreferențiale (v. metashare.infoiasi.ro).

<sup>11</sup> Se preferă editorul NotePad++ pentru diversele funcționalități (spre ex. recunoașterea diacriticelor limbii române).

<sup>12</sup> Informatic vorbind o expresie regulată este un șir de caractere șablon care descrie mulțimea cuvintelor posibile, formate cu acele caractere, respectând anumite reguli.

<category>conectivitate pentru serviciile de date</category>

<value =,1”>bună</value>

Aceste expresii regulate folosesc paranteze (rotunde, pătrate, acolade) prin care formează regulile de formare a cuvintelor. Utilitatea cea mai frecventă a unei expresii regulate constă în a recunoaște dacă un șir conține sau nu cuvinte sau sub-șir care pot fi formate prin expresia regulată respectivă.

De exemplu: șirul de caractere m[ăi]r poate fi interpretat în *măr* și *mir*.

#### 4.2. Metodologia de lucru

- se adnotează manual textul cu ajutorul aplicației Palinka<sup>13</sup> un corpus de texte pentru a construi o ontologie<sup>14</sup> de entități, categorii și valori;
- textul este preprocesat cu ajutorul unelei POS-tagger pentru limba română (Simionescu, 2011);
- se recunosc și se adnotează grupurile nominale de interes pentru aplicația cu ajutorul NP-chunker (Simionescu, 2011);
- se extrag automat numele proprii de entități, cu ajutorul instrumentului NER (Name Entity Recognition);
- se extrag legăturile de tip anaforic din text cu ajutorul instrumentului RARE (*Robust Anaphora Resolution Engine*);
- se recunosc în text entitățile, categoriile și valorile din ontologiile deja construite;
- se scrie un set de reguli pentru recunoașterea valorilor și se realizează legăturile de tip entitate-categorie-valoare;

În lucrarea de față, identificarea orientărilor de opinii referitoare la o trăsătură a unei entități dintr-un text se bazează pe un dicționar de opinii, rezultat din adnotarea unui corpus (format din cuvinte și sintagme cu încărcătură emoțională) pentru a determina orientarea opiniilor identificate. Pe lângă acest dicționar, important de tratat sunt negațiile și clauzele de tipul „dar”, „cu excepția”, „exceptând”.

Concret, parcurgem următorii pași:

- se identifică cuvintele și frazele de opinie;
- fiecărui cuvânt pozitiv i se atribuie un scor de opinie de (+1), iar fiecarui cuvânt negativ, (-1);
- cuvintele dependente de context vor avea scorul 0.

Luăm spre exemplificare următorul fragment: *Imaginea aparatului de fotografiat nu este grozavă, dar bateria ține mult.*

Astfel, prima propoziție, *Imaginea aparatului de fotografiat nu este grozavă* primește scorul (+1), deoarece cuvântul *grozavă* este un cuvânt pozitiv, urmând ca în propoziția a doua, *dar bateria ține mult* să primească scorul (0), cuvântul *mult* fiind considerat dependent de context frazal. Remarcăm faptul că nu au fost luate în vedere cele două cuvinte, *nu*, și *dar*, primul fiind o negație și al doilea fiind un conector pragmatic.

- manipularea negației – cuvintele de negație sunt folosite pentru revizuirea scorurilor de opinie obținute la pasul anterior, pe baza unor reguli. După acest pas prima propoziție, *Imaginea aparatului de fotografiat nu este grozavă* va primi scorul (-1) urmare a considerării

<sup>13</sup> <http://clg.wlv.ac.uk/trac/palinka/>

<sup>14</sup> În limbaj natural, ontologia reprezintă cunoștințele ca un set de concepte dintr-un anumit domeniu, folosind un vocabular comun pentru a indica tipurile, proprietățile și relațiile dintre acele concepte.

negației *nu*.

- este cunoscut rolul conectorilor pragmatici (dar, doar, pentru că etc.) care introduc un nou act de enunțare, opinie contrară. Astfel, propoziția a doua, *dar bateria ține mult* primește scorul (+1) urmare a luării în considerare a conectorului *dar*.

- în ultima etapă a aplicației se realizează sumarea formelor de expresii sentiment, pentru a determina orientarea finală a opiniei rezultate din text.

Revenind la exemplul nostru, scorul frazei: *Imaginea aparatului de fotografiat nu este grozavă* (-1), *dar bateria ține mult* (+1), va consta în sumarea celor două valori:  $-1+1=0$ . Cu alte cuvinte, fraza are o intensitate emoțională neutră.

## 5. Concluzii și direcții viitoare de cercetare

Lucrarea prezintă o metodă automată capabilă să depisteze și să explice opinii referitoare la anumite entități (persoane, companii, produse etc.) identificate într-un text, indiferent de natura lui, bazată pe un dicționar de opinii rezultat din adnotarea manuală a unui corpus inițial (format din cuvinte și sintagme cu încărcătură emoțională). Mai mult, pe lângă acest dicționar, am avut în vedere rolul semantic al negațiilor și conectorilor pragmatici de tipul „dar”, „cu excepția”, „exceptând”. Această aplicație vine să sprijine dezvoltarea unei resurse lexicale complexe, necesară interpretării aprecierilor calitative întâlnite în orice fel de text, care poate fi de real folos managerilor de marketing, firmelor de PR, politicienilor, cumpărătorilor on-line, dar și specialiștilor în domeniul prelucrării limbajului natural, lingviști etc.

Din cele relatate, observăm faptul că operația de însumare a scorurilor pentru fiecare propoziție cu o anumită încărcătură emoțională nu este suficient acoperitoare (v. exemplul precedent când apare variant de neutralizare a sentimentelor), motiv pentru care ne propunem să adăugăm diferențieri de intensitate în exprimarea opiniilor. În limba română, folosirea superlativului amplifică semantic convingerile unui opinent. Spre exemplu: *Vodafone România oferă cea mai bună conectivitate pentru serviciile de date dintre toate rețelele mobile GSM/ UMTS/ CDMA din România*.

*Mihai este cel mai rău student de la facultatea noastră.*

Superlativul ne determină să lărgim scala de valori. Astfel exprimări ca *cea mai* poate prelua gradul de pozitivitate sau negativitate. În primul exemplu, *cea mai bună* primește valoarea (+2), iar în exemplul al doilea *cel mai rău* primește valoarea (-2).

## Referințe bibliografice

Ardeleanu, I.: *Extragerea de opinii din texte*, lucrare de licență coord. de prof.univ.dr. Dan Cristea, Universitatea “Alexandru Ioan Cuza” din Iași, 2013.

Dave, K., Lawrence, S. and Pennock D.M.: *Mining the peanut gallery: Opinion extraction and semantic classification of product reviews*, Proceedings of WWW, 2003.

Das, S. and Chen, M.: *Yahoo! For Amazon: Extracting market sentiment from stock message boards*. In Proceedings of the Asia Pacific Finance Association Annual Conference (APFA), 2001.

Ding, X., Liu, B. and Zhang, L.: *Entity discovery and assignment for opinion mining applications*. In Proceedings of ACM SIGKDD International Conference on Knowledge

- Discovery and Data Mining, 2009.
- Dragut, E.C., Yu, C., Sistla, P. and Meng, W.: *Construction of a sentimental word dictionary*. In Proceedings of ACM International Conference on Information and Knowledge Management, 2010.
- Esuli, A. and Sebastiani, F.: *SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining*. Proceedings from International Conference on Language Resources and Evaluation (LREC), Genoa, 2006.
- Esuli, A. and Sebastiani, F.: *Determining term subjectivity and term orientation for opinion mining*. In Proceedings of EACL-06, 11th Conference of the European Chapter of the Association for Computational Linguistics, Trento, IT. Forthcoming, 2006.
- Gîfu, D.: *Violența simbolică în discursul electoral*, Casa Cărții de Știință, Cluj-Napoca, 2011.
- Gîfu, D. and Cristea, D.: *Multi-dimensional analysis of political language* in “Future Information Technology, Application, and Service” (volum 1/164), James J. (Jong Hyuk) Park, Victor C.M. Leung, Cho-Li Wang, Taeshik Shon (editors), Springer, 2012.
- Hatzivassiloglou, V. and McKeown, K.: *Predicting the semantic orientation of adjectives*. In Proceedings of the Joint ACL/EACL Conference, 1997.
- Hu, M. and Liu, B.: *Mining opinion features in customer reviews*. In Proceedings of AACL, 2004.
- Ignat, E.: *RARE-UAIC (Robust Anaphora Resolution Engine)*, resursă gratuită pe META-SHARE, Universitatea “Alexandru Ioan Cuza” din Iași, 2011.
- Jindal, N. and Liu, B.: *Identifying comparative sentences in text documents*. In Proceedings of ACM SIGIR Conf. on Research and Development in Information Retrieval, 2006.
- Kamp, H. & Reyle, U.: *From Discourse to Logic*, Dordrecht: Kluwer, 1993.
- Liu, B.: *Web data mining; Exploring hyperlinks, contents, and usage data*. In Opinion Mining, Springer, Heidelberg, 2006.
- Liu, B.: *Sentiment analysis and subjectivity*. Handbook of Natural Language Processing. N. Indurkha and F.J. Damerau, eds., 2010.
- Liu, B.: *Sentiment analysis and opinion mining*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2012.
- Mihalcea, R., Banea C., and Wiebe, J.: *Learning Multilingual Subjective Language via Cross-Lingual Projections*. 45th Annual Meeting of the Association for Computational Linguistics (ACL-2007).
- Narayanan, R., Liu, B. and Choudhary, A.: *Sentiment analysis of conditional sentences*. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (Singapore). Association for Computational Linguistics, 2009.
- Pang, B. and Lee, L.: *Opinion mining and sentiment analysis*. Foundations and Trends in Information Retrieval 2, 1-2, 2008.
- Pang, B., Lee, L. and Vaithyanathan, S.: *Thumbs up? Sentiment Classification using machine learning techniques*. In Proceedings of EMNLP-02, 7th Conference on Empirical Methods in Natural Language Processing (Philadelphia, PA). Association for Computational Linguistics, Morristown, NJ, 2002.
- Pang, B. and Lee, L.: *A Sentimental Education: Sentiment Analysis using Subjectivity Summarization based on minimum cuts*. In Proceedings of the Association for Computational Linguistics, 2004.

- Peng, W. and Park, D.H.: *Generate adjective sentiment dictionary for social media sentiment analysis using constrained nonnegative matrix factorization*. In Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, 2011.
- Simionescu, R.: *UAIC Romanian Part of Speech Tagger*, lucrare de disertație coord. de prof.univ.dr. Dan Cristea, Universitatea “Alexandru Ioan Cuza” din Iași, nlptools.info.uaic.ro, 2011.
- Simionescu, R.: *NP-chunker (Noun Phrase chunking)*, instrument implementat la Universitatea “Alexandru Ioan Cuza” din Iași, nlptools.info.uaic.ro, 2011.
- Taboada, M., J. Brooke, J., Tofiloski, M., Voll, K. and Stede, M.: *Lexicon-based methods for sentiment analysis*. Computational Linguistics 37, 2, 2011.
- Turney, P.: *Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews*. In Proceedings of the Association for Computational Linguistics, 2002.
- Tong, R.M.: *An operational system for detecting and tracking opinions in on-line discussion*. Workshop note, SIGIR 2001 Workshop on Operational Text Classification.
- Yu, H. and Hatzivassiloglou, V.: *Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences*. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2003.