Evoluţia opiniilor în forum-uri folosind o abordare semantică de sumarizare

Dumitru Clementin Cercel
Universitatea Politehnica Bucureşti
Bd. Splaiul Independenţei, nr. 313, Bucureşti
clementin.cercel@gmail.com

Ştefan Trăuşan-Matu
Universitatea Politehnica Bucureşti
Bd. Splaiul Independenţei, nr. 313, Bucureşti
Institutul de Cercetări pentru Inteligenţa Artificială al Academiei Române
Calea 13 Septembrie nr.13, Bucureşti
stefan.trausan@cs.pub.ro

REZUMAT

Creşterea numărului de comentarii a devenit o problemă pentru consumatorii de produse, care trebuie să citească un număr semnificativ de comentarii pentru a identifica punctele tari şi punctele slabe ale unui produs. Sumarizarea opiniilor este necesară pentru a ajuta consumatorii să înţeleagă, într-un mod simplificat, o cantitate mare de opinii. Grafurile reprezintă un model adecvat pentru a reprezenta forumurile şi un mod eficient de a urmări evoluţia opiniilor cu privire la o anumită entitate. În acest articol noi vom prezenta un algoritm de vizualizare a evoluţiei opiniilor într-un forum de discuţii, ce foloseşte o abordare de sumarizare. Nu este vorba de o evoluţie în timp a opiniilor din forum, ci de o evoluţie a structurii grafului, din punct de vedere al opiniilor ce caracterizează fiecare nod.

Cuvinte cheie

Evoluţie opinii, Polaritate cuvânt, Polaritate propoziţie, Sumarizare opinii, Stanford Typed Dependencies.

Clasificare ACM

H5.2. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCERE

Resursele web cuprind două categorii de informaţii: fapte şi opinii [23]. Faptele presupun o prezentare obiectivă a realităţii, în timp ce opiniile reprezintă afirmaţii subiective despre evenimente, produse sau persoane. Extragerea de opinii ("Opinion mining") a reprezentat un domeniu de mare interes pentru cercetători odată cu creşterea popularităţii Web-ului, care generează, şi în continuare, cantităţi mari de date textuale electronice. Volumul de resurse web, cum ar fi forum-urile, blog-urile, comentariile de pe site-uri gen e-commerce, reprezintă surse de opinii care pot fi folosite atât de persoanele fizice, cât şi de organizaţii.

Analiza sentimentelor („Sentiment analysis") are ca scop detectarea opiniilor exprimate în resursele web şi clasificarea opiniilor în funcţie de sentimentul exprimat în opinii pozitive şi opinii negative. Abordările propuse de cercetători pentru a rezolva această ultimă sarcină denumită clasificarea sentimentelor („sentiment classification") folosesc clasificarea sentimentelor la nivel de document, la nivel de propoziţie şi la nivel de cuvânt [3, 7] .

Clasificarea propoziţiilor în propoziţii subiective sau obiective este numită în literatura de specialitate clasificarea subiectivităţii („subjectivity classification"). Subiectivatea nu este echivalentă cu existenţa unei opinii [17]. Putem avea propoziţii subiective care să nu exprime opinii, dar putem avea cazuri în care propoziţiile obiective implică opinii. Rezultatele prezentate în [18] arată că rezultatele analizei sentimentelor se pot îmbunătăţii dacă această sarcină este ulterioară cea a clasificării propoziţiilor.

În acest articol vom prezenta un algoritm de vizualizare a evoluţiei opiniilor într-un forum de discuţii, care simplifică graful cu mesajele ce conţin opinii. În graful final, fiecare nod rezultat prin agregarea opiniilor de aceeaşi polaritate de la nodurile vecine conţine polaritatea opiniei nodului şi cuvintele care dau respectiva polaritate. Soluţia propusă foloseşte o abordare de sumarizare de opinii. Am descompus problema în următoarele subsarcini:

- filtrarea grafului de mesaje în scopul de a elimina mesajele care nu se referă la entitatea avută în vedere, respectiv a elimina mesajele care nu conţin opinii despre această entitate;
- identificarea cuvintelor de opinie într-un mesaj. Vom considera că opiniile pot fi exprimate prin cuvinte precum adjective, adverbe şi verbe;
- calcularea polarităţii mesajului dintr-un nod al grafului;
- agregarea nodurilor (pe orizontală şi pe verticală) cu aceeaşi polaritate din graf.

Trăsăturile părţilor de vorbire ce caracterizează cuvintele dintr-o propozitie determină anumite relaţii semantice între cuvinte. În algoritmul propus relaţiile gramaticale sunt folosite pentru a determina dependenţa între cuvântul a cărui evoluţie a opiniilor dorim să o urmărim şi cuvântul ce descrie opinia despre el. Astfel, relaţii de dependenţă semnificative au fost identificate folosind parserul Stanford [16] pentru a extrage propoziţiile care conţin opinii referitoare la o entitate ţintă.

Expunerea continuă cu descrierea cadrului şi a abordărilor importante din domeniile existente pentru rezolvarea

subsarcinilor, acestea fiind urmate de formularea problemei şi prezentarea metodei propuse. Rezultatele experimentelor, discuţia despre analiza erorilor şi concluziile sunt expuse în ultimele trei părţi prezentate în lucrare.

STAREA ACTUALĂ A DOMENIULUI

Clasificarea sentimentelor

Detectarea sentimentului exprimat de un cuvânt de opinie poate fi făcută de un operator uman, folosind un corpus sau cu ajutorul unui dicţionar [13].

Analiza perechilor de adjective ce sunt legate prin conjuncţii precum „and", „or", „but", „either-or" sau „neither-nor" este abordată în [8] pentru a deduce orientarea unui adjectiv. Folosind un set de documente, algoritmul de învătare supervizat constă din următorii paşi:

- extragerea tuturor conjuncţiilor dintre adjectivele setului de documente;
- folosirea unui model de regresie log-linear pentru a clasifica perechile de adjective legate prin conjuncţii ca având aceeaşi orientare sau diferită;
- folosirea unui algoritm de grupare pentru a împărţii adjectivele în două clase. Clasa pentru care adjectivele componente au frecvenţa medie de apariţie cea mai mare în setul de documente este etichetată ca fiind pozitivă.

Acuratatetea rezultatelor în [8] este de 78,08% pentru un set de 657/679 de adjective de polaritate pozitivă, respectiv negativă.

Soluţia propusă de autori în [10] foloseşte legăturile semantice între cuvintele din WordNet [21] pentru a defini o măsură a distanţei între două cuvinte, dată de cea mai scurtă cale între ele. Pe baza acestei măsuri, orientarea semantică a unui adjectiv este calculată în funcţie de distanţa între respectivul adjectiv şi cuvintele „good" şi „bad".

Metoda propusă în [19, 20] deduce orientarea semantică a unui cuvânt din legatura sa semantică cu alte cuvinte. În acest scop, autorii definesc două mulţimi de cuvinte ce au orientare semantică în general independendă de context şi cuvintele componente sunt antonime: o mulţime de cuvinte cu polaritate pozitiva P_{words} = {"good", "nice", "excellent", "positive", "fortunate", "correct", "superior"} şi o altă mulţime de cuvinte cu polaritate negativă N_{words} = {"bad", "nasty", "poor", "negative", "unfortunate", "wrong", "inferior"}. Măsurile folosite pentru a calcula distanţa semantică între două cuvinte sunt: PMI („informaţie reciprocă punctuală") şi LSA („analiza semantică latentă). Astfel, pentru un cuvânt de intrare "word" orientarea sentimentului exprimată de el este dată de expresia:

$$SO - A(word) = \sum_{pword \in Pwords} A(word, pword) - \sum_{nword \in Nwords} A(word, nword)$$

Sumarizarea opiniilor

Sumarizarea opiniilor este diferită faţă de problema clasică de sumarizare de text, prin faptul că prima este centrată pe trăsăturile entităţii asupra cărora consumatorii şi-au exprimat punctul de vedere (opinia) şi pe sentimentele trezite de aceste trăsături. Scopul sumarizării de opinii este acela de a ajuta utilizatorii să înţeleagă mulţimea de opinii care se găseşte pe Web într-un mod simplificat.

Metodele de sumarizare de opinii pot fi împărţite astfel: metode de sumarizare bazate pe aspect si metode de sumarizare care nu sunt bazate pe aspect [11]. Prima categorie implică identificarea trăsăturilor care caracterizează un obiect (de exemplu, pentru obiectul "iPod" trăsăturile pot fi "durată baterie", "preţ", "calitate sunet"), determinarea polarităţii opiniilor exprimate şi prezentarea opiniilor sumarizate pentru fiecare trăsătură [22].

Un alt algoritm de sumarizare [12] extrage principalele subiecte din corpus prin atribuirea unei ponderi pentru fiecare cuvânt bazat pe tehnica tf-idf (term frequency - inverse document frequency), atât la nivel de paragraf, cât şi la nivel de document. Sumarizarea se face per subiect şi constă în extragerea de propoziţii care conţin conceptele relevante obţinute în pasul anterior, concepte care exprimă o opinie. Polaritatea unei propoziţii se obţine printr-o funcţie proprie care combină sentimentele cuvintelor de opinie. Testarea algoritmului este făcută folosind un corpus de ştiri şi blog-uri.

Ganesan et al. [6] propun un algoritm de sumarizare de opinii, bazat pe graf care elimină opiniile redundante.

Alţi autori [14] propun o metodă de sumarizare folosind rating-ul global al comentariilor unui produs. Fiecare comentariu este reprezentat de o mulţime de perechi (w_m, w_h), unde w_m reprezintă o trăsătură a entităţii respective şi expresia w_h este opinia cu privire la trăsătură. Algoritmul este descris în trei paşi:

- identificarea celor mai importante trăsături folosind PLSA („analiza semantică latentă probabilistică") [9] şi aplicarea algoritmului clasic de clustering k-means [15] pentru gruparea trăsăturilor;
- predicţia rating-ului pentru fiecare trăsătură. Este calculată printr-o măsură de agregare propusă de autori folosind rating-ul frazelor care fac parte din acelaşi cluster, prin urmare având asociate aceeaşi trăsătură;
- extragerea frazelor reprezentative pentru fiecare trăsătură. Rating-ul unei fraze este calculat folosind două euristici: fiecare frază dintr-un comentariu are acelaşi rating ca şi rating-ul global al comentariului, respectiv fiecare frază are rating-ul global al tuturor comentariilor.

FORMULAREA PROBLEMEI

Un forum F constă dintr-o mulţime de fire de discuţii: F = $\{T_1, ...T_k\}$. Un forum este caracterizat de o topică, ce reprezintă subiectul discuţiei. Astfel, un forum este caracterizat de setul de mesaje componente M, de setul de autori al mesajelor A şi de setul de opinii pozitive şi negative O, despre o entitate ţintă. Firele de discuţie componente ale unui forum au aceeaşi structură: un set de comentarii ale utilizatorilor, ca răspuns la unul din comentariile anterioare sau la subiectul discuţiei. Putem spune că acestea formează un graf orientat G = (V, E) unde:

- $V = \{ v_i \mid v_i = (v_i^m, v_i^a, v_{ij}^{op}, v_{ij}^{st}, v_i^t), v_i^m \in M, v_i^a \in A, v_{ij}^{op} \subset O, j \in \{0, 1, ..., n_O\}, n_O$ = numărul total de opinii în $v_i^m, i \in \{1,2, ..., n_M\}, n_M =$ numărul total de mesaje în M $\} \cup \{ v_0\}$ este setul de noduri. Nodul v_0 este un nod special introdus, fiind caracterizat doar de subiectul de discuţie;
- $E = \{e_1, e_2, ..., e_{nM}\}$ este setul de arce, astfel încât dacă $e = (v_i, v_j) \in E$ atunci mesajul v_i este replică la mesajul v_j.

Fiecare nod din graf reprezintă un mesaj v_i^m din respectivul fir, mesaj scris de un autor v_i^a şi conţine opiniile v_{ij}^{op} despre entitatea respectivă. Variabila v_i^t este momentul de timp când mesajul a fost scris.

Fiecare opinie din v_{ij}^{op} este caracterizată de tăria sentimentului v_{ij}^{st} pe care îl exprimă. Intensitatea opiniilor este exprimată printr-o funcţie $s_{op} : O \rightarrow (-1, 1) \setminus \{0\}$, unde $s_{op}(o) \in (0, 1)$ dacă opinia este pozitivă şi $s_{op}(o) \in (-1, 0)$ dacă opinia este negativă. Asociem fiecărui mesaj din graful iniţial un scor $msgScore(v_i^m)$, dat de suma dintre puterea sentimentelor exprimate de opiniile pozitive şi opiniile negative, ce sunt extrase din mesaj împărţit la numărul total de opinii din mesaj:

$msgScore(v_i^m) : M \rightarrow (-1, 1), n_0^k + n_0^l = n_0$

$$msgScore(v_i^m) = \frac{\sum_{k=1}^{n_O^k} s_{op}(v_{ik}^{op}) + \sum_{l=1}^{n_O^l} s_{op}(v_{il}^{op})}{n_O}$$

Definiţie (agregare noduri pe verticală): Două noduri din graf $v_i = (v_i^m, v_i^a, v_{ij}^{op}, v_{ij}^{st}, v_i^t)$ şi $v_k = (v_k^m, v_k^a, v_{kl}^{op}, v_{kl}^{st}, v_k^t)$ vor fuziona dacă opiniile v_{ij}^{op} şi v_{kl}^{op} au aceeaşi polaritate (pozitivă sau negativă) şi $(v_i, v_k) \in E$, rezultatul fiind un singur nod $v_i' = (v_{rs}^m, v_{rs}^a, v_{rs}^{op}, v_{rs}^{st}, v_{rs}^t)$, caracterizat prin $v_{rs}^m = v_i^m \cup v_k^m$, $v_{rs}^a = v_i^a \cup v_k^a$, $v_{rs}^{op} = v_{ij}^{op} \cup v_{kl}^{op}$, $v_{rs}^t = v_i^t \cup v_k^t$ şi $v_{rs}^{st} = v_i^{st} \cup v_k^{st}$.

Definiţie (agregare noduri pe orizontală): Două noduri din graf $v_i = (v_i^m, v_i^a, v_{ij}^{op}, v_{ij}^{st}, v_i^t)$ şi $v_k = (v_k^m, v_k^a, v_{kl}^{op}, v_{kl}^{st}, v_k^t)$ vor fuziona dacă opiniile v_{ij}^{op} and v_{kl}^{op} v_{kl}^{op} au aceeaşi polaritate (pozitivă sau negativă) şi $\exists v_k^i \in V$ astfel încât $(v_i, v_k^i) \in E$ şi $(v_k, v_k^i) \in E$ rezultatul fiind un singur nod $v_i' = (v_{rs}^m, v_{rs}^a, v_{rs}^{op}, v_{rs}^{st}, v_{rs}^t)$, caracterizat prin $v_{rs}^m = v_i^m \cup v_k^m$, $v_{rs}^a = v_i^a \cup v_k^a$, $v_{rs}^{op} = v_{ij}^{op} \cup v_{kl}^{op}$, $v_{rs}^t = v_i^t \cup v_k^t$, $v_{rs}^{st} = v_i^{st} \cup v_k^{st}$.

Pentru fiecare nod rezultat prin agregare (pe verticală sau pe orizontală) se va calcula valoarea normalizată:

$$msgScore(v_i') = \frac{msgScore(v_i) + msgScore(v_k)}{2}$$

Vizualizarea opiniilor într-un forum de discuţii se poate face în mod simplificat într-un graf final G' = (V', E') unde:

- $V' = \{ v'_i \mid v'_i = (M_i, A_i, O_i, S_i, T_i), M_i \subset M, A_i \subset A, O_i \subset O, i \in \{1,2, ..., n_M'\}, n_M' < n_M \}$. Un nod din graf este un ansamblu de mesaje M_i scrise de un set de autori A_i, la momentele de timp T_i. Opiniile O_i despre o entitate considerată având tăria sentimentelor S_i, corespund la aceste mesaje.
- E' = setul de arce, astfel încât dacă $e' = (v'_i, v'_j) \in E'$ atunci setul de mesaje corespunzătoare nodului v'_i sunt replică la mesajele nodului v'_j.

Graful final rezultat va fi caracterizat de următoarea proprietate importantă: orice cale de la nodul v_0 la un nod frunză este o alternanţă de noduri de sentimente de polaritate pozitivă şi negativă.

SISTEMUL PROPUS

Algoritmul propus primeşte ca intrare corpusul de forumuri şi entitatea pentru care dorim să urmărim evoluţia opiniilor. Entitatea poate fi un substantiv comun sau propriu. Pentru entităţile care constau din substantive comune, algoritmul va ţine cont atât de entitatea în sine, cât şi de sinonimele sale. Este folosit în acest scop WordNet pentru a obţine o listă de sinonime pentru un substantiv.

Folosind parserul Stanford vom obţine, pentru un mesaj din graf, relaţiile de dependenţă care evidenţiază perechile compuse din entitatea avută în vedere şi cuvântul care conţine opinia despre entitate. Dependenţele între cuvinte sunt relaţii gramaticale binare între un cuvânt conducător („governor") şi un cuvânt dependent (de exemplu, pentru dependenţa „nsubj (interesting, movie)" adjectivul "interesting" este cuvântul conducător, iar substantivul "movie" este partea de vorbire dependentă). Parserul Stanford poate detecta 55 de relaţii gramaticale care pot fi prezente într-o propoziţie. De asemenea, obţinem în acest pas părţile de vorbire pentru fiecare cuvânt (de exemplu, JJ – adjectiv; VBN – verb participiu trecut). Vom lua în considerare următoarele relaţii de dependenţă:

- Nsubj („nominal subject"). Am considerat cazul în care cuvântul conducător al relaţiei este orice adjectiv şi cuvântul dependent este entitatea ţintă. De exemplu, în propoziţia: "The movie is interesting", reprezentarea dependenţelor Stanford sunt: det (movie, The), nsubj (interesting, movie), cop (interesting, is), root (ROOT, interesting);
- Acomp („adjectival complement"). Cuvântul conducător al relaţiei acomp este un verb la orice timp şi partea dependentă este un adjectiv. De asemenea, verbul este în relaţie Nsubj cu entitatea urmărită. Verbul prezintă starea, acţiunea entităţii, în timp ce adjectivul exprimă opinia cu privire la această funcţie. Pentru propoziţia: "She looks very beautiful." relaţiile de dependenţă sunt: nsubj (looks, She), root (ROOT, looks), advmod (beautiful, very), acomp (looks, beautiful);
- Advmod („Adverbial Modifier"). Am considerat cazurile în care cuvântul conducător al relaţiei advcomp este un verb la orice timp sau adjectiv şi partea dependentă este un adverb.
- Neg („negation modifier"), preconj(preconjunct), conj_nor (conjuct negation). În aceeaşi etapă considerăm adverbe („not" and „neither nor") care schimbă polaritatea de opinie. În cazul în care adjectivul este determinat de o negaţie, polaritatea adjectivului este schimbată. De exemplu, în propoziţa: "The movie is not interesting.", relaţiile de dependenţă sunt: det (movie, The), nsubj (interesting, movie), cop (interesting, is), neg (interesting, not),

root (ROOT, interesting), în timp ce în propoziţia: "The movie is neither amazing nor funny.", dependenţele sunt: det (movie, The), nsubj (amazing, movie), nsubj (funny, movie), cop (amazing, is), preconj (amazing, neither), root (ROOT, amazing), conj_nor (amazing, funny).

Cu privire la cuvintele de opinie găsite, determinăm polaritatea şi tăria polarităţii lor folosind SentiWordNet [4]. SentiWordNet este o resursă lexicală în care pentru fiecare mulţime de sinonime („synset") s din WordNet este asociat trei scoruri numerice: Obj(s) (indică gradul de obiectivitate al cuvântului), Pos(s) (indică gradul de pozitivitate al cuvântului), şi Neg(s) (indică gradul de negativitate al cuvântului). Cele trei scoruri variază de la 0 la 1 şi suma lor este egală cu 1. Valorile SentiWordNet au fost obţinute prin utilizarea unui algoritm de clasificare semi-supervizată şi analiza definiţiilor conceptelor („glosses") ce sunt asociate cu mulţimile de sinonime SentiWordNet.

Pentru obţinerea polarităţii unui mesaj este nevoie de a calcula scorul său folosind formulele din prezentarea problemei. Polarităţile mesajelor pot fi reprezentate ca o funcţie msgPolarity : M -> {-1, 0, +1}. Funcţia are valoarea -1 când opinia mesajului este negativă, valoarea 0 când opinia este neutra şi +1 când opinia este pozitivă. Astfel, mesajul unui nod al grafului poate fi clasificat în mesaj de polaritate pozitivă sau negativă.

Algoritmul ce identifică polaritatea opiniei unui mesaj din graf este prezentat mai jos:

Date intrare: entitatea considerată e, un mesaj v_i^m din forum, n relaţii de dependenţă de forma relaţie$_i$(g_i, d_i) între cuvintele mesajului v_i^m

Date ieşire: opinie mesaj forum

```
scor_opinie_msg <- 0
for i = 1 to n
    if relaţie_i = nsubj şi g_i = adjectiv
    şi d_i = e
          scor_opinie_msg +=intensitate
          opinie g_i obţinută folosind
          SentiWordNet
    end if
    if relaţie_i = acomp şi g_i = verb şi
    d_i = adjectiv şi există relaţie_j(g_j,
    d_j) astfel încât relaţie_j = nsubj şi
    g_j = verb şi d_j = e şi g_i = g_j
          scor_opinie_msg += intensitate
          opinie d_i obţinută folosind
          SentiWordNet
    end if
    if relaţie_i = advmod şi (g_i = verb sau
    g_i = adjectiv) şi d_i = adverb şi
    există relaţie_j(g_j, d_j) astfel încât
    relaţie_j = nsubj şi g_j = adverb şi
    d_j = e şi d_i = g_j
          scor_opinie_msg += intensitate
          opinie d_i obţinută folosind
          SentiWordNet
    end if
    if relaţie_i = neg şi g_i = adjectiv şi
    există relaţie_j(g_j, d_j) astfel încât
    relaţie_j = nsubj şi g_j = adjectiv şi
    d_j = e şi g_i = g_j
          scor_opinie_msg += (-1) *
          intensitate opinie g_i obţinută
          folosind SentiWordNet
    end if
    if relaţie_i = conj_nor şi g_i =
    adjectiv şi d_i = adjectiv şi există
    relaţie_j(g_j, d_j) astfel încât relaţie_j
    = nsubj şi g_j = adjectiv şi d_j = e şi
    g_i = g_j şi există relaţie_K(g_K, d_K)
    astfel încât relaţie_k = nsubj şi
    g_k = adjectiv şi d_k = e şi d_i = g_k
          scor_opinie_msg += (-1) *
          (intensitate opinie g_i +
          intensitate opinie g_k obţinute
          folosind SentiWordNet)
    end if
end for
if scor_opinie_msg > 0
      polaritate pozitivă mesaj
end if
if scor_opinie_msg < 0
      polaritate negativă mesaj
end if
if scor_opinie_msg = 0
      polaritate neutră mesaj
end if
```

Primul pas al algoritmului constă în eliminarea nodurilor care nu se referă la entitatea avută în vedere, precum şi cele care nu conţin opinii despre entitate. Celelalte noduri din graf nu vor fi modificate. Aplicarea celor două definiţii de reducere a nodurilor va determina o serie de grafuri intermediare, în care nodurile reprezintă mesaje ce conţin opinii despre entitate. După aplicarea unei reduceri vom restabili legăturile dintre nodurile afectate. Graful final, obţinut când cele două definiţii nu mai pot fi aplicate, este folosit pentru vizualizarea evoluţiei opiniilor din forumul iniţial.

REZULTATE EXPERIMENTALE

Corpusul pe care l-am folosit la evaluarea sistemului propus este o colecţie de 5 forum-uri din domeniul politicii [5] extrase de pe site-ul The Huffington Post. Întrucât articolul este un studiu preliminar în acest subiect am ales sa folosim o mulţime de forum-uri de dimensiune redusă iar etichetarea mesajelor cu opinii necesită un volum mai redus de timp.

Fiecare mesaj din forumurile considerate a fost adnotat manual dacă conţine sau nu opinii despre entitatea urmărită. Atât în algoritmul propus cât şi în etichetarea manuală a forumurilor suntem interesaţi de propoziţiile care conţin opinii la modul general despre o entitate. Nu vom căuta şi opinii despre trăsături ale sale. De asemenea, nu vom trata cazurile când opiniile sunt exprimate prin informaţie non-textuală (de exemplu emoticoane).

Entitatea avută în vedere în experimente este "Barack Obama". Tabelul 1 arată numărul de noduri ale grafului după eliminarea mesajelor cu opinii, care nu se referă la entitatea vizată (notată ca fiind condiţia 1), eliminarea nodurile care conţin informaţii fără opinie despre entitate (notată ca fiind condiţia 2), precum şi agregarea de noduri

care includ opinii de aceeaşi polaritate (notată ca fiind condiţia 3). În tabelul 2 vom prezenta precizia şi acurateţea între adnotarea manuală de opinii pe graful iniţial şi opiniile calculate de algoritm pentru graful obţinut după aplicarea condiţiei 2.

Tabelul 1. Rezultate dimensiuni grafuri obţinute

	Număr de noduri graf iniţial	Număr de noduri aplicând cond. 1	Număr de noduri aplicând cond. 2	Număr de noduri aplicând cond. 3
Forum 1	100	24	9	4
Forum 2	100	11	5	3
Forum 3	100	18	6	3
Forum 4	100	13	5	2
Forum 5	100	25	7	4

Tabelul 2.Performanţele analizei sentimentelor în forum-urile considerate

	Precizie	Acurateţe
Forum 1	0.54	0.52
Forum 2	0.66	0.64
Forum 3	0.63	0.54
Forum 4	0.75	0.63
Forum 5	0.45	0.55

DISCUŢIE

Analiza erorilor procesului de extragere de opinii este influenţată de mai mulţi factori:

- polaritatea cuvintelor care exprimă sentimente depinde de domeniul şi de contextul în care sunt utilizate [17]. Într-un context specific, adjectivele pot avea un sens pozitiv, dar într-un alt context dat, sensul poate fi negativ sau neutru. Aceeaşi expresie poate avea diferite polarităţi ale sentimentului exprimat în diferite domenii. De exemplu, în ceea ce priveşte expresia "citeşte cartea!", într-un domeniu despre cărţi sensul este favorabil, dar într-un domeniu despre filme sensul este negativ. Chiar şi în acelaşi domeniu, depinde de trăsătura urmărită (de exemplu: expresia „durata mare a bateriei telefonului" comparativ cu expresia „timpul de incarcare mare al telefonului");
- în aceeaşi propoziţie pot fi opinii multiple care exprimă aceeaşi polaritate, dar pot fi prezente şi opinii multiple de polarităţi diferite [17];
- detectarea opiniilor implicite, date de expresii retorice, metafore (de exemplu: „Mâncarea a fost ca o piatră"), dubla negaţie. Un cuvânt pozitiv poate avea un sens negativ într-un context metaforic sau ironic [2].

CONCLUZII ŞI DIRECŢII VIITOARE

Sistemele de extragere de opinii au fost dezvoltate îndeosebi pentru comentariile produselor aflate pe site-uri de comerţ electronic, fiind mai puţine în cazul discuţiilor interactive. Pentru a identifica într-un text opiniile şi a le clasifica în: pozitive, negative sau neutre este necesar să se dezvolte metode complexe.

În viitor dorim să îmbunătăţim rezultatele metodei propuse pentru vizualizarea evoluţiei opiniilor într-un forum, prin propunerea unui algoritm îmbunătăţit de detectare a polarităţii pentru o propoziţie în care sunt exprimate opinii de sentimente diferite, precum şi a unui algoritm de identificare a opiniilor implicite, luând în calcul contextul în care apar opiniile. De asemenea, dorim să efectuăm o analiză a extragerii de opinii din perspectiva modelului polifonic [1] deoarece este posibil ca unele opinii să se influenţeze reciproc.

MULŢUMIRI

Adresăm mulţumiri profesorului Julien Velcin, Université Lumière Lyon2, pentru ajutorul său în realizarea acestei metode.

Rezultatele prezentate în acest articol au fost obţinute cu sprijinul Ministerului Muncii, Familiei şi Protecţiei Sociale prin Programul Operaţional Sectorial Dezvoltarea Resurselor Umane 2007-2013, Contract nr. POSDRU/107/1.5/S/76903

REFERINŢE

1. M.M. Bakhtin, Speech Genres & Other Late Essays, University of Texas Press, Austin, 1986.
2. D. Davidov, O. Tsur, A. Rappoport, Semi-upervised Recognition of Sarcastic Sentences in Twitter and Amazon, 14th Conference on Computationl Natural Language Learning, Uppsala, Sweden, 2010.
3. X. Ding, B. Liu, and P. S. Yu, A holistic lexicon-based approach to opinion mining, In Proceedings of the 2008 International Conference on Web Search and Data Mining, Palo Alto, California, USA, 2008.
4. A. Esuli, and F. Sebastani, SENTIWORDNET: A publicly available lexical resource for Opinion Mining, LREC, 2005.
5. M. Forestier, J. Velcin, D. A. Zighed, Extracting Social Networks to Understand Interaction, International Conference on Advances in Social Network Analysis and Mining (ASONAM'11), vol. (2011), pp. 213-219, 2011.
6. K. Ganesan, C. Zhai, and J. Han, Opinosis: a graph-based approach to abstractive summarization of highly redundant opinions, In Proceedings of the 23rd International Conference on Computational Linguistics. COLING '10. Association for Computational Linguistics, Stroudsburg, PA, USA, 2010, pp. 340–348.
7. A. B. Goldberg and X. Zhu, Seeing stars when there aren't many stars: graph-based semi-supervised learning for sentiment categorization, In Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing, 2006.
8. V. Hatzivassiloglou and K. R. McKeown. Predicting the semantic orientation of adjectives. In Proceedings of ACL 1997, pp.174-181, 1997.
9. T. Hofmann, Probabilistic latent semantic indexing, In Proceedings of SIGIR '99, pp. 50-57, 1999.
10. J. Kamps and M. Marx, Words with attitude. In Proceedings of the First International Conference on Global WordNet, pp.332-341, 2002.

11. H.D. Kim, K. Ganesan, P. Sondhi, and C. Zhai, Comprehensive Review Of Opinion Summarization, 2011.
12. L.-W, Ku, Y.-T, Liang, and H.-H, Chen, Opinion extraction, summarization and tracking in news and blog corpora, In AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW), pp. 100–107, 2006.
13. B. Liu, M Liu B, Tutorial on sentiment analysis, based on Chapter 11 of the book "Web Data Mining – Exploring Hyperlinks, Contents and Usage Data", 2007.
14. Y. Lu, D. Huizhong, W. Hongning, and Z. ChengXiang, Exploiting Structured Ontology to Organize Scattered Online Opinions, In Proceedings of Interntional Conference on Computational Linguistics (COLING-2010), 2010.
15. J. B, MacQueen, Some methods for classification and analysis of multivariate observations, In Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 281-297. University of California Press, 1967.
16. M.-C, de Marnee, and C.D. Manning, Stanford Typed Dependencies Manual, http://nlp.stanford.edu/software/ dependencies/manual.pdf, 2008.
17. B. Pang, and L. Lee, Opinion mining and sentiment analysis, Foundations and Trends in Information Retrieval, vol. 2, pp. 1–135, 2008.
18. B. Pang and L. Lee. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. In Proceedingsof ACL, 2004.
19. P.D. Turney. Thumbs up or thumbs down: Semantic orientation applied to unsupervised classification of reviews. In Proceedings of ACL 2002, pp.417-424, 2002
20. P.D, Turney, and M.L. Littman, Measuring praise and criticism: inference of semantic orientation from association, ACM TOIS 21(4), 2003.
21. C. Fellbaum, WordNet: an Electronic Lexical Database, MIT Press, 1998.
22. B. Liu, M. Hu, and J. Cheng. Opinion observer: Analyzing and comparing opinions on the web, In Proceedings of International Conference on World Wide Web (WWW-2005), 2005.
23. B. Liu, Opinion Mining, In Proceedings of the 17th International World Wide Web Conference, Beijing, 2008.