NATURAL LANGUAGE COMPLEXITY. SENTIMENT ANALYSIS

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

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) şi, 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., conştienţi de beneficiile constituirii unor seturi de date privitoare la opiniile receptorilor. Focalizarea acestei lucrări asupra complexităţii limbajului natural şi relaţiilor logice ale acestuia este deosebit de importantă dacă se priveşte această abordare prin prisma translatării unei descrieri în limbaj natural într-o descriere riguroasă şi neambiguă.

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

Premisa prezentului studiu este aceea că estimarea orientării emoţionale prezente în documentul-text ajută în construirea unei baze de date cu informaţii referitoare la subiecte, servicii, produse etc. de mare interes public, care poate servi implementării unui instrument de procesare a limbajului natural, util prezicerii nevoilor potenţialului 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 menţionăm aspecte de modelare a percepţiei compoziţionale, cu accent pe metoda propusă. Ultimul capitol evidenţiază concluziile prezentului studiu, menţionând şi unul dintre proiectele de cercetare viitoare pe care Grupul de Cercetare în Tehnologii ale Limbajului Natural al Facultăţii de Informatică de la Universitatea "Alexandru Ioan Cuza" din Iaşi (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) şi Tong (2001) cu privire la opiniile exprimate pe piaţa de vânzare. De reţinut, rămâne clasificarea gradului de pozitivitate al unui text la nivel de document, propoziţie sau caracteristici, constând în cuvinte ce exprimă emoţii (ex: *furios*, *supărat*, *fericit*). O abordare a determinării sentimentelor într-un text, spre exemplu, electoral (Gîfu, 2011) a constat în clasificarea cuvintelor cu încărcătură emoţională în două clase: pozitive şi negative. Mai mult, există abordări care iau în considerare şi 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) şi chiar mai mult, de la -1 la +1 (metoda prezentei lucrări).

În termeni generali, AS constă în extragerea de opinii dintr-un text. Întâlnită şi ca analiza subiectivităţii[1] 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 atribute ale produsului (calitate, caracteristici etc.) şi agregarea opiniilor pentru fiecare dintre ele (sărăcăcios, combinat, bun)". Mai mult, AS a fost interpretată ca incluzând diverse tipuri de analiză şi evaluare (Liu, 2006).

De la cercetarea subiectivităţii, AS s-a axat şi pe cercetarea obiectivităţii într-un text, la final rezultând o clasificare a textelor în două clase: obiectiv şi 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 uriaşă azi de „sentimente", disponibile de la social media, inclusiv Twitter, Facebook, forumuri, bloguri etc. Analiza sentimentelor oferă organizaţiilor posibilitatea de a monitoriza opiniile cu privire la produsele/serviciile şi reputaţia acestora (aşa-numitul feedback)[2], de pe diferite site-uri de social media în timp real şi să acţioneze în consecinţă.

O primă clasificare a textelor analizate (propoziţii/ fraze) ar putea ţine seama de criteriul subiectivităţii. Vorbim de două clase principale: obiective, care conţin informaţii concrete, şi subiective, care conţin opinii explicite, credinţe şi opinii cu privire la entităţile specifice. Atenţia noastră se îndreaptă, mai mult, pe analiza textelor subiective. Cu toate acestea, şi

[1] Au mai existat termeni ca *extragerea de recenzii sau calcul al afecţiunii*. (Ioana Ardeleanu, *Extragerea de opinii din texte*, lucrare de licenţă coord. de prof.univ.dr. Dan Cristea, 2013).

[2] Există pachete statistice majore, cum ar fi SAS şi 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 cazaţi în vacanţa de iarnă: *Pensiunea **X** dispune de camere spaţioase,* curate*, bine încălzite,* unde eşti întâmpinat de un *p*ersonal amabil etc.

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

Amintim câţiva autori reprezentativi cu preocupări şi comentarii importante în privinţa 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 propoziţie/frază;
- Aspecte bazate pe analiza de sentiment;
- Analize de sentiment comparativă;
- Achiziţia lexiconului - sentiment.

3.1. Analiza sentimentelor la nivel de document

Este cea mai simplă formă de analiză a sentimentului şi se presupune că documentul conţine 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* şi *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* şi *negativ*. Desigur, poate fi adăugată şi o clasă *neutru* sau poate fi luată în considerare o scală numerică, la care să fie raportat documentul (de exemplu: *SentiWordNet*[3]). Având în vedere datele de antrenare, sistemul învaţă un model de clasificare, folosind un algoritm de clasificare, cum ar fi SVM (*Support Vector Machines)*[4] sau KNN (K-nearest neighbors)[5]. 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 şi atunci când fiecare document este reprezentat ca un grup de cuvinte (*bag of words*). Reprezentări mai avansate utilizează POS (partea de vorbire), lexicoane sentiment şi structurile de parsare (segmentare la nivel de unitate lexicală[6]).

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

[4] SVM (Support Vector Machine) mapează vectorul de intrare într-un spaţiu cu mai multe dimensiuni unde se construieşte un hiperplan liniar de separaţie. Aceste clasificatoare se pot reprezenta prin funcţii kernel (nucleu) ce definesc similarităţi între perechi de date.

[5] In 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 apropiaţi (vot majoritar). Pentru k=1 obiectul va fi atribuit clasei vecinului celui mai apropiat.

[6] unităţi lexicale pot fi: morfem, cuvânt, clauză, propoziţie, 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 selecţie 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 şi expresii (Taboada et. al, 2011).

3.2. Analiza sentimentelor la nivel de propoziţie/ frază

Un singur document poate conţine mai multe opinii chiar şi cu privire la aceeaşi entitate. Când vrem să avem o imagine mai clară a opiniilor exprimate cu privire la o entitate (lucru, fiinţă, organizaţie, localitate etc.) trebuie să trecem la nivelul propoziţiei/ frazei.

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

Amintim abordarea bazată pe reduceri minimale, propusă în Pang şi Lee (2004). Principala premisă este că propoziţiile vecine ar trebui să aibă aceeaşi 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 interogaţiile, 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 funcţionează 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 entităţi care au mai multe trăsături (atribute) şi, desigur, opiniile diferă. Acest lucru se întâmplă de multe ori în comentarii cu privire la produse, servicii etc. sau în forumuri de discuţii dedicate diverselor categorii de entităţi (cum ar fi actori politici, maşini, aparate foto, smartphone-uri şi chiar produse farmaceutice).

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

Clasificarea acestui enunţ, pozitiv (*Becali a ajutat mult săracii, construindu-le case, dându-le bani*) şi negativ (*dar nimeni nu a ştiut exact cum a făcut atâţia 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 propoziţiile pe rând.

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

a acelora a căror frecvenţă este peste un prag învăţat experimental (Hu and Liu, 2004). De asemenea, analiza bazată pe încărcătura emoţională se concentrează pe recunoaşterea tuturor formelor de expresii-sentiment într-un document dat, cât şi pe entităţile la care se referă. Cu alte cuvinte, identificarea aspectelor evaluative care sunt menţionate în mod explicit în propoziţii. Există, totuşi, multe opinii care nu sunt menţionate în mod explicit în propoziţii, putând fi deduse din expresiile-sentiment menţionate 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 propoziţii care conţin opinii comparative, precum şi 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 puţin*, *uşoare* etc.), adjective şi adverbe superlative (*mai*, cel puţin, *cele mai bune*), clauze adiţionale (*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[7] pentru a filtra frazele care nu conţin opinii comparative. Un algoritm simplu pentru a identifica entităţile preferate pe baza tipului de comparaţii utilizate şi prezenţa negaţiei prezintă Ding, Liu and Zhang (2009).

3.5. Achiziţia lexiconului - sentiment

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

Există trei opţiuni pentru achiziţionarea lexiconului-sentiment:

a) *abordări manuale*, în care cercetătorii construiesc un lexicon manual, constând dintr-un set de cuvinte selectate din dicţionare explicative, ulterior, extins prin utilizarea de resurse lexicale deja existente (îmbogăţirea 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 dicţionar (a) este acela că lexiconul dobândit este prea general şi, prin urmare, nu surprinde particularităţile specifice unui domeniu specific (Dragut et al., 2010; Peng şi 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) evidenţiază conceptul de coerenţa sentimentelor, care permite o identificare a adjectivelor cu polaritate complexă (seed-adjectives). Altfel spus, un set de

[7] Spre exemplu, clasificatorul Naïve Bayes, o metodă statistică de clasificare şi recunoaştere a formelor, unde fiecare document este privit ca o colecţie de cuvinte, iar ordinea cuvintelor este considerată irelevantă.

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

Spre exemplu: *bărbat puternic şi armonios*.

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

4. Modelarea percepţiei compoziţionale

Există o necesitate pentru o mai bună modelare a percepţiei compoziţionale. 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 menţionat faptul că o entitate, pe parcursul textului, poate să apară sub forma unor structuri anaforice[8].

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

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

Când într-un document se fac referiri la mai multe persoane, este esenţial 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. Deşi 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 greşeli gramaticale, cu punctuaţie lipsă sau problematică, care abundă în argou etc.

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

4.1. Descrierea metodei

În prezenta lucrare, aplicaţia este capabilă să depisteze şi să expliciteze aprecieri calitative asupra unor entităţi (companii, produse, persoane etc.). În realizarea softului s-au parcurs următoarele etape (figura 1):

- *construirea unei antologii de entităţi, categorii şi valori,* fiind necesară pentru obţinerea unui rezultat cât mai corect şi complet;
- *preprocesarea textului*, etapă care se referă la adnotarea, împarţirea unui text în entităţi, în cazul de faţă 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)*[9], cu alte cuvinte împărţirea unui text în secvenţe de cuvinte corelate sintactic;
- *recuperarea legăturilor de natură anaforică*, moment în care ne propunem să nu pierdem nici o referire la o anume entitate. Extragerea legăturilor de natură anaforică reprezintă găsirea referinţei unui cuvânt/ unei expresii cu privire la o entitate din textul analizat;

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

[9] Noun Phrase chunking (NP-chunk) se referă la extragerea grupurilor/expresiilor substantivale dintr-o propoziţie.

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

În propoziţia de mai sus, -*i*-ul se referă la *Oamenii*. Pentru extragerea legăturii de natură anaforică, în aplicaţia noastră am folosit unealta informatică RARE (*Robust Anaphora Resolution Engine*)[10], implementată de Eugen Ignat (2011).

- *extragerea entităţilor*, etapă în care s-a utilizat modulul NER (*Name Entity Recognition*) în vederea extragerii automate a entităţilor. În fapt, pot fi recunoscute entităţi precum nume de persoane, organizaţii, localităţi etc. Se primeşte la intrare (*input*) un fişier de tip .txt[11] şi la ieşire (*output*) obţinem un fişier de tip .txt ce conţine doar entităţile menţionate în textul analizat.

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

Din acest text va rezulta un fişier care conţine următoarele entităţi: Vodafone, România, Vodafone România, GSM, UMTS, CDMA. În cazul unei entităţi care apare în textul preprocesat de mai multe ori, fişierul de la ieşire va reţine acea entitate doar o singură dată.

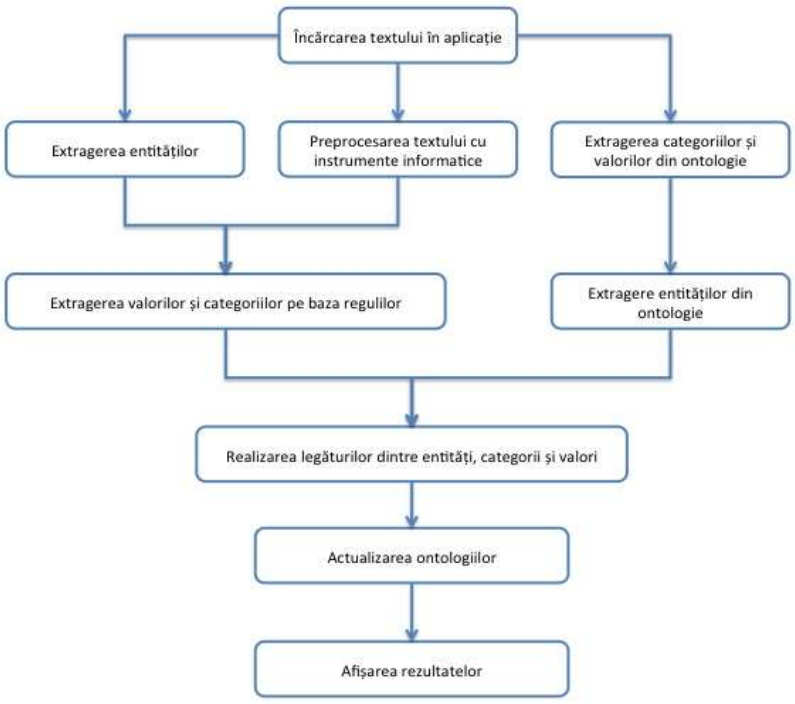

Figura 1 – Arhitectura programului informatic

- *recunoaşterea categoriilor, valorilor şi relaţionarea cu entităţile*. Având la dispoziţie fişierele rezultate în urma parcurgerii etapelor anterioare, se vor extrage automat categoriile, valorile şi relaţionarea cu entităţile cu ajutorul unui set de reguli[12].

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

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

[10] Scopul instrumentului RARE este de a extrage lanţuri coreferenţiale (v. metashare.infoiasi.ro).

[11] Se preferă editorul NotePad++ pentru diversele funcţionalităţi (spre ex. recunoaşterea diacriticelor limbii române).

[12] Informatic vorbind o expresie regulată este un şir de caractere şablon care descrie mulţimea 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 recunoaşte dacă un şir conţine sau nu cuvinte sau sub-şir care pot fi formate prin expresia regulată respectivă.

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

4.2. Metodologia de lucru

- se adnotează manual textul cu ajutorul aplicaţiei Palinka[13] un corpus de texte pentru a construi o ontologie[14] de entităţi, categorii şi valori;
- textul este preprocesat cu ajutorul uneltei POS-tagger pentru limba română (Simionescu, 2011);
- se recunosc şi se adnotează grupurile nominale de interes pentru aplicaţia cu ajutorul NP-chunker (Simionescu, 2011);
- se extrag automat numele proprii de entităţi, 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 entităţile, categoriile şi valorile din ontologiile deja construite;
- se scrie un set de reguli pentru recunoaşterea valorilor şi se realizează legăturile de tip entitate-categorie-valoare;

În lucrarea de faţă, identificarea orientărilor de opinii referitoare la o trăsătură a unei entităţi dintr-un text se bazează pe un dicţionar de opinii, rezultat din adnotarea unui corpus (format din cuvinte şi sintagme cu încărcătură emoţională) pentru a determina orientarea opiniilor identificate. Pe lângă acest dicţionar, important de tratat sunt negaţiile şi clauzele de tipul „dar”, „cu excepţia”, „exceptând”.

Concret, parcurgem următorii paşi:

- se identifică cuvintele şi 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 ţine mult*.

Astfel, prima propoziţie, *Imaginea aparatului de fotografiat nu este grozavă* primeşte scorul (+1), deoarece cuvântul *grozavă* este un cuvânt pozitiv, urmând ca în propoziţia a doua, *dar bateria ţine 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*, şi *dar*, primul fiind o negaţie şi al doilea fiind un conector pragmatic.

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

[13] http://clg.wlv.ac.uk/trac/palinka/

[14] În limbaj natural, ontologia reprezintă cunoştinţele ca un set de concepte dintr-un anumit domeniu, folosind un vocabular comun pentru a indica tipurile, proprietăţile şi relaţiile dintre acele concepte.

negaţiei *nu*.

- este cunoscut rolul conectorilor pragmatici (dar, doar, pentru că etc.) care introduc un nou act de enunţare, opinie contrară. Astfel, propoziţia a doua, *dar bateria ţine mult* primeşte scorul (+1) urmare a luării în considerare a conectorului *dar*.
- în ultima etapă a aplicaţiei 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 ţine mult* (+1), va consta în sumarea celor două valori: -1+1=0. Cu alte cuvinte, fraza are o intensitate emoţională neutră.

5. Concluzii şi direcţii viitoare de cercetare

Lucrarea prezintă o metodă automată capabilă să depisteze şi să expliciteze opinii referitoare la anumite entităţi (persoane, companii, produse etc.) identificate într-un text, indiferent de natura lui, bazată pe un dicţionar de opinii rezultat din adnotarea manuală a unui corpus iniţial (format din cuvinte şi sintagme cu încărcătură emoţională). Mai mult, pe lângă acest dicţionar, am avut în vedere rolul semantic al negaţiilor şi conectorilor pragmatici de tipul „dar", „cu excepţia", „exceptând". Această aplicaţie 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 şi specialiştilor în domeniul prelucrării limbajului natural, lingvişti etc.

Din cele relatate, observăm faptul că operaţia de însumare a scorurilor pentru fiecare propoziţie cu o anumită încărcătură emoţională nu este suficient acoperitoare (v. exemplul precedent când apare variant de neutralizare a sentimentelor), motiv pentru care ne propunem să adăugăm diferenţieri 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 reţelele 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* primeşte valoarea (+2), iar în exemplul al doilea *cel mai rău* primeşte valoarea (-2).

Referinţe bibliografice

Ardeleanu, I.: *Extragerea de opinii din texte*, lucrare de licenţă coord. de prof.univ.dr. Dan Cristea, Universitatea "Alexandru Ioan Cuza" din Iaşi, 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 Acific 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.: *Violenţa simbolică în discursul electoral*, Casa Cărţii de Ştiinţă, 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 AAAI, 2004.
Ignat, E.: *RARE-UAIC* (*Robust Anaphora Resolution Engine*), resursă gratuită pe META-SHARE, Universitatea "Alexandru Ioan Cuza" din Iaşi, 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. Indurkhya 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 disertaţie coord. de prof.univ.dr. Dan Cristea, Universitatea "Alexandru Ioan Cuza" din Iaşi, nlptools.info.uaic.ro, 2011.

Simionescu, R.: *NP-chunker* (*Noun Phrase chunking*), instrument implementat la Universitatea "Alexandru Ioan Cuza" din Iaşi, 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.