Analiza stărilor emoționale induse de citirea unei știri utilizând Analiza Semantică Latentă

Diana Lupan, Mihai Dascălu, Ștefan Trăușan-Matu, Traian Rebedea, Philippe Dessus, Maryse Bianco

Analiza stărilor emoționale induse de citirea unei știri utilizând Analiza Semantică Latentă

Diana Lupan, Mihai Dascălu, Ștefan Trăușan-Matu, Traian Rebedea

Universitatea Politehnica din București

313 Splaiul Independenței, 060042 București, România

diana.lupan@cti.pub.ro, mihai.dascalu@cs.pub.ro, stefan.trausan@cs.pub.ro, traian.rebedea@cs.pub.ro

Philippe Dessus, Maryse Bianco

Université Pierre-Mendès France

F-38040 Grenoble CEDEX 9, Franța

philippe.dessus@upmf-grenoble.fr

REZUMAT

Emoţiile pot fi identificate atât în comunicarea verbală cât şi în cea scrisă. Dacă în primul caz pot fi identificate mai uşor datorită unor trăsături specifice comunicării verbale (limbajul corpului, tonul vocii sau inflexiuni), în al doilea caz regăsirea acestora poate fi o adevărată provocare. Aşadar, propunem o metodă inedită de analiză automată a emoţiilor transmise prin intermediul comunicării scrise, mai exact, determinarea stării emoţionale a unei persoane în urma citirii unei ştiri. Cu alte cuvinte, scopul nostru este de a determina cum citirea unei ştiri afectează starea emoţională a cititorului şi să ajustăm aceste valori pe baza stării emoţionale curente a acestuia. Dintr-o perspectivă mai tehnică, sistemul dezvoltat (Emo2 – Emotions Monitor) combină o abordare independentă de context (evaluarea efectivă a ştirii utilizând tehnici de prelucrare a limbajului natural) cu influenţele determinate de starea emoţională curentă a utilizatorului. Astfel, scopul metodei propuse este de a obţine o estimare a stării emoţionale finale a utilizatorului cât mai apropiată de cea reală.

Cuvinte cheie

stare emoţională, analiza emoțiilor, analiza semantică latentă, analiza automată de articole de ştiri

Clasificare ACM

I.2.7 Natural Language Processing

INTRODUCERE

Emoţiile sunt unul dintre elementele definitorii ale naturii umane, ele fiind prezente în viaţa de zi cu zi şi aproape în orice context. Acestea adaugă valoare interacţiunilor umane într-o modalitate unică şi efectul indus de ele poate schimba complet înţelesul unui mesaj.

Cu toate că emoţiile pot fi identificate mai uşor în comunicarea faţă în faţă sau verbală, ele pot fi evidenţiate şi în mesajele scrise. Dacă în primul caz sunt exprimate preponderent prin limbajul corpului şi caracteristici specifice vocii (spre exemplu ton, ritm, frecvenţă), în al doilea caz toate aceste trăsături sunt eliminate deoarece canalul de comunicaţie nu le poate suporta. În consecinţă, identificarea emoţiilor în comunicarea scrisă se poate dovedi a fi o adevărată provocare datorită lipsei acestor caracteristici, singurele elemente de analiză rămânând cuvintele, grupate în propoziţii. Cu toate acestea, analiza emoţiilor din comunicarea scrisă poate oferi o înţelegere mai exactă a mesajului transmis şi poate prezice impactul cel mai probabil pe care îl va avea asupra persoanei care îl va citi.

Abordarea noastră urmăreşte analiza automată a modificării stării emoţionale şi a sentimentelor unei persoane în urma citirii unui articol. Ne vom concentra pe analiza unor articole de ştiri scurte formate din titlu şi o scurtă descriere, scrise cu intenţia de a „provoca” emoţii şi de a atrage atenţia cititorului.

În acest articol vom prezenta două abordări implementate în sistemul Emo2 (Emotions Monitor): o abordare independentă de context şi una dependentă de context. Prima abordare presupune evaluarea conţinutului ştirii utilizând diferite metode de prelucrare a limbajului natural, iar a doua ia în considerare starea emoţională a cititorului în momentului citirii articolului.

În secţiunea următoare este prezentată o descriere generală a arhitecturii sistemului, cu detalierea fiecărui modul. Cea de a treia secţiune se axează pe abordarea independentă de context cu cele 2 dimensiuni, semantică şi lexicală. Cea de a patra secţiune prezintă abordarea influenţată de context, iar ultima parte cuprinde rezultatele şi concluziile.

ABORDĂRI SIMILARE

În această secțiune vom prezenta câteva dintre cele mai importante proiecte similare cu subiectul propus. Proiectul UA-ZBSA [9] a fost dezvoltat pentru SemEval 2007 şi îsi propune să clasifice titlurile unor ştiri în 6 categorii, în funcţie de emoţiile pe care fiecare le exprimă: furie, dezgustare, frică, bucurie, tristeţe şi suprindere.

Pornind de la această idee, abordarea noastră cuprinde tot o clasificare după 6 emoții de bază, dar am adaptat lista de mai sus pentru a obține o clasificare cât mai exactă în raport cu subiectul propus. Acestea sunt: în controlul situației, frică, bucurie, tristețe, încântare, plictiseală. (descrise în detaliu în secțiunile următoare).

Clasificarea emoțiilor utilizată in proiectul UA-ZBSA este bazată pe frecvența şi pe apariția în același context a conceptelor. Se aplică ideea conform căreia cuvintele care apar de multe ori într-un context care exprimă o stare emoțională, au o probabilitate mare de a exprima starea emoțională respectivă.

Un sistem similar UPAR7 [10] a fost proiectat pornind de la ideea că este foarte probabil ca majoritatea conceptelor din titlul unei știri să exprime emoții. Obiectivul analizei

este de a identifica fragmentul care cuprinde mesajul deoarece are cea mai mare relevanță.

Totodată, s-au dezvoltat reguli pentru detecția unor anumitor tipuri de emoții (negația poate fi un indicator pentru surpriză).

Sistemul anterior a integrat SS-Tagger, Stanford Parser și utilizează WordNet, SentiWordNet și WordNet-Affect (resurse lexicale).

ARHITECTURA SISTEMULUI

Sistemul dezvoltat este format din mai multe module grupate pe nivele şi este reprezentat în Figura 1.

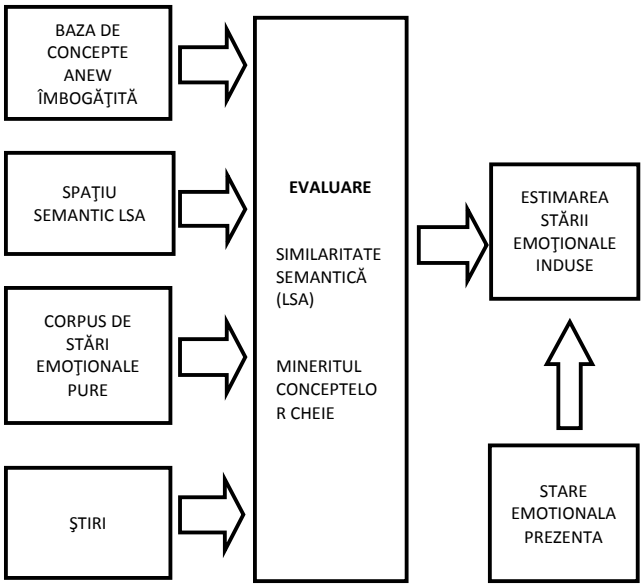

Figura 1. Arhitectura sistemului

Prima etapă constă în determinarea stării emoţionale curente a utilizatorului. Aşadar, aplicaţia noastră încearcă să o determine cât mai exact prin analizarea unui set de 5 ştiri care sunt evaluate în mod explicit de către utilizator. Aceste informaţii sunt utilizate într-o etapă ulterioară pentru a adapta şi personaliza rezultatele obţinute de sistem astfel încât acestea să reflecte cât mai exact impactul fiecărei ştiri asupra stării emoţionale curente a utilizatorului.

În termeni mai tehnici, analiza efectuată este divizată în 2 etape independente: o evaluare iniţială obiectivă, în cadrul modulului Evaluare, şi o evaluare subiectivă, adaptivă, de personalizare a rezultatelor luând în considerare starea emoţională curentă a utilizatorului (integrată în componenta Estimarea Stării Emoţionale Induse).

Metoda de evaluare propusă utilizează mai multe date de intrare: baza de concepte ANEW îmbogăţită, un spaţiu semantic latent de vectori (LSA), un corpus de articole care exprimă o stare emoţională pură şi ştirile în curs de evaluare. Baza de concepete ANEW îmbogăţită [1] este formată din concepte (cuvinte) care au asociate 3 valori ce sugerează starea emoţională transmisă raportată la 6 stări etalon (Tristeţe / Supărare, Plictiseală / Încântare, Controlat / În control). Prima dimensiune, de Tristeţe / Supărare exprimă aşa cum sugerează şi denumirile, o stare de fericire sau de supărare. Dimensiunea Plictiseală / Încântare reflectă gradul de interes acordat de către cititor informaţiei prezentate în articol, mai exact dacă are intenţia de a continua să citească despre subiectul prezentat sau nu. Ultima dimensiune redă starea de a te simţi în controlul situaţiei prezentate în cadrul ştirii versus a fi controlat de evenimentele (sau persoanele) descrise în articol. Aceste valori sunt în intervalul [1; 9], unde 1 este limita inferioară (polul negativ al fiecărei axe, cea mai tristă stare emoţională pentru axa Bucurie / Tristeţe) iar 9 cea superioară (polul pozitiv al fiecărei axe, bucurie, pentru axa Bucurie / Tristeţe). Baza de concepte iniţială (aproximativ 1000 de cuvinte) a fost îmbogăţită utilizând un algoritm format din 2 etape, prezentat în detaliu în secţiunea următoare.

Procesul de evaluare cuprinde 2 abordări: determinarea similarităţii semantice între texte (conţinutul ştirilor fiind comparat cu setul de documente ce exprimă o stare emoţională pură) şi mineritul cuvintelor cheie. Cea de a doua abordare constă în extragerea celor mai importante concepte din coţinutul şi titlul unei ştiri şi în combinarea valorilor lor "emoţionale" cu scopul de a obţine sentimentele induse per total de acestea. Prin urmare, setul de documente ce induc o stare emoţională pură (Tristeţe / Supărare, Plictiseală / Încântare, Controlat / În control) este utilizat pentru a obţine dimensiunea semantică a analizei, iar identificarea conceptelor cheie oferă dimensiunea lexicală. Mineritul conceptelor cheie este efectuat separat pentru titlul ştirii şi pentru conţinut deoarece se consideră că titlul are o încărcătură emoţională mai bogată decât restul ştirii. Titlul unei ştiri este construit pentru a fi scurt şi pentru a atrage atenţia cititorilor astfel încât să o citească integral; în alte cuvinte, se consideră că titlul exprimă, din perspectiva unui cititor, esenţa şi constituie o sinteză a conţinutului efectiv. Prin urmare, datorită proprietăţilor lor intrinseci, cuvintele cheie din titlu vor avea o contribuţie mai mare în formulele aplicate decât cele extrase din restul ştirii.

În evaluarea stării emoţionale induse utilizând analiza semantică latentă (LSA), rezultatul este determinat utilizând media valorilor documentelor similare cu conţinutul unei ştiri, ponderate de valoarea efectivă a similarităţii, astfel încât documentele mai similare să deţină un procentaj mai mare. Luând în considerare similiaritatea dintre conceptele care induc o anumită stare, se poate deduce că documentele cu o structură similară a conceptelor cheie determină o stare emoţională similară. Afirmaţia anterioară nu este validă în toate cazurile, însă atunci când se compară articole de ştiri dintr-un anumit domeniu, într-un interval scurt de timp, rezultatele obţinute au fost destul de precise. Cu alte cuvinte, ştirile din aceleaşi domeniu, transmit rareori mesaje care induc emoţii diferite într-un interval scurt de timp.

Valorile obţinute utilizând cele 2 abordări sunt combinate astfel încât rezultatele finale să aproximeze cât mai bine starea emoţională cea mai probabilă a unui cititor (iniţial cu o stare neutră) după citirea unei ştiri. Modulul „Stare emoţională prezentă" integrează informaţiile oferite de utilizator cu restul analizei prin estimarea stării sale emoţionale curente. În acest scop, utilizatorul trebuie să completeze un chestionar şi să evalueze un set de 5 ştiri. În experimentele efectuate, articolele de ştiri reprezintă intrarea principală şi sunt formate din titlu şi 2-3 fraze din

corpul ştirii. Pentru testare am utilizat ştiri extrase din fluxurile RSS aparţinând CNN.com.

PRIMA ETAPĂ: ABORDAREA INDEPENDENTĂ DE CONTEXT

Abordarea independentă de context cuprinde 2 etape diferite: prima determină similaritatea dintre un articol şi un set de articole predefinite care exprimă o stare emoţională pură, iar a doua compară conceptele extrase din textul analizat cu termenii din baza de date ANEW îmbogăţită. Fiecare termen are asociat 3 conotaţii care exprimă starea emoţională posibil indusă în momentul citirii articolul de o persoană.

Abordarea lexicală

Abordarea lexicală utilizează o bază de date pentru stocarea valorilor emoţionale asociate cu fiecare concept exprimate în 3 dimensiuni: Tristeţe / Supărare, Plictiseală / Încântare, Controlat / În control. Baza de date utilizată este generată printr-un proces de îmbogăţire care este descris în detaliu în paragrafele următoare.

Iniţial, numărul de termeni stocaţi în baza de concepte a fost de 1000 şi valorile asociate fiecărui concept sunt obţinute din proiectul ANEW [1]. În cadrul acestuia s-a dezvoltat un set de reguli de clasificare a unui număr semnificativ de cuvinte din limba engleză în funcţie de emoţiile pe care le induc. Baza de date asociată ANEW este îmbogăţită prin utilizarea unor tehnici de prelucrare a limbajului natural (WordNet şi analiza semantică latentă) pentru determinarea similarităţii dintre un concept considerat "cunoscut" şi unul "nou". Mai exact, am combinat metricile de similaritate dintre o ontologie lexicalizată cu cosinusul dintre conceptele reprezentate într-un spaţiu semantic latent de vectori.

Procesul efectiv de îmbogăţire a constat din 2 etape. Prima a presupus selectarea în mod aleatoriu a unui cuvânt şi determinarea sinonimelor utilizând *synset*-urile din WordNet, iar în a doua etapă sinonimele au fost determinate utilizându-se analiza semantică latentă [3]. Luând în considerare similaritatea semantică determinată de cosinus-ul dintre 2 concepte, putem introduce noi termeni în baza de date utilizând termeni deja cunoscuţi (care se află în baza de date) şi mai exact, valorile asociate cu aceştia şi similarităţile cu conceptele ce se doresc a fi adăugate. În plus, doar conceptele "cunoscute" care au un grad de similaritate peste o anumită limită sunt considerate sinonime şi sunt incluse în calcule. Deci, utilizând sinonimele identificate prin WordNet a conceptelor din baza de date ANEW şi luând în considerare doar cele mai similare k concepte, determinăm valorile associate termenului nou şi îl introducem în baza de date îmbogăţită. După rularea mai multor teste cu valori incrementale pentru k, s-a concluzionat să alegem pentru experimentele finale k=3. Aşadar vom alege cele mai dominante şi similare 3 concepte "cunoscute".

Abordarea semantică

Aşa cum a fost menţionat anterior, abordarea semantică constă în determinarea similarităţilor dintre conţinutul unui articol de ştiri şi un corpus de documente care exprimă o stare emoţională pură. Cu toate că noţiunea de "puritate" este subiectivă şi depinde de evaluarea efectuată de fiecare persoană în parte, documentele selectate în acest proiect exprimă o dimensiune dominantă, identificată şi acceptată unanim de mai mulţi evaluatori umani. Corpusul utilizat în experimentele efectuate este format din 10 documente asociate fiecărei stări emoţionale, selectate dintr-un domeniu specific şi predefinit (ştiri de interes general). De exemplu, următorul text sugerează un sentiment puternic de fericire:

"A small school was presented with a check for one hundred thousand dollars during an unforgettable assembly celebrating recycling. The students at Pascal school performed beyond expectations collecting many recyclable containers. They beat out every other school in the country to win the title and the generous check. The school plans to use the money on green projects like converting their computer center to run on solar energy."

Aşa cum s-a menţionat anterior, aceste texte au fost alese dintr-un domeniu de interes general pentru a cuprinde o gamă cât mai largă de emoţii. Algoritmul utilizat pentru calcularea similarităţii semantice dintre un articol de ştiri şi document ce exprimă o stare emoţională pură este analiza semantică latentă, descrisă în secţiunea următoare.

Analiza Semantică Latentă

Principala tehnică de prelucrare a limbajului natural din acest proiect este analiza semantică latentă, aceasta fiind utilizată atât în procesul de îmbogăţire a bazei de date, cât şi în abordarea semantică. Analiza semantică latentă [2] [3] este o metodă de determinare a similarităţii dintre cuvinte şi fraze prin analiza unui corpus de texte extins. Ideea fundamentală din spatele algoritmului este că totalitatea contextelor în care un cuvânt apare sau nu, oferă o modalitate de determinare a similarităţii dintre sensurile cuvintelor. Analiza semantică latentă creează o reprezentare vectorială a textelor din corpus şi determină similaritatea dintre perechile de texte prin compararea reprezentărilor vectoriale asociate într-un spaţiu de vectori de dimensiune mai mică, alcătuit din componentele principale cu cele mai mari valori singulare, numit spaţiu semantic latent. Principalul motiv pentru care utilizăm acest alogoritm este că s-a dovedit că aproximează bine comportamentul uman [4].

Prin urmare, analiza semantică latentă poate fi folosită cu succes pentru a extinde datele obţinute în urma experimentelor cu subiecţi umani (subiecţii au evaluat diferite cuvinte după 3 dimensiuni: plăcere ,încântare şi dominanţă) şi pentru a estima modul în care statea emoţională este afectată după citirea unei ştiri [6]. Algoritmul utilizează un corpus extins de texte, deci primul pas a fost colectarea unui număr mare de date şi separarea lor în documente. Motivul principal pentru care este necesar acest pas este datorită abordării de tip "sac de cuvinte" ("bag of words") pe care o are algoritmul, însemnând că o propoziţie este reprezentată fără a se ţine cont de reguli de gramatică sau de ordinea cuvintelor [7].

În general, corpusul trebuie să conţină texte relevante şi similare ca formă şi structură subiectului analizei deoarece un corpus specific va conţine un procentaj mai mare de cuvinte relevante şi nu va irosi din "puterea sa de reprezentare" pe cuvinte care, cel mai probabil, nu vor exista în contextul aplicaţiei. În cadrul acestui proiect, am utilizat corpusul de ştiri Reuters [8].

Prin urmare, referitor la procesul iniţial de învăţare al algoritmului, fişierele originale au fost procesate pentru a respecta condiţiile de bază impuse de acesta. În primul rând, fiecare document trebuie să conţină între 50 şi 100 de cuvinte. Mai jos sunt regulile care s-au utilizat pentru divizarea fişierelor iniţiale:

- propoziţiile nu pot fi fragmentate;
- este recomandat ca fiecare paragraf să aparţină unui singur document. Dacă nu este posibil, se aplică prima regulă;
- fiecare document ar trebui să fie împărţit în părţi aproximativ egale;
- documentele iniţiale mai mici decât o limită sunt eliminate (limita de cuvinte < 50);
- părţile rezultate din divizarea unui document care sunt mai mici decât 50 de cuvinte sunt combinate pentru a se îndeplini numărul minim de cuvinte.

Documentele astfel obţinute sunt prelucrate astfel încât zgometele să fie eliminate (numere, semne de punctuaţie, separatori). În final, fiecare document conţine doar cuvinte cu litere mici (doar litere) separate prin spaţii (“ ”).

Evaluarea articolelor

Rezultatul final este calculat prin combinarea valorilor celor două metode (semantică şi lexicală). În primul rând conceptele din titlu sunt considerate a fi mai încărcate emoţional decât cele din conţinut, deci acestea au o pondere mai mare. Formula evaluării din perspectiva abordării lexicale este:

$$valoare_{concepte-cheie} = p * valoareTitlu + (1-p) * valoareConţinut \quad (1)$$

unde $valoareTitlu$ şi $valoareConţinut$ sunt o combinaţie a valorilor conceptelor cheie în care toate conceptele au ponderi egale.

Valoarea lui p a fost determinată experimental folosind mai multe iteraţii în intervalul [0,55; 0,9] cu un increment de 0,05. Prin aceste experimente s-a descoperit că valoarea 0,6 este cea care combină cel mai bine mesajul exprimat în titlu cu cel exprimat în conţinut.

Conţinutul unui articol este considerat a fi similar cu un text din setul de documente dacă depăşeşte un prag de 0,2 (similaritatea ia valori între -1 şi 1). Pentru fiecare stare emoţională se calculează media aritmetică între valorile care îndeplinesc această condiţie. Pasul următor este reprezentat de normalizarea valorilor pentru a le încadra în intervalul [1; 9]. Rezultatul final este calculat folosind următoarea formulă, care a fost ponderată pentru a accentua valoarea obţinută din analiza conceptelor cheie în defavoarea rezultatelor din analiza semantică (aceasta fiind limitată intrinsec de abordarea de tip sac de cuvinte utilizata):

$$rezultatFinal = p * valoare_{concepte-cheie} + (1-p) valoare_{setDeDocumente} \quad (2)$$

Valoarea lui p a fost determinată experimental testând valori tot din intervalul [0,55; 0,9]. Experimentele au determinat că p=0,65 determină combinarea cea mai favorabilă a valorilor asociate cuvintelor cheie cu a similarităţii semantice dintre ştiri şi documentele aferente stărilor emoţionale pure.

A DOUA ETAPĂ: ABORDAREA DEPENDENTĂ DE CONTEXT

Abordarea dependentă de context reprezintă o ajustare a valorilor obţinute în prima etapă a proiectului (independentă de context) şi este utilizată pentru personalizarea rezultatelor obţinute. Am pornit de la ideea că starea emoţională a unei persoane va fi influenţată diferit de o ştire depinzând de starea sa curentă şi de trăsăturile specifice asociate personalităţii sale. Spre exemplu, o persoană care este iniţial fericită şi citeşte o ştire tristă va fi mai puţin afectată şi deci mai puţin tristă decât o persoană care era iniţial tristă. Mai mult, efectele citirii unei ştiri depind şi de cât de aproape cititorul este din punct de vedere geografic de evenimentele prezentate.

Luând în considerare aceste observaţii, au fost introduse câteva elemente care evaluează aceste caracteristici. În primul rând, starea emoţională curentă este estimată prin evaluarea de către utilizator a unui set de 5 ştiri care sunt selectate astfel încât să fie cât mai similare cu ştirile care vor fi analizate automat de către aplicaţie. De asemenea, utilizatorul va trebui să completeze un chestionar referitor la locaţia sa geografică prezentă. Cele 5 ştiri din setul iniţial sunt determinate prin sortarea tuturor ştirilor din baza de date după similarităţile cu ştirile principale şi selectarea primelor 5 cele mai similare.

Pentru a evalua o ştire, utilizatorul are la dispoziţie 2 opţiuni pentru fiecare dintre dimensiuni. Aceste valori sunt combinate în funcţie de similaritatea dintre fiecare ştire de evaluat (din setul de 5) cu ştirile principale (din setul de 10), astfel încât cu cât similaritatea cu o ştire principală este mai mare, cu atât valoarea corespunzătoare ştirii de evaluat va avea o proporţie mai mare în rezultatul final.

Rezultatele obţinute în această etapă sunt combinate cu cele obţinute în etapa precedentă (independentă de context) în proporţii egale. Interfaţa grafică disponibilă utilizatorului a fost îmbogăţită fiind introduse opţiunile descrise mai sus. Toate cele 5 ştiri trebuie să fie evaluate pentru a se obţine cele mai exacte rezultate.

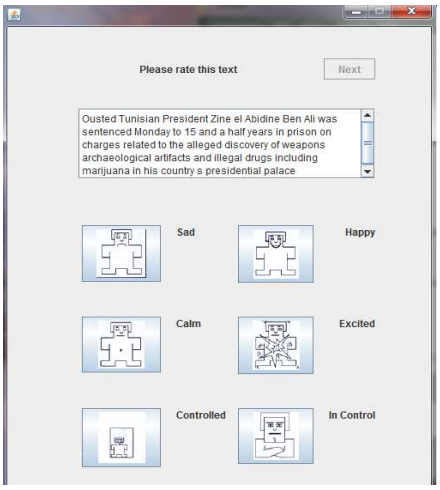

Figura 2. Interfaţa grafică pentru evaluarea iniţială a ştirilor

Locaţia curentă a utilizatorului este folosită pentru corectarea rezultatelor, urmărind idea că evenimentele

care se desfăşoară în apropierea unei persoane au o probabilitatea mai mare de a-l afecta (influenţa). Pentru determinarea acesteia, se testează dacă conceptele din conţinutul ştirii se află pe continentul introdus de utilizator. În caz afirmativ, rezultatele sunt modificate astfel: valorile mai mari decât 5 (valoarea mediană a domeniului de definiţie) sunt crescute, iar cele mai mici sunt reduse şi mai mult.

Pentru testarea acestor aspect referitoare la locaţie am utilizat serviciul web GeoNames care oferă acces la o bază de date geografică ce conţine peste 8 milioane de denumiri de locaţii şi funcţii dintre ele, pentru a putea testa legăturile dintre ele.

REZULTATE

Aşa cum a fost menţionat în secţiunile anterioare, sistemul proiectat a fost testat utilizând feed-uri RSS aparţinând CNN.com. Pentru a vizualiza rezultatele aplicaţiei, utilizatorul are la dispoziţie o fereastră care afişează o ştire şi 3 grafice ce exprimă cea mai probabilă stare emoţională indusă de ştire. Fiecare grafic este poziţionat între 2 reprezentări a câte o stare emoţională pură pentru o reprezentare cât mai sugestivă a rezultatelor. Figura 3 reprezintă o mostră a rezultatelor aplicaţiei.

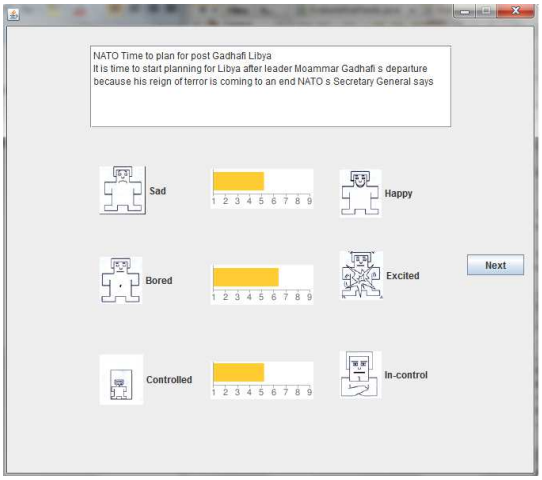

Figura 3. Interfaţa grafică reprezentând rezultatele aplicaţiei

Pentru o reprezentare de ansamblu se poate genera şi următorul grafic radial corespunzător celor 3 dimensiuni. De exemplu, pornind de la valorile: fericit/trist = 4, încântat/plictisit = 7 şi controlat/în-control = 8 se obţine graficul din Figura 4.

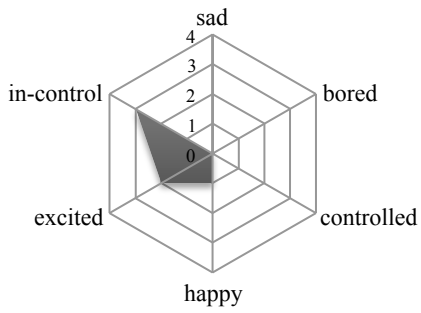

Figura 4. Exemplu de graf generat

Pentru a valida rezultatele obţinute, valorile din prima şi a doua etapă a proiectului au fost comparate cu valori obţinute în urma unui sondaj care a inclus un număr de 10 participanţi (5 femei şi 5 bărbaţi) cu vârste cuprinse între 20 – 25 de ani din domenii diverse (calculatoare, respectiv medical). Sondajul a conţinut 2 etape, fiecare având 5 ştiri pentru evaluarea iniţială (aşa cum a fost menţionat anterior aceste ştiri sunt selectate astfel încât să fie cât mai similare cu ştirile evaluate automat de aplicaţie) pentru a se determina starea emoţională curentă a participantului, o întrebare referitoare la poziţia sa geografică şi un alt set de 10 ştiri. Prin urmare setul de date a conţinut 20 de elemente din fluxul RSS (reprezentate de ştirile principale care sunt evaluate automat de aplicaţie). În Figura 5 sunt prezentate rezultatele sintetice obtinute in urma sondajului:

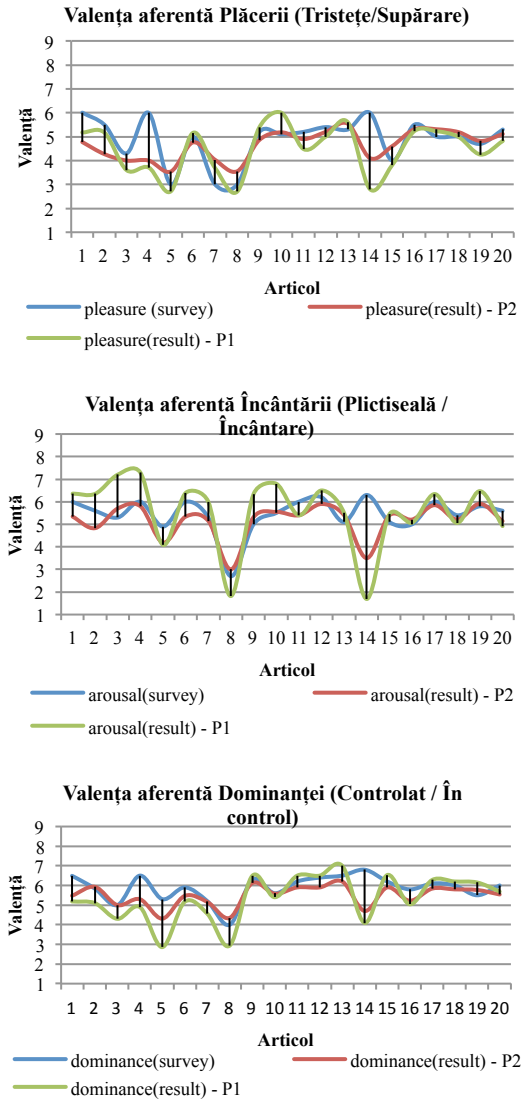

Figura 5. Valenţele Plăcere, Încântare, Dominanţă; P1 reprezintă abordarea independentă de context; P2 reprezintă abordarea dependentă de context

Fiecare grafic este asociat uneia dintre cele 3 dimensiuni ale analizei (valenţele Plăcere, Încântare, Dominanţă). Axa verticală corespunde cu valorile efective (rezultatele) în intervalul [1; 9], iar axa orizontală reprezintă articolele de ştiri analizate, în număr de 20. Corelaţiile dintre cele 2

abordări şi sondajul, pentru toate dimensiunile sunt prezentate în Tabelul 1.

Tabelul 1. Corelaţiile dintre rezultatele aplicaţiei vs. sondaj

	Plăcere	Încântare	Dominanţă	Media pe cele 3 dimensiuni
Sondaj-P1	0,55	0,48	0,61	0,55
Sondaj – P2	0,54	0,58	0,56	0,56
P1 – P2	0,86	0,91	0,96	0,91

Aşa cum poate fi dedus din tabelul anterior, valenţa Plăcere obţine corelaţii proape egale între cele 2 abordări şi mediile corelaţiilor pe cele 3 dimensiuni sunt foarte similare. Dar, luând în considerare natura fiecărei dimensiuni, putem extrapola din experimentul efectuat că dimensiunea Încântare este mai dependentă de context, în sensul că poate varia mai mult decât celelate dimensiuni depinzând de stările emoţionale anterioare, în timp ce dimensiunea Dominanţă este mai mult influenţată de contextul actual şi de conţinutul ştirii. În plus, am obţinut corelaţii foarte mari între cele 2 abordări, ceea ce a fost anticipat datorită faptului că P2 rafinează rezultatele iniţiale (P1).

Totuşi, există situaţii în care a doua abordare obţine rezultate mai inexacte în comparaţie cu sondajul (ce exprimă starea emoţională curentă a utilizatorului) şi faţă de cealaltă abordare. În afară de subiectivitatea ce apare atunci când o persoană evaluează propria sa stare emoţională, o posibilă explicaţie poate fi faptul că ştirile din chestionar nu au fost suficient de similare cu cele evaluate în mod automat de aplicaţie; astfel, în loc să ajusteze valorile din prima etapă pentru a fi mai aproape de realitate, algoritmul este indus în eroare şi le modifică incorect, determinând ca rezultatele finale să fie mai slabe decât cele obţinute în prima etapă. O posibilă soluţie la această problemă presupune îmbogăţirea bazei de ştiri cu mai multe texte pentru ca ştirile noi să fie evaluate în funcţie de ştiri cu un grad de similaritate mai mare.

Cu toate că locaţia introduce un grad de precizie mai mare în cadrul analizei noastre, ne propunem să adăugăm şi alte elemente în versiunile viitoare ale sistemului: sex, vârstă şi alţi factori similari care pot fi utilizaţi pentru identificarea diferitelor categorii de utilizatori.

CONCLUZII

Scopul nostru a presupus dezvoltarea unei aplicaţii care evaluează modul în care este afectată starea emoţională a unei persoane după citirea unei ştiri. În esenţă abordarea utilizată a constat în 2 etape: una dependentă şi una independentă de context. Am utilizat tehnici de prelucrare a limbajului natural, din care cea principală a fost analiza semantică latentă. Rezultatele au fost validate prin compararea cu un sondaj şi acesta a demonstrat că putem considera metoda utilizată promiţătoare şi relevantă, într-un context în care subiectivitatea joacă un rol important. Totuşi există şi aspecte care pot fi îmbunătăţite, prin urmare plănuim să efectuăm o analiză mai amănunţită a textelor scrise şi să modelăm cât mai exact evoluţia stării emoţionale a unei persoane.

Drept direcţii viitoare de cercetare plănuim să efectuăm sondaje mai complexe, cu un număr mai ridicat de fluxuri RSS şi un număr mai mare de persoane astfel încât feedback-ul obţinut să fie cât mai explicit şi exact. Totodată îmbogăţirea corpusului de documente ce exprimă o stare emoţională pură şi includerea mai multor factori psihologici în modelul nostru adaptiv ar putea conduce la rezultate mai bune.

REFERINŢE

1. Bradley, M. M., & Lang, P. J.. *Affective norms for English words (ANEW): Stimuli, instruction manual and affective ratings*. Gainesville (FL): The Center for Research in Psychophysiology, University of Florida, Tech. Report 1999.
2. Foltz, P. W.. Latent semantic analysis for text-based research. *Behavior Research Methods, Instruments and Computers, 28*(2), 197–202 1996.
3. Landauer, T. K., & Dumais, S. T.. A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. *Psychological Review, 104*(2), 211–240 1997.
4. Landauer, T. K., Foltz, P. W., & Laham, D.. An introduction to Latent Semantic Analysis. *Discourse Processes, 25*(2/3), 259–284 1998.
5. Manning, C., & Schütze, H.. *Foundations of statistical Natural Language Processing*. Cambridge (Mass.): MIT Press 1999.
6. Wang, L. & Wan, Y.. Sentiment classification of documents based on Latent Semantic Analysis. In S. Lin & X. Huang (Eds.) *Advanced Research on Computer Education, Simulation And Modeling Communications in Computer and Information Science* (CESM 2011, Vol. 176, pp. 356–361). New York: Springer 2011.
7. Wiemer-Hastings, P. (1999). How latent is Latent Semantic Analysis? *Proc. 16th International Joint Conference on Artificial intelligence (IJCAI'99)* – Volume 2.
8. Reuters-21578 Text Categorization Collection http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html 2012.
9. Kozareva, Z., Navarro, B., Vazquez, S., Montoyo, A.. UA-ZBSA: A Headline Emotion Classification through Web Information. In Proceeding of the 4th International Workshop on Semantic Evaluations SemEval '07, pg 334-337, 2007.
10. Chaumartin, F.-R.. UPAR7: A knowledge-based system for headline sentiment tagging. In Proceeding of the 4th International Workshop on Semantic Evaluations SemEval '07, pg 422-425, 2007.