

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

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

Cuvinte cheie

stare emoțională, analiza emoțiilor, analiza semantică latentă, analiza automată de articole de știri

Clasificare ACM

I.2.7 Natural Language Processing

INTRODUCERE

Emoțiile sunt unul dintre elementele definitorii ale naturii umane, ele fiind prezente în viața de zi cu zi și aproape în orice context. Acestea adaugă valoare interacțiunilor umane într-o modalitate unică și efectul indus de ele poate schimba complet înțelesul unui mesaj.

Cu toate că emoțiile pot fi identificate mai ușor în comunicarea față în față sau verbală, ele pot fi evidențiate și în mesajele scrise. Dacă în primul caz sunt exprimate preponderent prin limbajul corpului și caracteristici specifice vocii (spre exemplu ton, ritm, frecvență), în al doilea caz toate aceste trăsături sunt eliminate deoarece canalul de comunicație nu le poate suporta. În consecință, identificarea emoțiilor în comunicarea scrisă se poate dovedi a fi o adevărată provocare datorită lipsei acestor caracteristici, singurele elemente de analiză rămânând cuvintele, grupate în propoziții. Cu toate acestea, analiza emoțiilor din comunicarea scrisă poate oferi o înțelegere

mai exactă a mesajului transmis și poate prezice impactul cel mai probabil pe care îl va avea asupra persoanei care îl va citi.

Abordarea noastră urmărește analiza automată a modificării stării emoționale și a sentimentelor unei persoane în urma citirii unui articol. Ne vom concentra pe analiza unor articole de știri scurte formate din titlu și o scurtă descriere, scrise cu intenția de a „provoca” emoții și de a atrage atenția cititorului.

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

În secțiunea următoare este prezentată o descriere generală a arhitecturii sistemului, cu detalierea fiecărui modul. Cea de a treia secțiune se axează pe abordarea independentă de context cu cele 2 dimensiuni, semantică și lexicală. Cea de a patra secțiune prezintă abordarea influențată de context, iar ultima parte cuprinde rezultatele și 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 și își propune să clasifice titlurile unor știri în 6 categorii, în funcție de emoțiile pe care fiecare le exprimă: furie, deznădăjire, frică, bucurie, tristețe și surprindere.

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 în proiectul UA-ZBSA este bazată pe frecvența și 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 și este reprezentat în Figura 1.

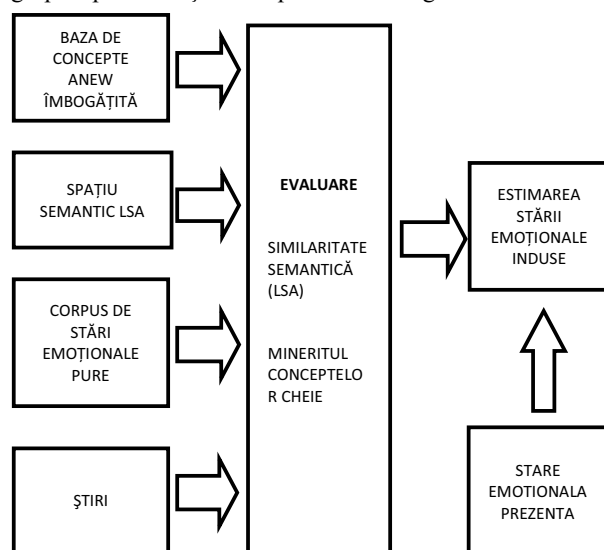


Figura 1. Arhitectura sistemului

Prima etapă constă în determinarea stării emoționale curente a utilizatorului. Așadar, aplicația noastră încearcă să o determine cât mai exact prin analizarea unui set de 5 știri care sunt evaluate în mod explicit de către utilizator. Aceste informații sunt utilizate într-o etapă ulterioară pentru a adapta și personaliza rezultatele obținute de sistem astfel încât acestea să reflecte cât mai exact impactul fiecărei știri asupra stării emoționale curente a utilizatorului.

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

Metoda de evaluare propusă utilizează mai multe date de intrare: baza de concepte ANEW îmbogățită, un spațiu semantic latent de vectori (LSA), un corpus de articole care exprimă o stare emoțională pură și știrile în curs de evaluare. Baza de concepte ANEW îmbogățită [1] este formată din concepte (cuvinte) care au asociate 3 valori ce sugerează starea emoțională transmisă raportată la 6 stări etalon (Tristețe / Supărare, Plictiseală / Încântare, Controlat / În control). Prima dimensiune, de Tristețe / Supărare exprimă așa cum sugerează și denumirile, o stare de fericire sau de supărare. Dimensiunea Plictiseală / Încântare reflectă gradul de interes acordat de către cititor informației prezentate în articol, mai exact dacă are

intenția de a continua să citească despre subiectul prezentat sau nu. Ultima dimensiune redă starea de a te simți în controlul situației prezentate în cadrul știrii 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 emoțională pentru axa Bucurie / Tristețe) iar 9 cea superioară (polul pozitiv al fiecărei axe, bucurie, pentru axa Bucurie / Tristețe). Baza de concepte inițială (aproximativ 1000 de cuvinte) a fost îmbogățită utilizând un algoritm format din 2 etape, prezentat în detaliu în secțiunea următoare.

Procesul de evaluare cuprinde 2 abordări: determinarea similarității semantice între texte (conținutul știrilor fiind comparat cu setul de documente ce exprimă o stare emoțională pură) și mineritul cuvintelor cheie. Cea de a doua abordare constă în extragerea celor mai importante concepte din conținutul și titlul unei știri și în combinarea valorilor lor "emoționale" cu scopul de a obține sentimentele induse per total de acestea. Prin urmare, setul de documente ce induc o stare emoțională pură (Tristețe / Supărare, Plictiseală / Încântare, Controlat / În control) este utilizat pentru a obține dimensiunea semantică a analizei, iar identificarea conceptelor cheie oferă dimensiunea lexicală. Mineritul conceptelor cheie este efectuat separat pentru titlul știrii și pentru conținut deoarece se consideră că titlul are o încărcătură emoțională mai bogată decât restul știrii. Titlul unei știri este construit pentru a fi scurt și pentru a atrage atenția cititorilor astfel încât să o citească integral; în alte cuvinte, se consideră că titlul exprimă, din perspectiva unui cititor, esența și constituie o sinteză a conținutului efectiv. Prin urmare, datorită proprietăților lor intrinseci, cuvintele cheie din titlu vor avea o contribuție mai mare în formulele aplicate decât cele extrase din restul știrii.

În evaluarea stării emoționale induse utilizând analiza semantică latentă (LSA), rezultatul este determinat utilizând media valorilor documentelor similare cu conținutul unei știri, ponderate de valoarea efectivă a similarității, astfel încât documentele mai similare să dețină un procentaj mai mare. Luând în considerare similaritatea dintre conceptele care induc o anumită stare, se poate deduce că documentele cu o structură similară a conceptelor cheie determină o stare emoțională similară. Afirmatia anterioară nu este validă în toate cazurile, însă atunci când se compară articole de știri dintr-un anumit domeniu, într-un interval scurt de timp, rezultatele obținute au fost destul de precise. Cu alte cuvinte, știrile din aceeași domeniu, transmit rareori mesaje care induc emoții diferite într-un interval scurt de timp.

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

corpul știrii. Pentru testare am utilizat știri extrase din fluxurile RSS aparținând CNN.com.

PRIMA ETAPĂ: ABORDAREA INDEPENDENTĂ DE CONTEXT

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

Abordarea lexicală

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

Inițial, numărul de termeni stocați în baza de concepte a fost de 1000 și valorile asociate fiecărui concept sunt obținute 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 funcție de emoțiile pe care le induc. Baza de date asociată ANEW este îmbogățită prin utilizarea unor tehnici de prelucrare a limbajului natural (WordNet și analiza semantică latentă) pentru determinarea similarității dintre un concept considerat “cunoscut” și unul “nou”. Mai exact, am combinat metricile de similaritate dintre o ontologie lexicalizată cu cosinusul dintre conceptele reprezentate într-un spațiu semantic latent de vectori.

Procesul efectiv de îmbogățire a constat din 2 etape. Prima a presupus selectarea în mod aleatoriu a unui cuvânt și 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 cunoscuți (care se află în baza de date) și mai exact, valorile asociate cu aceștia și similaritățile 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 și sunt incluse în calcule. Deci, utilizând sinonimele identificate prin WordNet a conceptelor din baza de date ANEW și luând în considerare doar cele mai similare k concepte, determinăm valorile asociate termenului nou și îl introducem în baza de date îmbogățită. După rularea mai multor teste cu valori incrementale pentru k , s-a concluzionat să alegem pentru experimentele finale $k=3$. Așadar vom alege cele mai dominante și similare 3 concepte “cunoscute”.

Abordarea semantică

Așa cum a fost menționat anterior, abordarea semantică constă în determinarea similarităților dintre conținutul unui articol de știri și un corpus de documente care exprimă o stare emoțională pură. Cu toate că noțiunea de “puritate” este subiectivă și depinde de evaluarea efectuată

de fiecare persoană în parte, documentele selectate în acest proiect exprimă o dimensiune dominantă, identificată și acceptată unanim de mai mulți evaluatori umani. Corpusul utilizat în experimentele efectuate este format din 10 documente asociate fiecărei stări emoționale, selectate dintr-un domeniu specific și predefinit (știri 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.”

Așa cum s-a menționat anterior, aceste texte au fost alese dintr-un domeniu de interes general pentru a cuprinde o gamă cât mai largă de emoții. Algoritmul utilizat pentru calcularea similarității semantice dintre un articol de știri și document ce exprimă o stare emoțională pură este analiza semantică latentă, descrisă în secțiunea 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 îmbogățire a bazei de date, cât și în abordarea semantică. Analiza semantică latentă [2] [3] este o metodă de determinare a similarității dintre cuvinte și 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 similarității dintre sensurile cuvintelor. Analiza semantică latentă creează o reprezentare vectorială a textelor din corpus și determină similaritatea dintre perechile de texte prin compararea reprezentărilor vectoriale asociate într-un spațiu de vectori de dimensiune mai mică, alcătuit din componentele principale cu cele mai mari valori singulare, numit spațiu semantic latent. Principalul motiv pentru care utilizăm acest algoritm 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 obținute în urma experimentelor cu subiecți umani (subiecții au evaluat diferite cuvinte după 3 dimensiuni: plăcere ,încântare și dominantă) și pentru a estima modul în care starea emoțională este afectată după citirea unei știri [6]. Algoritmul utilizează un corpus extins de texte, deci primul pas a fost colectarea unui număr mare de date și 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 propoziție este reprezentată fără a se ține cont de reguli de gramatică sau de ordinea cuvintelor [7].

În general, corpusul trebuie să conțină texte relevante și similare ca formă și structură subiectului analizei deoarece un corpus specific va conține un procentaj mai mare de cuvinte relevante și nu va irosi din “puterea sa de reprezentare” pe cuvinte care, cel mai probabil, nu vor exista în contextul aplicației. În cadrul acestui proiect, am utilizat corpusul de știri Reuters [8].

Prin urmare, referitor la procesul inițial de învățare al algoritmului, fișierele originale au fost procesate pentru a respecta condițiile de bază impuse de acesta. În primul rând, fiecare document trebuie să conțină între 50 și 100 de cuvinte. Mai jos sunt regulile care s-au utilizat pentru divizarea fișierelor inițiale:

- propozițiile nu pot fi fragmentate;
- este recomandat ca fiecare paragraf să aparțină unui singur document. Dacă nu este posibil, se aplică prima regulă;
- fiecare document ar trebui să fie împărțit în părți aproximativ egale;
- documentele inițiale mai mici decât o limită sunt eliminate (limita de cuvinte < 50);
- părțile 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 obținute sunt prelucrate astfel încât zgometele să fie eliminate (numere, semne de punctuație, separatori). În final, fiecare document conține doar cuvinte cu litere mici (doar litere) separate prin spațiu (" ").

Evaluarea articolelor

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

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

unde *valoareTitlu* și *valoareConținut* sunt o combinație a valorilor conceptelor cheie în care toate conceptele au ponderi egale.

Valoarea lui *p* a fost determinată experimental folosind mai multe iterații 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 conținut.

Conținutul unui articol este considerat a fi similar cu un text din setul de documente dacă depășește un prag de 0,2 (similaritatea ia valori între -1 și 1). Pentru fiecare stare emoțională se calculează media aritmetică între valorile care îndeplinesc această condiție. 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 obținută din analiza conceptelor cheie în defavoarea rezultatelor din analiza semantică (aceasta fiind limitată intrinsec de abordarea de tip sac de cuvinte utilizată):

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

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

similarității semantice dintre știri și documentele aferente stărilor emoționale pure.

A DOUA ETAPĂ: ABORDAREA DEPENDENTĂ DE CONTEXT

Abordarea dependentă de context reprezintă o ajustare a valorilor obținute în prima etapă a proiectului (independentă de context) și este utilizată pentru personalizarea rezultatelor obținute. Am pornit de la ideea că starea emoțională a unei persoane va fi influențată diferit de o știre depinzând de starea sa curentă și de trăsăturile specifice asociate personalității sale. Spre exemplu, o persoană care este inițial fericită și citește o știre tristă va fi mai puțin afectată și deci mai puțin tristă decât o persoană care era inițial tristă. Mai mult, efectele citirii unei știri depind și de cât de aproape cititorul este din punct de vedere geografic de evenimentele prezentate.

Luând în considerare aceste observații, au fost introduse câteva elemente care evaluează aceste caracteristici. În primul rând, starea emoțională curentă este estimată prin evaluarea de către utilizator a unui set de 5 știri care sunt selectate astfel încât să fie cât mai similare cu știrile care vor fi analizate automat de către aplicație. De asemenea, utilizatorul va trebui să completeze un chestionar referitor la locația sa geografică prezentă. Cele 5 știri din setul inițial sunt determinate prin sortarea tuturor știrilor din baza de date după similaritățile cu știrile principale și selectarea primelor 5 cele mai similare.

Pentru a evalua o știre, utilizatorul are la dispoziție 2 opțiuni pentru fiecare dintre dimensiuni. Aceste valori sunt combinate în funcție de similaritatea dintre fiecare știre de evaluat (din setul de 5) cu știrile principale (din setul de 10), astfel încât cu cât similaritatea cu o știre principală este mai mare, cu atât valoarea corespunzătoare știrii de evaluat va avea o proporție mai mare în rezultatul final.

Rezultatele obținute în această etapă sunt combinate cu cele obținute în etapa precedentă (independentă de context) în proporții egale. Interfața grafică disponibilă utilizatorului a fost îmbogățită fiind introduse opțiunile descrise mai sus. Toate cele 5 știri trebuie să fie evaluate pentru a se obține cele mai exacte rezultate.

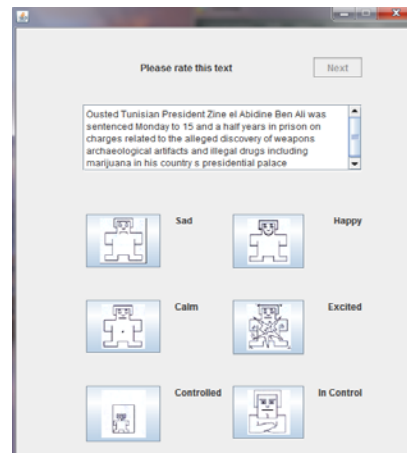


Figura 2. Interfața grafică pentru evaluarea inițială a știrilor

Locația curentă a utilizatorului este folosită pentru corectarea rezultatelor, urmărind ideea că evenimentele

care se desfășoară în apropierea unei persoane au o probabilitate mai mare de a-l afecta (influența). Pentru determinarea acesteia, se testează dacă conceptele din conținutul știrii se află pe continentul introdus de utilizator. În caz afirmativ, rezultatele sunt modificate astfel: valorile mai mari decât 5 (valoarea mediană a domeniului de definiție) sunt crescute, iar cele mai mici sunt reduse și mai mult.

Pentru testarea acestor aspect referitoare la locație am utilizat serviciul web GeoNames care oferă acces la o bază de date geografică ce conține peste 8 milioane de denumiri de locații și funcții dintre ele, pentru a putea testa legăturile dintre ele.

REZULTATE

Așa cum a fost menționat în secțiunile anterioare, sistemul proiectat a fost testat utilizând feed-uri RSS aparținând CNN.com. Pentru a vizualiza rezultatele aplicației, utilizatorul are la dispoziție o fereastră care afișează o știre și 3 grafice ce exprimă cea mai probabilă stare emoțională indusă de știre. Fiecare grafic este poziționat între 2 reprezentări a câte o stare emoțională pură pentru o reprezentare cât mai sugestivă a rezultatelor. Figura 3 reprezintă o mostră a rezultatelor aplicației.

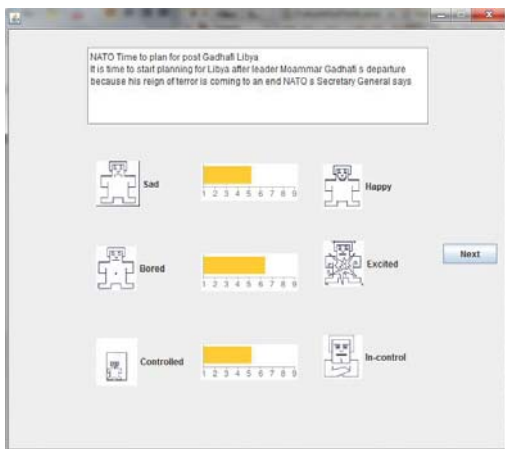


Figura 3. Interfața grafică reprezentând rezultatele aplicației

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

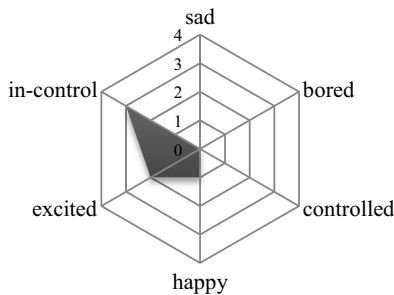


Figura 4. Exemplu de graf generat

Pentru a valida rezultatele obținute, valorile din prima și a doua etapă a proiectului au fost comparate cu valori obținute în urma unui sondaj care a inclus un număr de 10 participanți (5 femei și 5 bărbați) cu vârste cuprinse între 20 – 25 de ani din domenii diverse (calculatoare, respectiv medical). Sondajul a conținut 2 etape, fiecare având 5 știri pentru evaluarea inițială (așa cum a fost menționat anterior aceste știri sunt selectate astfel încât să fie cât mai similare cu știrile evaluate automat de aplicație) pentru a se determina starea emoțională curentă a participantului, o întrebare referitoare la poziția sa geografică și un alt set de 10 știri. Prin urmare setul de date a conținut 20 de elemente din fluxul RSS (reprezentate de știrile principale care sunt evaluate automat de aplicație). În Figura 5 sunt prezentate rezultatele sintetice obținute în urma sondajului:

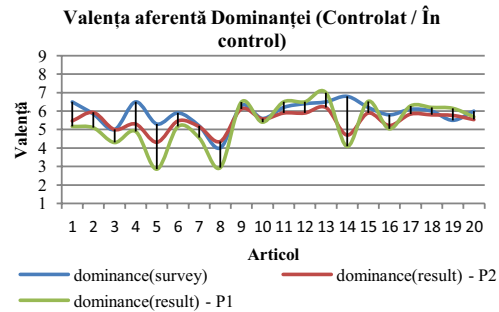
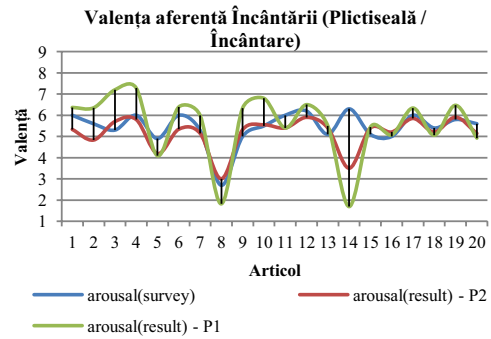
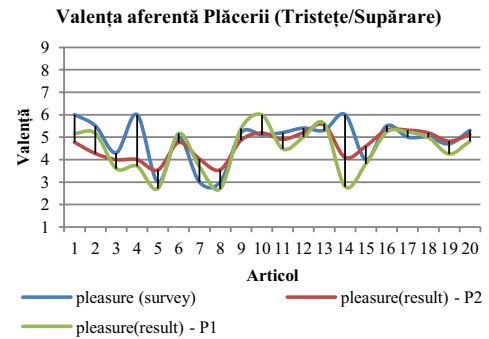


Figura 5. Valențele Plăcere, Încântare, Dominanță; P1 reprezintă abordarea independentă de context; P2 reprezintă abordarea dependentă de context

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

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

Tabelul 1. Corelațiile dintre rezultatele aplicației vs. sondaj

| | Plăcere | Încântare | Dominanță | 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 |

Așa cum poate fi dedus din tabelul anterior, valența Plăcere obține corelații proape egale între cele 2 abordări și mediile corelațiilor 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 emoționale anterioare, în timp ce dimensiunea Dominanță este mai mult influențată de contextul actual și de conținutul știrii. În plus, am obținut corelații foarte mari între cele 2 abordări, ceea ce a fost anticipat datorită faptului că P2 rafinează rezultatele inițiale (P1).

Totuși, există situații în care a doua abordare obține rezultate mai inexacte în comparație cu sondajul (ce exprimă starea emoțională curentă a utilizatorului) și față de cealaltă abordare. În afară de subiectivitatea ce apare atunci când o persoană evaluează propria sa stare emoțională, o posibilă explicație poate fi faptul că știrile din chestionar nu au fost suficient de similare cu cele evaluate în mod automat de aplicație; astfel, în loc să ajusteze valorile din prima etapă pentru a fi mai aproape de realitate, algoritmul este indus în eroare și le modifică incorect, determinând ca rezultatele finale să fie mai slabe decât cele obținute în prima etapă. O posibilă soluție la această problemă presupune îmbogățirea bazei de știri cu mai multe texte pentru ca știrile noi să fie evaluate în funcție de știri cu un grad de similaritate mai mare.

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

CONCLUZII

Scopul nostru a presupus dezvoltarea unei aplicații care evaluează modul în care este afectată starea emoțională a unei persoane după citirea unei știri. În esență abordarea utilizată a constat în 2 etape: una dependentă și 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 și acesta a demonstrat că putem considera metoda utilizată promițătoare și relevantă, într-un context în care subiectivitatea joacă un rol important. Totuși există și aspecte care pot fi îmbunătățite, prin

urmare plănuim să efectuăm o analiză mai amănunțită a textelor scrise și să modelăm cât mai exact evoluția stării emoționale a unei persoane.

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

REFERINȚE

- 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.
- Foltz, P. W.. Latent semantic analysis for text-based research. *Behavior Research Methods, Instruments and Computers*, 28(2), 197–202 1996.
- 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.
- Landauer, T. K., Foltz, P. W., & Laham, D.. An introduction to Latent Semantic Analysis. *Discourse Processes*, 25(2/3), 259–284 1998.
- Manning, C., & Schütze, H.. *Foundations of statistical Natural Language Processing*. Cambridge (Mass.): MIT Press 1999.
- 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.
- Wiemer-Hastings, P. (1999). How latent is Latent Semantic Analysis? *Proc. 16th International Joint Conference on Artificial intelligence (IJCAI'99) – Volume 2*.
- Reuters-21578 Text Categorization Collection <http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html> 2012.
- 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.
- 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.