Recunoaşterea umorului în texte

Ţifrea Oana
Facultatea de Informatică,
Universitatea Alexandru Ioan Cuza,
Iaşi,
coordonator: Dan Cristea,

29 iunie 2008

Cuprins

Listă de figuri

Listă de tabele

Capitolul 1

Introducere

"Raţionamentul corect se naşte din experienţă. Experienţa se naşte din raţionament eronat. Concluzie: Raţionamentul corect se naşte din raţionament eronat."(anonim)

La înţelegerea acestor cuvinte suntem contrariaţi de absurditatea logicii si izbucnim în râs.

Umorul are un caracter specific (literar, subtil sau fin) prin soluţii neaşteptate, caraghioase care pot produce ilaritate. Persoanele cu umor sunt acele persoane care prin comportare sau prin vorbe, în anumite contexte, declanşează râsul.

Simţul umorului este influenţat de tradiţiile, cultura, istoria unui popor, sau diferă după poziţia pe scara ierarhiei sociale sau după etate. Nu numai că variază de la o persoană la alta, dar se poate ca aceaşi persoană să găsească o glumă ca fiind amuzantă într-o zi şi în altă zi nu, depinzând de starea de spirit a persoanei, de evenimentele recent petrecute în viaţa persoanei respective.

Umorul diferă de asemenea după anumite perioade istorice, multe glume din trecut nemaifiind actualedeoarece a dispărut contextul care permitea perceperea lor ca având un anumit haz. Umorul poate fi usturător prin- satiră, ironie, batjocură; cinic, sau blând, binevoitor, plin de înţelegere, autocritic.

Lucrarea de faţă îşi propune să depisteze umorul fără a înţelege sensul mesajului. Se vor încerca algoritmii clasici de clasificare de texte si diferite atribute ale textelor umoristice si neumoristice determinate euristic.
Umorul computaţional este un domeniu în care există unele abordări de găsire a unui şablon universal pentru generarea şi recunoaşterea umorului în texte.

Experimentele testate în această lucrera cu privire la recunoaşterea umorului în texte sunt comparabile cu cele existente în domeniu.

1.1 O privire de ansamblu asupra umorului

Comunicarea om calculator nu mai constituie demult un deziderat iluzoriu al inteligenţei artificiale. Pentru ca această comunicare să fie una cât mai apropiată de comunicarea interumană, calculatorul (va trebui nu numai să recunoască, ci şi să folosească şi umorul. Mai mult, umorul oferă profunzimi ale limbajului uman- referindu-ne la cel real, complex, un limbaj creativ nu doar o mulţime de propoziţii standard. Reuşind să modelăm înţelegerea şi generarea umorului de către calculatoare, câştigăm o mai bună imagine asupra modului în care creierul uman funcţionează nu doar în privinţa umorului ci a limbajului şi cunoşterii în general. Sunt multe situaţii în interacţiunea interumană unde umorul joacă un rol important permiţând continuitatea conversaţiei, întărind relaţiile interumane. Urmaţi de paradigma CASA (Computers are Social Actors)(Calculatoarele sunt actori sociali) ne putem aştepta ca umorul să joace un rol similar şi în interacţiunea om-calculator.

În accepţiunea populară, calculatoarele nu vor putea să folosească şi să aprecieze umorul. Calculatoarele ficţionale şi roboţii au fost mereu imaginaţi ca fiind fără de umor chiar dacă pot utiliza limbajul natural. Dar şi şahul a fost odată considerat ca fiind un domeniu al oamenilor iar acum calculatoarele joacă la nivel grandmaster.

Agenţii sociali şi inteligenţi au devenit o paradigmă pentru rezolvarea şi descrierea problemelor în stilul oamenilor. Cercetările în privinţa acestor agenţi includ capacitatea de percepţie a dorinţelor, credinţelor şi a intenţiilor. Dar precum spune şi Roddz Cowie, *Dacă vom arăta emoţii la aceşi agenţi cu siguranţă ne vom aştepta ca ei să aibă şi un pic de simţ al umorului* [Binsted et al., 2006].

Pentru a înţelege umorul trebuie să-l plasăm în mediul său natural, care este societatea, trebuie mai întâi să-i determinăm funcţia utilă, care este o funcţie socială. Dintre **influenţele pozitive ale umorului** putem menţiona:

- afectează atenţia şi memoria [Baym, 1995];
- facilitează interacţiunile sociale [Binsted et al., 2006];
- ameliorează problemele de comunicare [Bergson, 1992];

- ajută la armonizarea unei conversa ţii[Bergson, 1992];
- poate stabili un punct comun între partenerii de discuţie [Hewitt, 2002];
- face conversaţia plăcută [Nijholt, 2005];
- contribuie la motivare, la atenţie, la înţelegerea şi captarea informaţiilor şi dezvoltarea unui sentiment afectiv a mesajului [Nijholt, 2006] [Binsted et al., 2006];
- poate înlesni problemele de comunicare ce pot apărea în interacţiunea dintre agenţi şi om, pentru că umorul este un mecanism primar de stabilire a individualităţii, întărindu-se raporturile acestei comunicări (omul se simte apreciat atunci când alţii îi recunoscc glumele ceea ce îmbunătăţeşte [Baym, 1995] [Black and Forro, 1999];
- umorul înlesneşte comunicarea şi învăţarea individuală şi ajută la sincronizarea învăţării în grup atât la adulţi cât şi la tineri [Binsted et al., 2006];
- facilitează crearea unei înţelegeri comune, ajută la generarea solidarităţii şi a identităţii de grup [Binsted et al., 2006];
- reduce stressul [Binsted et al., 2006];
- stimulează creativitatea şi îmbunătăţeşte comunicarea, morala şi productivitatea [Stock and Strapparava, 2006];
- atragerea atenţiei auditoriului [Stock and Strapparava, 2006];
- ajută la memorare [Stock and Strapparava, 2006];

Umorul trebuie să răspundă unor exigenţe ale vieţii în comun, având o semnificaţie socială. Nerespectarea acestor exigenţe de utilizare a umorului poate genera efecte contrare pentru care umorul a fost iniţial creat. Pot apărea astfel tensiuni de ordin personal şi social:

- câteodată este greu de înţeles glumele celorlalţi deoarece indivizii, deşi au acelaşi set de cunoştinţe, au raţionamente total diferite. Mediul social al umorului este vast şi umorul poate fi văzut ca o parte din alte multe acţiuni sociale. Potrivit sociologilor impactul unei glume şi a umorului este foarte mare în viaţa de zi cu zi;

- în cadrul grupului -umorul se bazează pe normele grupului, cunoştinţe, practici şi probleme, generând identitatea umană dar şi pe cea de grup. Vorbim despre o identitate individuală, pentru că fiecare din cei implicaţi transmit părţi de mesaje care apoi sunt filtrate de fiecare individ în parte. Deci se poate spune că umorul utilizat în grup defineşte grupul respectiv. Unii oameni se pot simţi frustraţi datorită neînţelegerii umorului celorlalţi, şi, deşi umorul ar trebui să detensioneze, utilizat greşit poate înrăutăţi situaţia;
  Din influenţele negative ale umorului putem aminti:

- poate jigni;

- poate inhiba comunicarea din cauza stilurilor diferite de a face glume;

- poate crea tensiuni din punct de vedere profesional

[Black and Forro, 1999]

1.2 Tipuri de umor

Există numeroase forme şi tipuri de umor. Dintre acestea, câteva sunt adecvate pentru folosire în dezbateri, fiind deopotrivă şi cele mai des întâlnite:

- **Anecdota**: orice întâmplare interesantă care ajută vorbitorul să clarifice o chestiune. De multe ori cei care dezbat sunt încurajaţi să folosească studii de caz sau să construiască scenarii. Unele dintre ele pot căpăta o nuanţă comică, nu doar pentru a capta atenţia,ci şi pentru a accentua o anumită idee.
  Acordorul: Bună dimineaţa, am venit să vă acordez pianul.
  Domnul Bergovici: Dar eu nu am cheamat nici un acordor.
  Acordorul: Ştiu, vecinul dumneavoastră m-a chemat [Solomivici, 2002].

- **Exagerarea (Hiperbola)**: exagerarea unor trăsături, defecte sau a neadecvării unei acţiuni.
  Simon şi Dov stau de vorbă la cafenea
  -În timpul călătoriei mele în Africa, povesteşte Simon, am întâlnit un negru atât de negru, încât a trebuit să aprind lumina în plină zi, ca să-l pot vedea.

-Asta nu-i nimic, spune Dove, când am fost în Spania am văzut un om atât de slab că trebuia să intre de două ori într-o încăpere ca să-ţi poţi da seama că e acolo [Solomivici, 2002].

- **Ironia**: folosirea cuvintelor pentru a exprima altceva decât în mod normal. De obicei vorbitorul spune opusul la ce se gândeşte sau a ce aşteaptă publicul de la el.
  Rabinul nu este deloc mulţumit de atitudinea plină de umilinţă a lui Gelb.
  -Măi Gelb, nu eşti tu atât de mare pe cât de mic vrei să pari [Solomivici, 2002].

- **Revenirea**: vorbitorul comite o eroare în mod intenţionat, numai pentru a reveni şi a corecta repede acea eroare (în cazul unei dezbateri aceasta metodă poate fi folosită pentru a da dreptate iniţial oponenţilor, şi a reveni apoi pentru a arăta cât de greşit ar fi fost acest lucru).
  -Am auzit că băiatul tău s-a căsătorit. Din dragoste sau pentru bani?
  -Din dragoste... pentru bani [Solomivici, 2002].

- **Satira**: o formă de sarcasm care scoate în evidenţă lipsurile unei idei/acţiuni/persoane.
  Ştii bancul cu statuia? Statu-ia tot...

- **Subestimarea**: transformarea a ceva mare sau important în ceva normal sau chiar mult mai mic/mai puţin important decât este în realitate
  Doi ardeleni tăiau o bombă cu fierăstrăul. Vine al treilea şi îi întreabă:
  - Mă, dar dacă explodează?
  La care ceilalţi:
  -Apăi nu-i bai, că mai avem una!

- **Umorul de situaţie**: umor care vine din experienţa proprie. Puteţi fi siguri că publicul nu are de unde să cunoască situaţia descrisă.
  Aflat în delegaţie constată că şi-a uitat papucii acasă. Se aşează la o masă şi îi scrie o scrisoare soţiei:
  Dragă Lea,
  Trimite-mi numaidecât papucii tăi. Am scris "papucii tăi" fiindcă dacă aş fi scris "papucii mei" ai fi citit papucii mei şi mi-ai fi trimis papucii tăi. Ce să fac eu cu papucii tăi? Aşa că îţi scriu foarte clar ca să înţelegi-"papucii tăi" şi nu "papucii mei", ca să citeşti papucii tăi şi să îmi trimiţi papucii mei...[Solomivici, 2002]

1.2.1 Gramatica umorului

Varietatea elementelor şi tipurilor care contribuie la crearea diferitelor forme de umor este nemăsurată, totuşi aceste elemente trebuie să conţină un ingredient de bază şi indispensabil: un impuls sau o urmă de agresivitate sau chiar răutate. Asta este tocmai ce filozoful francez Henri Bergson numea o anestezie măreaţă a inimii [Bergson, 1992]. Mulţi teoreticieni ai umorului au încercat să determine logica sau gramatica umorului. Să luăm de exemplu următoarele glume:

1. *Un masochist e o persoană căreia, deşi îi place să facă un duş fierbinte dimineaţa, face unul cald.*

2. *Domnule aş vrea să vă cer mâna fiicei dumneavoastre. De ce nu? Deja ai avut restul.*

Analizând cele două glume, dorinţa noastră este de a identifica un şablon care subliniază cele două situaţii. Gluma masochistului care se pedepseşte şi nu îşi satisface plăcerea zilnică este guvernată de regula care prezintă o **inversare a situaţiei logice**. Mâna fetei este percepută în primul caz în sensul metaforic dar apoi în sensul de bază. Altfel ascultătorul percepe situaţia în **două cadre incompatibile**. Mintea lui trebuie să opereze simultan în două lungimi de undă. Evenimentul nu este numai asociat unui singur cadru (cum este îin cazul normal), ci împărţit în două.

În umor, atât creaţia cât şi percepţia unei glume implică plăcerea de a face salt de la un sens la altul. Şablonul comun care subliniază cele două glume este percepţia unei situaţii în două cadre incompatibile, care îşi găsesc un punct comun, ceea ce implică inteligenţă. Totuşi umorul vine de cele mai multe ori din percepţia unei relaţii între două contexte consistente, dar incompatibile şi acest tip de activitate mentală pare să fie încântătoare pentru fiinţele umane. Tot ce este **contrar** nu numai **obişnuinţei** dar şi **simţurilor** sau **raţiunii** sau ceva care **lasă loc de interpretare** este amuzant. Sensul logic comun este neviabil pentru uzul practic. Nu poate fi reperat, de aceea trebuie să încercăm să evităm funcţionările cele mai comune. Umorul joacă un rol important în învăţarea şi comunicarea acestui aspect [Minski, 1981].

Umorul în mod sigur se bazează pe complicaţie. În primul rând sunt două scenarii, cel natural şi cel neaşteptat şi de obicei acestea două sunt în opoziţie. Un cuvânt cheie sau un declanşator îndreaptă mintea ascultătorului spre o altă viziune.

1.3 Teoriile umorului

Teoriile umorului sunt o încercare de a găsi un şablon comun pe baza căruia amuzamentul este construit. Marile şcoli de filozofie au încercat să găsească răspunsuri la scopul şi originea umorului, precum şi modul de manifestare. E.M. Blistein identifică trei mari categorii: [Blistein, 1964]

- Teoria de superioritate

  Potrivit teoriei de superioritate a umorului chiar şi cel mai subtil umor îşi are originea într-o anumită cruzime şi plăcere, altfel umorul vine dintr-un sentiment de superioritate faţă de cei de care râdem. Pasiunea pentru amuzament nu este altceva decât un sentiment de măreţie crescută în noi dintr-o concepţie de superioritate în comparaţie cu inferioritatea celorlalţi. Adesea absurditatea, măreţia, ciudăţenia, infirmitatea provoacă râsul.

  Această teorie conţine trei parţi:

  - orice situaţie amuzantă are un câştigător; Pe parcursul istoriei oamenii au folosit umorul pentru a concura cu alte persoane, făcându-i ţinta glumelor lor. Câştigătorul este cel care reuşeşte să îl ia în râs pe celălalt. Potrivit teoriei de superioritate oamenii care nu respectă un standard al societăţii tind să fie ridicoli.
  - nepotrivirea este mereu prezentă în umor;
  - umorul necesită un elemet surpriză;

- Teoria de nepotrivire

  Când glumele sunt examinate din lumina teoriei de nepotrivire pot fi observate două obiecte prezentate sub un singur concept sau cadru. Conceptul poate fi aplicat ambelor obiecte şi obiectele devin similare. Pe măsură ce gluma progresează, discrepanţa dintre aceste obiecte devine tot mai evidentă. Teoria de nepotrivire se axează pe elementul surpriză. Umorul este creat dintr-un conflict între ceea ce se aşteaptă şi ceea ce se întâmplă de fapt într-un text scris. Aceasta are în vedere cel mai evidente forme de umor: ambiguitate, dublul înţeles care în mod deliberat induce în eroare auditoriul, urmată de ultima frază a unei glume. În acest fel umorul încalcă o regulă importantă a utilizării limbii: Comunicarea trebuie sa fie cât mai clară.

- Teoria de eliberare

  Teoria de eliberare este de natură fiziologică şi psihofiziologică [Rutter, 1997]. în afară că exprimă o bătălie între două persoane sau un grup din societate, umorul poate exprima şi un conflict cu noi înşine, ne eliberează de grija de a fi noi înşine . Seriozitatea este un stress obişnuit asupra minţii în faţa ordinii evenimentelor urmate cu o regularitate, pe cînd umorul este o pierdere neaşteptată sau o domolire a acestui stress [Hazlitt, 1963].

Oamenii care sunt tensionaţi izbucnesc în râs de îndată ce tensiunea dispare. Potrivit teoriei de eliberare, elementul central în umor nu este nici sentimentul de superioritate nici noutatea ci sentimentul de eliberare care vine din eliberarea inhibiţiilor.

Să luăm ca exemplu următoarea glumă:

Un om intră într-o cofetărie şi comandă o prăjitură, dar cum i se aduce prăjitura cere un pahar cu lichior în locul ei. A băut paharul şi voia să plece fără să fi plătit pentru el. Propritetarul l-a oprit:
-Nu aţi plătit lichiorul.
-Dar v-am dat prăjitura în locul lui.
-Dar nu aţi plătit nici pentru aceasta.
-Dar nici nu am mâncat-o [Freud, 1957].

Rezultatul unei asemenea glume este că ne simţim contrariaţi apoi în încurcătură; după aceea râdem şi găsim absurditatea logicii amuzantă. Personificarea perfectă pentru această teorie poate fi găsită în teoria lui Sigismund Freud. Freud priveşte umorul ca o metodă de a se elibera de cenzură - termen folosit pentru inhibiţiile interne care ne previn să ne eliberăm de toate impulsurile noastre naturale. Prin cenzură nu se referă doar la impulsurile de natură sexuală, ci şi la cele răutăcioase. Teoria de superioritate a lui Freud se bazează pe răutate. El distinge între glumele nevinovate si cele tendenţioase şi deşi afirmă că există şi glume inocente, acestea foarte rar reuşesc să provoace o izbucnire în râs ca cele răutăcioase. El explică de ce glumele tind să fie compacte şi condensate, cu dublu înţeles. Este pentru a-i păcăli pe cenzorii care văd doar partea nevinovată a înţelesului şi nu reuşesc să perceapă dorinţele interzise [Freud, 1957]. În cele din urmă nu explică umorul doar după sensurile acestor dispozitive comice; ele ar fi fără

sens dacă omul nu ar putea, sub acoperire, să dea frâu liber dorinţelor reprimate. Glumele sunt comparate cu visele, pentru că datorită lor putem afla mai multe despre activitatea normală a psihicului fiecăruia. Scopul umorului este de a readuce în lumea adulţilor plăcerea care caracteriza copilăria noastră, când nu aveam nevoie de glume pentru a ne simţi fericiţi.

1.4 Criterii pentru determinarea succesului umorului

Criteriul prin care se determină dacă umorul este bun sau rău depinde în mare măsură de gustul şi preferinţele personale,şi, în particular, depinde de stil şi tehnica umoristică. Totuşi, se pare că sunt trei criterii de care depinde succesul umorului: **originalitatea, accentul şi economia**. Meritul originalităţii este de la sine înţeles; oferă elementul esenţial: surpriza, care depăşeşte aşteptările noastre. Accentul se obţine prin exagerare sau simplificare, două criterii care conduc la efectele umorului. Economia în umor nu înseamnă curaj, ci mici aluzii în locul atacului frontal, indicii implicite în locul declaraţiilor explicite. Fiecare din aceste trei teorii pot explica câteva dintre tipurile de umor, dar nu pot constitui explicaţiile pentru toate tipurile de umor. Dar umorul este în mod sigur cu mai multe feţe. Poate fi agresiv sau poate batjocoritor, poate fi inteligibil, poate exprima eliberarea din tensiune sau libertatea de griji, poate fi jucăuş, sau inteligent, chiar poate deveni serios, dar nu poate fi fals. Umorul, chiar şi la cel mai exagerat nivel de răutate, **nu poate abandona adevărul**.

Sunt trei condiţii pentru existenţa umorului [Veatch, 1998]:

- V Violarea unui anumit angajament al ascultătorului despre cum lucrurile trebuie să fie.
- N Receptorul are sentimentul dominant că o anumită situaţie este normală.
- S Simultaneitate: cele două condiţii anterioare N şi V trebuie să fie prezente în mintea receptorului în acelaşi timp.

În altă ordinne de idei, o situaţie este amuzantă cînd situaţia pare normală dar în acelaşi timp ceva apare în neregulă.

Veatch descrie condiţia V ca violare a moralităţii subiective, o violare a lucrului de care este ataşat perceptorul. Poate fi descrisă cel mai bine în acest fel, pentru că depinde de ceea ce perceptorul experimentează ca violare. Depinde de angajamentul sau ataşamentul său faţă de o situaţie dată. Veatch continuă explicând că gradul de ataşament al perceptorului este important deoarce o stuaţie i se poate părea amuzantă, jignitoare sau neremarcantă. Descrie acest fenomen cu trei niveluri de angajament şi consecinţe, relaţia ataşament-umor fiind descrisă mai bine în Tabelul 1.

			Perceptorul		
Nivel	Logica	Ataşament	O înţelege	îl jigneşte	Observă umorul
Nivelul 1	non-V	nimic	nu	nu	nu
Nivelul 2	V şi N	slab	da	nu	da
Nivelul 3	V şi non-V	puternic	da	da	nu

Tabela 1.1: Scara nivelurilor de de ataşament

Factorul de normalitate (totalitatea informaţiilor normale dintr-o glumă) împreună cu nivelul de ataşament, au o profundă importanţă în felul cum receptorul percepe experienţa violării.

1.5 Umorul din punct de vedere social

Simţul umorului poate fi:

- individual (doar unor indivizi li se pot părea amuzante unele situaţii);
- universal (de ex: clovnii pentru majoritatea oamenilor sunt stimulatori de umor);
- cultural (de ex: vestimentaţia africanilor pot provoca "'şocuri "'culturale europenilor);

Ceea ce poate pentru un individ este foare amuzant pentru altcineva poate să nu însemne nimic. Recunoaşterea umorului nu este doar subiectvă, ci este şi personală. Umorul caracterizează interacţiunea persoanelor în societate şi răspunsul nostru la acestă interacţiune trebuie înţeles în plin context [Mulder and Anton, 2001]. în construcţia unui mesaj, vorbitorii şi scriitorii

pornesc cu nişte cunoştinţe pe care ei le presupun a fi cunoscute de mai mulţi destinatari şi la care noua informaţie (cea pe care o vor sa o transmită) poate fi ataşată. Presupunerea este bazată pe cunoştinţe acumulate ale lumii sau pe experienţele cu membrii comunităţii. Când comunicarea are succes, aşteptările vorbitorului sau ale scriitorului sunt în concordanţă cu percepţia destinatarului şi noua informaţie se poate lega cu celelalte.

Umorul şi comunicarea în sine depind de acumularea de înţelesuri împărţite, de stilul oratoric al oamenilor aparţinând unei anumite culturi. Comunicarea se rupe când nivelul de cunoştinţe anterior acumulate de către vorbitor/scriitor si ascultător/cititor nu sunt la fel. Aceasta este adevărat pentru orice tip de comunicare dar ruperea este în mod special evidentă în cazul umorului, a cărui percepţie depinde direct de concurenţa între fapte şi impresii disponibile atât vorbitorului/scriitorului cât şi ascultătorului/cititorului.

Sunt multe glume care semantic pot însemna acelaşi lucru , dar în termeni de pragmatism şi cultură pot apărea neînţelegeri care pot împiedica înţelegerea glumei. Acestea sunt, ceea ce numim "glume culturale"sau "glume etnice" şi ar fi dificil să spunem că există glume universale. Poate o glumă universală este o glumă biculturală. Multe glume etnice sunt interschimbabile, depinzând totuşi, de audienţă şi de cel ce spune gluma. Bancuri despre olteni pot deveni despre moldoveni pentru oamenii din diferite regiuni ale ţării.

Capitolul 2

Cercetările în domeniul umorului computaţional

Studierea prin mijloace computaţionale a umorului este un domeniu ale cărui baze au început să se pună abia în ultimii ani, neexistând o teorie general acceptată (ca de exemplu Centering Theory în Teoria Discursului). În continuare sunt trecute în revistă principalele cercetări dedicate acestui domeniu pe plan mondial.

2.1 Generare de umor

2.1.1 LBJOJG

LBJOJG(Light Bulb Joke Generator) a fost dezvoltat de Attardo si Raskin în 1993 si după cum îi spune şi numele, este un generator de glume de tipul *"De câţi (substantiv) este nevoie pentru a (verb)"*, însă era foarte limitat deoarece nu ansamblează sau analizează atribute ale glumelor [Attardo and Raskin, 1994].
De exemplu şablonul:

How many [substantiv] does it take to change a light bulb?
[număr1].[număr1-număr2] to [activitate1] and [număr2] to [activitate2].

cu intrarea:
(poles (activitate1 hold the light bulb) (număr1 five) (activity2 turn the table he's standing on)(număr2 four))

va genera următoarea glumă:
How many Poles does it take to change a light bulb? Five. One to hold the light bulb and four to turn the table he's standing on.

2.1.2 Jape

Cercetările lui Binsted si Ritchie au condus la model semantic si sintactic de reguli ce a produs generaratorul de glume numit JAPE. JAPE foloseşte substituţia cuvintelor, a silabelor pentru a crea ambiguitate fonologică. Este alcătuit dintr-un vocabular format din 59 de cuvinte omofone, 14 şabloane şi un validator de glume [Binsted and Ritchie, 1997].
Exemplu: *What do you call a quirky quantifier? An odd number.*

2.1.3 HAHAcronym

Un alt proiect a fost HAHAcronym dezvoltat de Stock si Straparava pentru un sistem care generează versiuni amuzante ale acronimelor deja existente. Efectul comic s-a obţinut mai ales prin exploatarea teoriei de nepotrivire. Algoritmul urmăreste prima dată parsarea acronimelor. Se va păstra o parte din cuvintele care definesc acronimul iar înlocuirea celorlalte cuvinte se face prin:

- utilizarea unui câmp semnatic opus
- păstrarea literei iniţiale, a ritmului şi rimei

Exemplu: ACM(Association for Computing Machinery) devine Association for Confusing Machinery [Stock and Strapparava, 2003]

2.1.4 WISCRAIC

WISCRAIC este un generator de glume utlizat pentru cei ce învaţă limba engleză. Un exempu de glumă generată de acesta este:
What bird is lowest in spirits? A bluebird
(ambiguitate deoarece *blue* poate fi culoarea sau poate însemna depresie) [Binsted-McKay, 2000]

2.1.5 MSG

MSG este un program care converteşte cuvintele alfanumerice din parole în propoziţii amuzante. Programul ia ca argument 8 caractere si va trebui să genereze propoziţii uşor de memorat. De exemplu, folosind şablonul:

(cuv1=Nume Persoană)+(cuv2=Verb Pozitiv)+(cuv3=NumePersoană+"'s")+(cuv4=Substantiv Comun)+,while+(cuv5=Nume Persoan'ua)+(cuv6=Verb Negativ)+(cuv7=Nume Persoan'ua+"'s")+(cuv8=Substantiv comun)

cu şirul alfabumeric *AjQA3Jtv* se obţine *Arafat joined Qauayle's Ant, while Tarar Jeopardized thurmond's vase.* [McDonough, 2001]

2.1.6 Jester

Jester este un sistem online de recomandare de glume. În funcţie de alegerile efectuate se determină statistic printr-un algoritm de tipul cel mai apropiat vecin, care sunt gusturile utilizatorului în privinţa glumelor. [Goldberg et al., 2000]

2.2 Recunoaşterea umorului

2.2.1 Recunoaşterea umorului fără înţelegera sensului

Această secţiune se bazează pe articolul publicat de [Sjobergh and Araki, 2007]. Pentru a recunoaşte umorul fără a face toate conexiunile dintre cuvintele unei propoziţii se calculează valori ale unor atribute ce pot caracteriza un text. În funcţie de valorile acestor caracteristici se încearcă o clasificare a textelor. Se folosesc combinaţii de atribute pentru a vedea dacă acestea sunt suficiente pentru recunoaşterea umorului sau nu.

Caracteristicile

- **Cea mai apropiată glumă**: Din datele de antrenament se caută gluma cea mai apropiată de textul pe care îl testăm. Gradul de apropiere

dintre 2 texte se calculează după numărul de cuvinte comune celor 2 texte.

- **Cea mai apropiată non-glumă**: Pe acelaşi principiu se determină cea mai apropiată non-glumă.
- **Cele mai apropiate 5 vecine**: Se determină cele mai apropiate 5 propoziţii în datele de antrenament, fiecare cu o pondere egală cu numărul de cuvinte care se suprapun. Glumele au semn pozitiv şi non-glumele au semn negativ.
- **Cuvinte amuzante**: S-a observat că unele cuvinte sunt comune dar unele sunt specifice doar glumelor. Pentru a suprprinde acest aspect, cuvintele care apar măcar de 5 ori în datele de antrenament şi dacă apar de 5 ori mai des în glume decât în non-glume sunt păstrate într-o listă. Fiecărui cuvânt îi este asignat o pondere care este frecvenţa relativă cuvîntului printre glume (numărul de glume în care apare cuvântul/numărul total de glume) împărţită la frecvenţa relativă a cuvântului printre non-glume.
- **Ambiguitatea cuvintelor**: se calculează uitându-ne într-un dicţionar online şi numărând sensurile cuvintelor.
  - **Ambiguitatea medie**: numărul mediu de ambiguitaţi într-o propoziţie (media numărului de sensuri pentru fiecare cuvânt);
  - **Ambiguitatea maximă**: cea mai mare valoare a numărului de sensuri pentru un cuvânt dintr-o propoziţie.
- **Cuvinte murdare**: numărul de cuvinte murdare prezente în propoziţii. O listă cu 2500 de cuvinte murdare downlodată de pe Internet a fost folosită pentru a se decide dacă un cuvânt este murdar sau nu.[Sjobergh and Araki, 2007]
- **Numărul de cuvinte prezente din lista**: *you, your, I, me*, *my, man, woman, he, she, his, her, guy*, şi *girl*; şi negaţiile, numărul de apariţii ale lui *not sau nt*.
- **Pronunţia**: folosind Dicţionarul CMU[1] de Pronunţie pentru a găsi pronunţia cuvintelor au fost calculate:

[1]ftp://ftp.cs.cmu.edu/afs/cs.cmu.edu/data/anonftp/project/fgdata/dict/

  - **Rime**: numărul de cuvinte perechi care au cel puţin 4 litere, cel puţin una din ele este vocală, pronunţată la fel la sfârşitul cuvântului;
  - **Similarităţi** la fel ca şi rimele dar folosind începutul cuvintelor în locul sfărşitului;
  - **Cuvintele noi**: cuvintele care nu apar în Dicţionarul de traduceri CMU[2]

- **Cuvintele care se repetă**: numărul de cuvinte care apar de mai mult decât odată în propoziţie şi care au mai mult de 5 litere;
- **Subşiruri care se repetă** cel mai lung subşir din propoziţie care este prezent mai mult decât odată. Media cuvintelor care se repetă şi a subşirurilor este aceeşi, dar divizată după lungimea propoziţiilor;
- **Antonimia**: caută antonimia unui cuvânt în dictionary.com. Dacă oricare din aceste antonime listate este prezent în propoziţie, un scor de 1 împărţit la numărul de antonime posibile îi este atribuit.
  - **Scorul maxim de antonimie**: cea mai mare valoare (perechea de antonime cu cele mai puţine alte valori de antonime);
  - **Antonimia**: suma tuturor valorilor antonimelor calculate;

Pentru fiecare caracteristică este calculată o funcţie de prag astfel încât media entropiei[3] acestora să fie cât mai mică posibil. Pentru a clasifica o dată de test, şi a vedea cărui grup aparţine, este verificată pentru fiecare caracteristică proporţia dintre exemplele pozitive şi cele negative. Proporţia exemplelor pozitive pentru fiecare grup la care aparţine este înmulţită şi comparată apoi cu produsul proporţiilor caracteristicilor negative. Dacă produsul exemplelor pozitive este mai mare atunci exemplul este pozitiv. Metoda are

[2] ftp://ftp.cs.cmu.edu/afs/cs.cmu.edu/data/anonftp/project/fgdata/dict/

[3] Măsură statistică care calculează gradul de împrăştiere a datelor. Dacă S este un exemplu de date de antrenament $p_{\oplus}$este proporţia datelor pozitive, iar $p_{\ominus}$ este cea a datelor negative atunci

$$Entropia(S) = -p_{\ominus} log_2 p_{\ominus} - p_{\oplus} log_2 p_{\oplus}$$

	Cu	fără
Toate atributele	85.4%	50.0%
Similaritate	75.7%	83.8%
Cuvinte prezente în glume	84.1%	76.8%
Ambiguitate	62.5%	84.8%
Stil	59.1%	85.4%
Idiomuri	63.5%	85.0%

Tabela 2.1: Recunoaşterea umorului prin caracteristici

avantajul de a fi foarte rapidă.

Deoarece sunt foarte multe caracteristici care reprezintă aproape aceeaşi informaţie, autorii încercă să elimine caracteristicile redundante sau nefolositoare pentru a mări performanţa.(Tabelul 2.1) Eliminarea unor caracterisitici este importantă deoarece asa putem afla ce informaţie este utilă în detectarea umorului. Procesul se desfăşoară astfel: se elimină caracterisiticile pe rând. Atributul, care atunci când nu este prezent dă cel mai bun rezultat, este eliminat. Când toate caracteristicile sunt eliminate, cel ma bun rezultat (împreună cu structura caracteristicilor prezente) este păstrat şi după acest rezult se poate vedea care caracteristici sunt mai importante.

Corpusul

Este format din 6800 de glume colectate de pe Internet si 6800 de propoziţii colectate din BNC (British National Corpus). Toate propoziţiile au lungimea cuprinsă între şi 80 de cuvinte, cu o medie de 15 cuvinte.

2.2.2 Aplicarea clasificatorilor de texte şi a unor atribute pentru recunoaşterea umorului

Se poate încerca recunoaşterea umorului folosind învăţarea automată, utilizând clasificatorul Bayes naiv si Maşini cu vector suport.

Rada Mihalcea şi Carlo Strapparava au ales să îşi restricţtioneze studiul la *one-linere* [Mihalcea, May 2006], această secţiune bazându-se pe studiile celor doi.

Un *one-liner* este o propoziţie cu efect comic şi cu o structură lingvistică interesantă: sintaxă simplă, folosirea deliberată a unor intrumente retorice

(e.g.aliteraţii, rime) şi utlizarea frecventă a unor structuri lingvistice menite să atragă atenţia cititorilor. În timp ce glumele mai lungi tind să producă umor printr-o structură narativă mai complexă, onelinerele produc efectul comic dintr-o lovitură, cu foarte puţine cuvinte. Acest lucru face ca acest tip de text să fie folosit pentru recunoaşterea automată a umorului, deoarece efectul comic se produce în prima şi singura propoziţie.

Exemplu: *Everyone has a photographic memory. Not everyone has film.*

Datele umoristice au fost alcătuite din onlinere colectate de pe Internet folosind procesul de bootstraping. Datele nonumoristice au fost selectate şi structurate astfel incât să fie structural şi stilistic similare onelinerelor. [Mihalcea and Pulman, 2007]

Corpusul

- **Datele negative**:
  Au testat pe 3 tipuri diferite de exemple negative, o propoziţie având în medie 10-15 cuvinte. Colecţia cuprinde:

  1. **Titluri Reuters** extrase din (8.20.1996-8.19.1997). Titlurile sunt formate din propoziţii scurte menite să atragă atenţia la fel ca cele din onelinere.
  2. **Proverbe** extrase dintr-o colecţie de proverbe. Proverbele sunt texte care transmit de obicei într-o propoziţie, fapte importante sau experienţe care sunt considerate adevărate de oameni. Proprietatea lor de a fi condensate şi de transmite un mesaj într-o singură propopoziţie le face foarte asemănătoare cu onelinerele. Defapt unele glume încearcă să reproducă proverbele, cu un efect comic, ca în exemplul următor:
     Beauty is in the eye of the beer holder, preluată din *Beauty is in the eye of the beholder*.
  3. **British National Corpus (BNC)** propoziţii extrase din BNC un corpus balansat care acoperă stiluri, genuri şi domenii diferite. Propoziţiile au fost colectate astfel încât să fie similare ca şi conţinut cu onelinerele: s-a folosit un sistem de colectare a informaţiilor implementând un model vectorial pentru a identifica propoziţiile cele mai asemănătoare cu fiecare din cele 16000 de onelinere.

- **Datele pozitive**:
  Pentru a colecta foarte multe date este destul de dificil deoarece majoritatea site-urilor Web sau liste de mail fac public nu mai mult de 50-100 de glume. Pentru a depăşi această problemă, s-a implementat o tehnică de colectare automată a glumelor pornind cu câteva glume manual identificate. Algoritmul identifică apoi în mod automat o listă de pagini web în care se găseşte gluma respectivă. Paginilor astfel găsite li se aplică 2 constrângeri.

  1. **constrângerea implementată**: este un set de cuvinte cheie care au legătură cu tema;
     Setul de cuvinte cheie folosit în implementare este alcătuit din 6 cuvinte cheie care au legătură cu tematica căutată: *oneliner, one-liner, one-liner, humor, humour, joke,funny.*
     De exemplu: http://www.berro.com/Jokes
     http://www.mutedfaith.com/funny/life.htm

  2. **constrângerea modelată**: exploatează structura HTML-urilor paginilor web;
     Aceasta se bazează pe ipoteza că paginile web tind să folosească enumeraţii atunci când au o colecţie de date de acelaşi tip. De exemplu dacă una din online.rele colectate manual este prezentă într-o pagină web precedată de un tag HTML de tip ¡li¿, atunci printre alte linii care au acelaşi tag se consideră că ar fi onelinere.

  [Mihalcea and Strapparava, 2006]

  După două iteraţii ale algoritmului de căutare, pornind de fiecare dată de la un set mic de câte 10 onlinere s-a ajuns la colectarea a 24000 posibile *onelinere*. După ce s-au eliminat duplicatele prin algoritmul de cea mai lungă secvenţă comună, s-a ajuns la un set de 16000 de onelinere, care pot fi folosite în experimente.

  Prin procesul automatic s-au identificat prin selecţia aleatorie 200 de glume, cu o probabilitate de 9% de eroare, ceea ce nu are un impact prea mare in procesul de invăţare [Mihalcea, May 2006].

Experimentele

Au avut în vedere tehnici automate de clasificare folosind euristici bazate pe atribute stilistice specifice umorului (aliteraţii, antonimii, cuvinte cu sens conotativ), cu un framework de învăţare formulat ca un clasificator tipic de text. S-a încercat identificarea unui set de atribute care să fie atât semnificative cât şi fezabil de implementat utilizând algoritmii existenţi.

- **Aliteraţia** : proprietăţile structurale şi fonetice ale glumelor sunt cel puţin la fel de importante ca şi conţinutul. *Onelinerele* se bazează pe producerea de efecte umoristice prin atragerea atenţiei cititorului prin aliteraţii, repetarea unor cuvinte, rime care au un efect comic.

  Următoarele onelinere sunt un exemple de glume care includ una sau mai multe lanţuri de aliteraţii.
  Veni, Vidi, Visa: I came, I saw, I did a little shopping.
  Infants dont enjoy infancy like adults do adultery.

  Pentru a extrage aceste caracteristici se identifică şi se numără aliteraţiile sau lanţurile de rime din fiecare set de date. Lanţurile sunt extrase automat utilizând un index creat cu ajutorul dicţionarului de pronunţii CMU.

- **Antonimia**: umorul se bazează de obicei pe nepotriviri, opuneri şi alte forme de contradiţii.
  Spre exemplu, efectul comic produs de următoarele onlinere este datorat prezenţei antonimelor. *A clean desk is a sign of a cluttered desk drawer. Always try to be modest and be proud of it!* Sursa lexicală folosită pentru identificarea antonimelor este WordNet[4].

- **Construcţii cu contotaţii sexuale**: de exemplu următoarele onelinere cuprind astfel de expresii. *The sex was so good that even the neighbors had a cigarette. Artificial Insemination: procreation without recreation.* Pentru a forma un lexicon pentru identificarea acestor atribute, s-au extras din Wordnet toate synset-urile marcate ca făcând parte din domeniul sexualităţii. Lista este apoi procesată pentru a elimina cele cu o polisemie mai mare de 4. Unele onelinere conţin toate cele 3 atribute:

[4]http://wordnet.princeton.edu/

Euristica	One-linere Reuters	One-linere BNC	One-liners proverbe
Aliteraţii	74.31%	59.34%	53.30%
Antonimii	55.65%	51.40%	50.51%
Toate	76.73%	60.63%	53.71%

Tabela 2.2: Rezultatele obţinute folosind euristicile stilistice

Clasificatorul	Reuters	BNC	Proverbe
Bayes Naiv	96.67%	73.22%	84.81%
SVM	96.09%	77.51%	84.48%

Tabela 2.3: Rezultatele obţinute folosind algoritmi de invăţare automată

> *Behind every great man is a great woman, and behind every great woman is some guy staring at her Behinds.*

Pe lângă aceste caracteristici stilistice experimentele au avut în vedere ş caracteristici de conţinut, în care recunoaşterea umorului este formulată ca o problemă de clasificare.

Rezultatele

Primul set de experimente (Tabela 2.2) a evaluat acurateţea de clasificare folosind atributele de mai sus: aliteraţii, antonimii, cuvinte cu conotaţie sexuală. Acestea sunt atribute numerice care se comportă ca euristici, şi singurul parametru este determinarea unui prag indicând valoarea minimă pentru a spune dacă un text poate fi clasificat ca fiind umoristic sau nu. Acest prag poate fi obţinut utilizând arbori de decizie aplicaţi pe un set de date mic de exemplu 1000, iar pe restul de 15000 să fie testat. Luând în considerare că titlurile de articole reprezintă indicatori stilistici, indicatorii din titlurile Reuters sunt cele mai diferite de onelinere. Pentru toate seturile de date atributul de aliteraţie pare să fie cel mai important.
Al doilea set de experimente (tabela 2.3) se ocupă cu evaluarea atributelor determinate de conţinutul lor. încurajaţi de rezultatele obţinute în cele două experimente s-a construit un al treilea experiment care încearcă să exploateze atributele de stil şi de conţinut. Toate evaluările folosesc "10-fold cross-validation" (validare încrucişată prin împărţirea în 10 părţi.

Capitolul 3

Experimente de identificare a umorului prin implementarea algoritmii clasici de clasificare de texte

3.1 Clasificarea de texte

Scopul clasificării de texte este de a asigna itemi uneia sau mai multor categorii predefinite pe baza conţinutului contextual.
Funcţiile optimale de categorizare pot fi învăţate din datele de antrenament. [Dumais, 1998]

Definiţie 1 *Un program* ***învaţă dintr-o experienţă E*** *dată de o clasă de activităţi T şi o măsură a performanţei P, dacă performanţa activităţii T, măsurată cu P, se îmbunătăţeşte odată cu experienţa E.*

[Mitchell, 1997]
În cazul nostru un program spunem că învaţă să recunoască umorul din texte dacă îşi îmbunătăţeşte performanţele măsurate prin abilitatea lui de a recunoaşte umorul ceea ce implică analiza textelor, obţinută prin filtrarea diferitelor tipuri de texte.

O problemă de detectare a umorului se poate formula astfel:

- **Obiectivul T**: recunoaşterea şi clasificarea umorului în texte;

- **Măsurarea performanţei P**: procentajul de situaţii umoristice corect clasificate;

- **Experienţa E**: o bază de date cu texte corect clasificate;

Atributele care influenţează alegerea antrenării corecte:

1. **Tipul de experienţă din care va învăţa sistemul**

   Tipul de experienţă din care învăţa programul are un impact foarte mare în succesul sau eşecul maşinii de învăţare. Un atribut cheie ar fi dacă antrenarea va oferi feedback:

   - **direct**: Sistemul va învaţă din exemple de antrenare directe constituite din cuvinte sau expresii şi probabilitatea lor de a produce situaţii amuzante.
   - **indirect**: Alternativ, poate avea doar informaţie indirectă alcătuită din propoziţii si clasificarea în amuzante sau nu. Aici cel care învaţă se confruntă cu o problemă în plus de credibilitate: determinarea gradului în care fiecare cuvânt are importanţă foarte mare în stabilirea deciziei finale. Credibilitatea poate fi o problemă extrem de dificilă deoarece un cuvânt poate schimba sensul unei propoziţii.

2. **Gradul în care cel care învaţă controlează secvenţa de exemple de antrenare**

   De exemplu cel care învaţă se poate baza pe un *profesor* care să selecteze un grup de propoziţii pentru filtrare. Dacă avem de a face cu un text pe tema naturii şi vrem sa ştim dacă este amuzant sau nu o euristică destul de intuitivă ar fi să căutăm în datele de antrenare acele propoziţii care sunt şi ele pe tema naturii. În mod alternativ, cel care învaţă poate selecta şi propune un număr de grupuri de cuvinte pe care le consideră dificile şi poate solicita răspunsul corect de la profesor.

3. **Cât de bine reprezentată este distribuţia exemplelor din care va învăţă algoritmul nostru**

   în general învăţarea este mult mai credibilă când datele de antrenament urmează o distribuţie similară datelor de test. În exemplu nostru de clasificare de texte este foarte probabil să nu ştim răspunsul corect.

Teoriile curente din învăţarea automată se bazează pe presupunerea că distribuţia datelor de antrenament este identică cu distribuţia exemplelor de test. Chiar dacă trebuie să facem această presupunere, în practică această regulă este încălcată de cele mai multe ori.

Definiţie 2 *Un clasificator de texte este o funcţie care primeşte la input un document şi îl încadrează într-o categorie y dintr-o mulţime predefinită de clase y1...yk.*

Secţiunile 3.3 şi 3.4 prezintă detaliat doi dintre cei mai folosiţi algoritmi.

3.2 Metode de reprezentare a documentelor

Din documentele în forma lor originală nu se poate învăţa. Ele trebuie să fie tranformate pentru a se potrivi cu formatul algoritmilor de învăţare. Deoarece cei mai mulţi din algoritmii de învăţare folosesc reprezentarea atribut-valoare, acest lucru înseamnă transformarea textului într-un vector. În primul rând toate documentele trebuie să fie **pre-procesate**. Acest lucru înseamnă de obicei eliminarea cuvintelor care nu prezintă importanţă, aducerea cuvintelor la forma de bază, transformarea în litere mici.

După acest proces are loc **transformarea**. Fiecare cuvânt va corespunde unei dimensiuni (cuvintele identice aparţin aceleaşi dimenisuni). Notaţie: cuvântul w_i corespunzând dimensiunii i a spaţiului vectorului.

Cea mai comună metodă este cea numită TF-IDF(Term Frequency Inverse Document Frequency). TFIDF(i,j) este a i-a coordonată a documentului j:

$$TFIDF(i,j) = TF(i,j)IDF(i) \tag{3.1}$$

$$IDF(i) = log\frac{N}{DF(i)} \tag{3.2}$$

TF(i,j) reprezintă de câte ori al i-lea cuvânt se găseşte în documentul j.

N numărul de documente

DF(i) numărul de documente care conţin cuvântul i măcar odată.

Documentele transformate formează împreună matricea termilor documentului. Este de dorit ca documentele de lungime diferite să aibă aceeasi lungime, care se realizează prin aşa numita normalizare a documentelor.

Dimensiunea unui spaţiu de vectori este foarte mare, ceea ce reprezintă un dezavantaj în învăţarea automată, de accea frecvent se apelează la metode

de reducere a dimensionalităţii. Sunt două posibilităţi: fie selectarea unui subset din atributele iniţiale, fie integrarea mai multor atribute în unul.

Această metodă de reprezentare are proprietatea că pe măsură ce un cuvânt este mai frecvent în toate documentele cu atât este mai puţin valoros (nu oferă informaţii utile pentru o clasificare). [Pilasyz, 2005]

3.2.1 N-gramele

Pentru a putea recunoaşte sau genera o glumă, programul trebuie să fie capabil să proceseze secvenţe de cuvinte. O metodă pentru acestă activitate ar putea fi N-gramele.

Definiţie 3 *Un N-gram este un model care foloseşte probabilitatea condiţionată de a prezice al N-lea cuvânt pe baza celor N-1 cuvinte anterioare.*

Se construiesc din statistici obţinute dintr-un corpus mare de text folosind co-ocurenţa cuvintelor din corpus pentru a determina secvenţa de probabilităţi. Probabilitatea într-un model statistic precum cel al N-gramelor este dependentă de corpusul din care se face antrenarea. Dacă acest corpus este prea specific domeniului sau activităţii, programul nu va fi capabil să generalizeze.

Un bigram este un N-gram în care N =2, iar pentru un trigram n este 3. Un bigram va folosi cuvântul anterior pentru a determina următorul cuvânt, iar un trigram va folosi 2 cuvinte anterioare. Probabilitatea bigramelor este una condiţionată, formula pentru probabilităţile bigramelor fiind:

$$p(A|B) = \frac{p(A \wedge B)}{p(B)} \tag{3.3}$$

Pentru a calcula p (B), următoarea formulă poate fi utilizată:

$$p(B) = \frac{\text{numărul de apariţii ale lui B în text}}{\text{numărul de cuvinte în text}} \tag{3.4}$$

în mod similar,

$$p(A \wedge B) = \frac{\text{numărul de apariţii ale lui B în text}}{\text{numărul de cuvinte în text)}} \tag{3.5}$$

Aceasta înseamnă că p (A — B) este:

$$p(A|B) = \frac{\text{numărul de apariţii ale lui A şi B în text}}{\text{numărul de apariţii ale lui B în text}} \tag{3.6}$$

De exemplu: *îşi doreşte să asiste la şedinţele cenaclului. Îşi doreşte să asiste nu numai ca simplu spectator, ci să se implice, să se desfăşoare, să se simtă util.* Pentru a afla care cuvânt este cel mai probabil să urmeze după *să* putem folosi bigrame. Avem perechile *să asiste* de 2 ori şi s*ă se* de 3 ori. Deci P(se/să) = 3/5 iar P(asiste/ să)=2/5. Astfel folosind acest corpus spunem că după *să* va urma *se*. Dacă de exemplu în corpus ar fi fost doar prima propoziţie ar fi fost alte rezultate.

Deci se poate observa necesitatea selectării unui corpus cât mai general, care să acopere datele de test.

3.3 Maşinile cu vector suport (Suport Vector Machine SVM)

Intuitiv, într-o problemă de clasificare ar fi ideal să folosim cât mai multe caracteristici cu putinţă ale datelor pentru a îmbunătăţi rezultatul clasificării. în acest caz, cele mai multe clasificări suferă de aşa numitul "small sample size effect". Adică, există un anumit număr optim de caracteristici de la care, dacă ne abatem, utilizând mai multe caracteristici în clasificare, perfomanţa ar avea foarte mult de suferit.

Metoda bazată pe vectori de suport este o tehnică concepută pentru eficientizarea aproximării funcţiilor multidimensionale. Ideea de bază a SVM-urilor este de a determina un clasificator care minimizează riscul empiric (eroarea setului de antrenare sau acurateţea acestuia) şi intervalul de încredere (erorile setului de test).

În 1965, Vapnik a propus o metodă de a găsi nişte hiperplane care să "despartă" optim două clase, şi care să nu depindă de estimarea unei probabilităţi. Acesta a fost baza teoriei maşinilor care învăţa bazându-se pe vectori de suport.

SVM-urile se bazează pe conceptul de plane de decizie (plane-hiperplane de separare) care definesc anumite "graniţe". Un plan de decizie este un plan care separă un set de obiecte ce aparţin unor clase diferite.

Figura 3.1 prezintă ideea care stă la baza SVM-urilor. În figură observăm obiectele originale (din partea stângă a desenului) mapate (rearanjate), folosind un set de funcţii matematice numite nuclee (kernels). Se poate vedea că obiectele mapate (din partea dreaptă a imaginii) sunt liniar separabile şi, astfel, în loc să construim o curbă ca în figura 3.1, pentru a separa

obiectele, putem să construim o linie "optimă" care să separe obiectele albastre de cele roşii. [1]

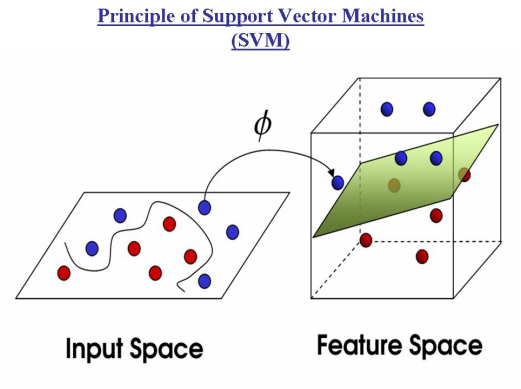


Figura 3.1: Maparea pe un alt spaţiu

Problema clasificării poate fi restricţionată, fără a restrânge generalitatea, la problema clasificării a doar două clase. În această problemă, obiectivul este să separăm cele două clase folosind o funcţie indusă de exemplele pe care le avem la dispoziţie. Scopul este de a obţine o clasificare care funcţionează bine şi pe exemplele necunoscute încă (adică generalizează bine).

Există mai mulţi clasificatori liniari care pot separa datele, dar numai unul dintre ei maximizează marginile (distanţă între linie şi cel mai apropiat punct din fiecare clasă). Acest clasificator liniar se numeşte **hiperplanul optim de separare**. Intuitiv, ne aşteptăm ca această separare optimă să ne ajute cel mai mult la problema generalizării. Punctele care se află în contact cu zona de separare (marginea) se numesc **vectori de suport**.

Hiperplanul de separare

Să considerăm problema separării unui set de vectori de antrenare, care fac parte din două clase (unde x_i sunt datele de intrare iar y_i sunt clasele)

$$\mathbf{D} = \{(x^1, y^1), \ldots, (x^l, y^l)\}, x \in \mathbf{R}, y \in \{-1, 1\} \tag{3.7}$$

separate cu un hiperplan.

$$\langle w, x \rangle + b = 0 \tag{3.8}$$

[1] http://www.imtech.res.in/raghava/rbpred/svm.jpg

Se spune că setul de vectori este **separat optim de hiperplan** dacă este separat fără eroare şi distanţa între vectorul cel mai apropiat şi hiperplan este maximă. Există o oarecare redundantă în ecuaţia 3.8, şi fără a restrânge din generalitate este mai indicat să considerăm un hiperplan canonic [Mukherjee and Vapnik, 1999],unde parametrii w, b satisfac relaţia:

$$\min_i |\langle w, x_i \rangle + b| = 1 \tag{3.9}$$

Cu alte cuvinte norma vectorului ar trebui să fie egală cu inversa distanţei între cel mai apropiat punct(obiect) din setul de date si hiperplan. Un hiperplan de separare în forma canonică tebuie să satisfacă următoarele relaţii,

$$\mathrm{y}^i[\langle w, x^i \rangle + b] \geq 1, i = 1 \ldots l. \tag{3.10}$$

Distanţa d(w,b; x) unui punct x faţă de hiperplanul (w, b) este,

$$d(w, b; x) = \frac{|\langle w, x^i \rangle + b|}{\|w\|}. \tag{3.11}$$

Hiperplanul optim este obţinut prin maximizarea marginii care suferă constrângerea dată de ecuaţia 3.9. Hiperplanul care separă datele în mod optim este cel care minimizează relaţia,

$$\phi(w) = \frac{1}{2} \|w\| \tag{3.12}$$

Relaţia este independentă de b, pentru că dacă ecuaţia este adevărată (este un hiperplan de separare), modificarea lui b va produce "mişcarea" hiperplanului în direcţia normală spre el însuşi. Marginea rămâne neschimbată, dar hiperplanul nu va mai realiza o separare optimă, în sensul că va fi mai apropiat de una din clase. În continuare presupunem că următoarea inegalitate este satisfăcută,

$$\|w\| < A. \tag{3.13}$$

atunci, din 3.10 şi 3.11,

$$\mathrm{d(w,b;x)} \geq \frac{1}{A} \tag{3.14}$$

şi deci hiperplanele nu pot fi la o distanţă mai mică de 1/A fată de oricare dintre puncte.

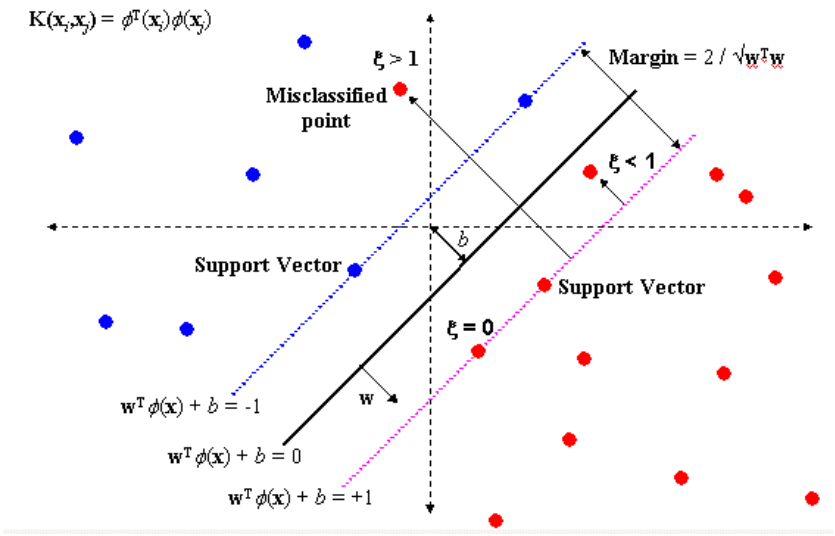


Figura 3.2: SVM-hiperplanul şi vectorii suport

Lagrangianul trebuie minimizat după w, b şi maximizat după $\alpha \geq 0$.(α_i sunt multiplicatori Lagrange):

$$\phi(w, b, \alpha) = \frac{1}{2} \|w\|^2 - \sum_{i=1}^{l} \alpha_i (y^i [\langle w, x^i \rangle + b] - 1) \tag{3.15}$$

Se poate arata ca ecuatia 3.15 se poate scrie si astfel:

$$\max_\alpha W(\alpha) = \max_\alpha - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle + \sum_{k=1}^{l} \alpha_k \tag{3.16}$$

Maparea liniară şi funcţiile Kernel

După cum am mai spus, când nu se poate găsi o separare liniară satisfăcătoare SVM-urile pot mapa vectorul de intrare, x, pe un spaţiu asociat, z, cu mai multe dimensiuni decât x. Astfel vom face maparea pe spaţiul z şi în acest spaţiu asociat vom realiza separarea liniară.

Într-adevăr, există nişte restricţii în maparea neliniară care se poate aplica, dar, surprinzător, cele mai folosite funcţii sunt acceptabile. Între acestea amintim pe cele polinomiale, aşa numitele funcţii RBF şi câteva funcţii

sinusoidale. Problema optimizării devine:

$$\max_\alpha W(\alpha) = \max_\alpha - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j K(x_i, x_j) + \sum_{k=1}^{l}\alpha_k \tag{3.17}$$

unde K(x,y) este o funcţie kernel (nucleu) care realizează maparea neliniară în spaţiul asociat (feature space) şi constrângerile rămân neschimbate.

3.4 Clasificatorul Bayes Naiv

Ideea de bază a algoritmului de învăţare naivă este de a estima probabilitatea ca un document dat să aparţină unei anumite categorii. Clasificatorul Bayes naiv presupune independenţa cuvintelor, dar în ciuda acestei simplificări el are performanţe destul de bune.

În general instanţele care trebuie clasificate sunt caracterizate printr-un vector de atribute $\vec{a} = (a_1, a_2, \ldots, a_n)$. Clasificatorul Bayes naiv asignează unei instanţe cea mai probabilă clasificare, sau o clasificare a posteoriori dintr-un set finit de clase C.

$$c_{MAP} \equiv \text{ argmax } P(c|\vec{a}) \ , \ c \in \mathbf{C} \tag{3.18}$$

care după ce este aplicată teorema lui Bayes devine

$$c_{MAP} \equiv \text{ argmax } P(c)P(\vec{a}|c) \ , \ c \in \mathbf{C} \tag{3.19}$$

Probabiliatea posteoriori $P(\vec{a}|c) = P(a_1, a_2, \ldots, a_n|c)$ poate fi estimată direct din datele de antrenament dar nu sunt în general fezabil de interpretat dacă datele nu sunt vaste. Totuşi presupunerea Bayes naivă că atributele ar fi condiţional independente unele de altele - va da următorul rezultat:

$$P(a_1, a_2, \ldots, a_n|c) = \prod_i P(a_i|c) \tag{3.20}$$

Astfel că această presupunere devine:

$$c_{NB} \equiv \text{ argmax } P(c)\prod_i P(a_i|c), \ c \in \mathbf{C} \tag{3.21}$$

[Mitchell, 1997]

În clasificarea de text se poate alege ca atribut poziţia fiecărui cuvânt într-un document. Acest lucru însemnă determinarea probabilitatăţii ca un anumit cuvânt w_k să apară la poziţia j, dată fiind clasificarea ţintă c_j, anume $P(a_j = w_k|c_j)$. Deoarece datele de antrenament sunt răzleţe introducem o nouă presupunere, anume: probabilitatea ca un anumit cuvânt w_k aflat la poziţia j este identică cu probabilitatea ca acelaşi cuvânt să fie la o altă poziţie m.

$$P(a_i = w_k|c_j) = P(a_m = w_k|c_j)\text{pentru toţi i,j,k,m} \tag{3.22}$$

Astfel estimăm probabilitatea $P(a_i = w_k|c_j)$ cu $P(w_k|c_j)$. Pentru a evita probabilitatea 0 se foloseşte aproximarea lui Laplace a probabilităţii.

$$P(w_k|c_j) = \frac{n_k + 1}{n + |Vocabulary|} \tag{3.23}$$

unde n_k reprezintă numărul de apariţii al cuvântului w_k în toate documentele de clasa c_j;

n_j reprezintă numărul total de poziţii în documentul de clasă c_j;

$|Vocabulary|$ reprezintă numărul de cuvinte distincte în toate documentele;

Cuvintele care nu se regăsesc în dicţionar sunt ignorate. Eliminarea cuvintelor care apar frecvent sau a celor care sunt foarte rare pare un motiv destul de întemeiat. Cuvintele care apar foare rar în documente este posibil să aibă un efect semnificativ la lungimea atributelor şi predicţiile nu ar trebui să se bazeze pe nişte observaţii rare. Eliminarea celor mai frecvente cuvinte este motivată de faptul că unele cuvinte, cum ar fi prepoziţiile, s-ar putea să nu ofere informaţii utile.(Hovold 2005)

Capitolul 4

Experimente în detectarea automată a umorului şi rezultatele obţinute

4.1 Corpusurile

Pentru a observa evoluţia algorimilor de învăţare automată am folosit mai multe corpusuri.

1. Texte scurte

- **Datele pozitive** Am folosit corpusul de 6805 de glume culese manual de către Jonas Sjoberg [Sjobergh and Araki, 2007]. Lungimea textelor acestui corpus variază între 1 şi 80, cu o medie de 13.

- **Datele negative** Au fost colectate urmărind să aibă aceeaşi structură ca cele umoristice (numărul de cuvinte şi media trebuie să fie aproximativ aceleaşi atât pentru datele pozitive cât şi pentru cele negative).

  Pentru non-glume am folosit date din **American National Corpus**

  http://americannationalcorpus.org/OANC/

  Textele sunt structurate pe 6 categorii: *journal* (jurnal), *technical* (tehnic), *travel guides* (ghiduri turistice), *non fiction* (nonficţiune), *fiction* (ficţiune), *letter* (scrisoare).

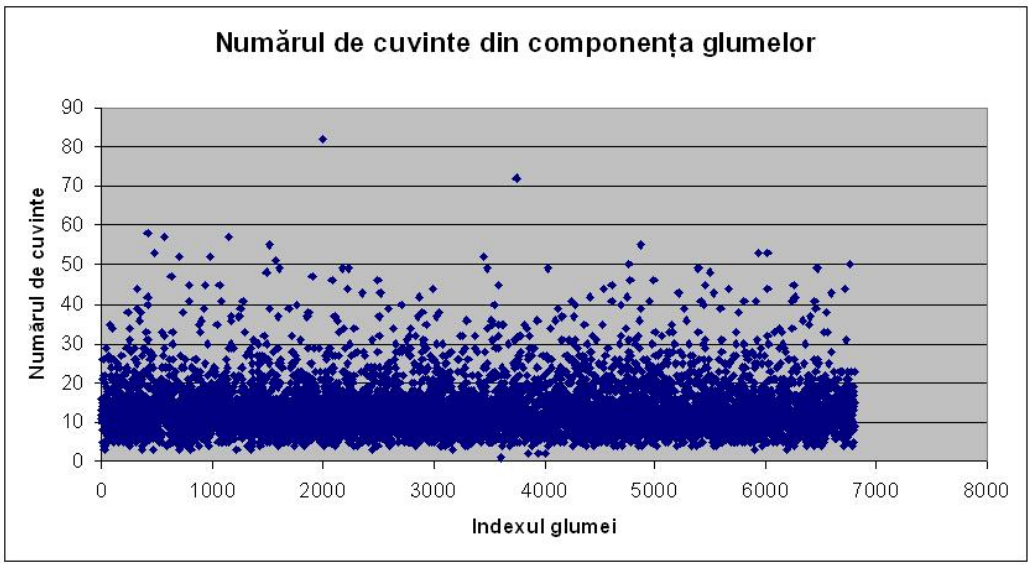


Figura 4.1: Numărul de cuvinte din componenţa glumelor

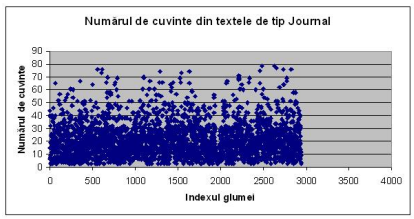
Datele au fost împărţite în propoziţii şi filtrate: au fost păstrate doar propoziţiile între 1 şi 80 de cuvinte.

Din figura se poate observa că majoritatea datelor pozitive sunt în intervalul 6-25. Astfel propoziţiile din acest interval ar trebui să aibă o probabilitate mai mare de a fi alese pentru datele negative.

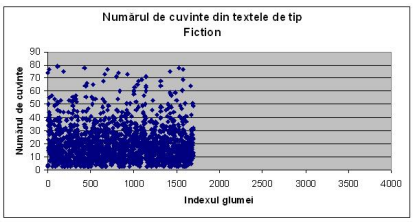
Dupa cum se poate observa din figura (media după numărul de cuvinte este 21, iar la datele umoristice este 13). Este nevoie de o nouă filtrare a datelor pentru a selecta mai multe date din intervalul 6-25. Dacă impunem

Clasificare	Media numărului de cuvinte	Numărul de texte
Journal	21.17453311	2945
Technical	15.25832609	3453
Travel Guides	20.37071886	2657
Non fiction	27.6681191	2754
Fiction	20.09540636	1698
Letter	18.62918248	2899

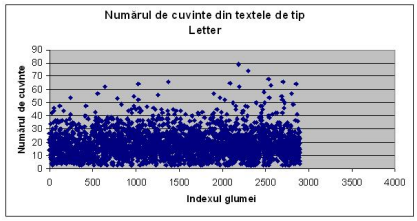
Tabela 4.1: Structura Datelor după prima filltrare



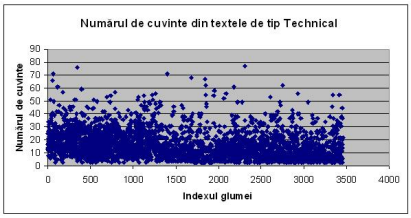
(a) Journal



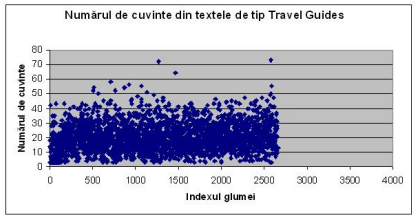
(b) Fiction



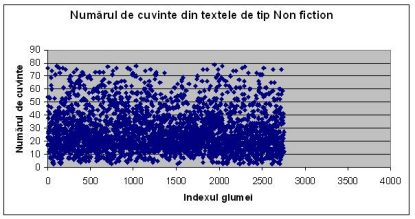
(c) Letter



(d) Technical



(e) Travel Guides



(f) NonFiction

Figura 4.2: Numărul cuvintelor pe categorii din componenţa textelor non-umoristice Corpus 1

Tip	Media
Journal	16.95258
Technical	14.61752
Travel Guides	16.74491
Non fiction	21.30893
Fiction	16.2572
Letter	16.61021

Tabela 4.2: Media numărului de cuvinte pentru datele nonumoristice

probabilitatea ca un text cu mai mult de 30 de cuvinte să fie 0.2, atunci obţinem rezultatele din tabelul 4.2.

Pentru a nu avea texte doar din cele 6 categorii al corpusului ANC am mai introdus un corpus de 1000 de propoziţii scurte cu media numărului de cuvinte egală cu 10. În concluzie avem pentru datele nonumoristice 8774 de texte cu o medie a numărului de cuvinte de 15.4.3

Datele pozitive şi negative pot fi folosite pentru a învăţa în mod automat modele computaţionale de recunoaştere a umorului şi pentru a evalua performanţele acestor modele. Se ştie că prezenţa unui număr mare de date de antrenament au potenţialul de îmbunătăţi acurateţea procesului de învăţare şi, în acelaşi timp, oferă informaţii despre modul în care datele de diumenisiune mai mare pot afecta precizia clasificării. [Witten, 2000]

2. Computers Humor/Computers NonHumor

Pentru datele pozitive am folosit date de pe site-ul [1] (un set de întâmplări amuzante ce implică prezenţa calculatorelor), iar datele negative de pe site-ul[2] - ce conţine articole despre calculatore. În total am colectat 1566 de exemple pozitive şi 1566 de exmple negative. Un exemplu de dată pozitivă

I was browsing the Internet when my friend came over and said he made a website. He told me to go a particular URL. When I went there, though, the browser said it was invalid. So I went to Google to search for it, and when I got to Google, he said, "Oh yeah, that's my web site."

iar unul de dată negativă

[1] http://linkenlim.vox.com/library/post/

[2] http://www.cs.cmu.edu/afs/cs/project/theo-11/www/

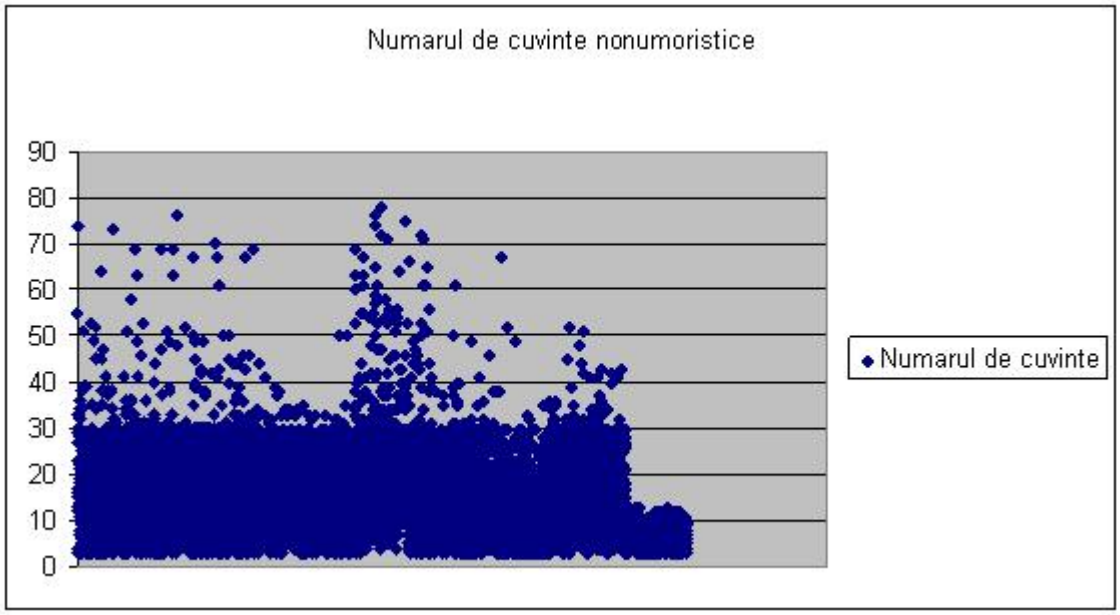


Figura 4.3: Distribuţia numărului de cuvinte penrtu datele nonumoristice după a doua filtrare

Has anybody bought a Pentium motherboard? If so or you where I can¿buy it, please send me a E-mail. Thank you in advance. Pentium processors / motherboards are not available to the general public as of yet. Intel has released them to companies such as Gateway and Dell to do testing, etc. It'll be a while...

3. Jewish humor

Pentru a vedea dacă umorul poate fi clasificat am creat o bază de date cu glume evreieşti împărţite în 14 categorii [3]. *Bar mitzvah* (31 de documente), *Bris and Mohel* (18 documente), *Divorce* (19 documente), *Donations* (7 documente), *Drinks and alcohol* (15 documente), *Food* (43 de documente), *Golf* (14 documente), *Israel* (65 de documente), *Getting married* (11 documente), *Pesach* (20 de documente), *Rabbis* (152 de documente), *Seventieth birthday* (15 documente), *Shadchen* (7 documente), *Shmuck* (12 documente), *Wedding and anniversaries* (14 documente). Exemplu: *Rabbi Morris has just resigned and Issy, the shul president, goes to visit him. "Rabbi," Issy says, "I've just heard the news. I'm really sorry that you've decided to leave us." "Don't worry," says Rabbi Morris, "you'll have nothing to worry about. I'm going to recommend a successor whom I believe will be better than me." "But that's exactly what's worrying me," says Issy, "your predecessor told me exactly the same thing."*

4. Dirty humor/Non dirty humor

Din baza de date a lui Jonas Sjoberg [Sjobergh and Araki, 2007] s-au selectat manual 1000 de glume pentru adulţi pentru a vedea dacă se poate face clasificarea umorului.

4.2 Experimentele

1. Fiecare cuvânt este un atribut sau rădăcina unui cuvânt este un atribut

Textele au fost parsate, s-au păstrat doar cuvintele care au frecvenţa de aparitie între 3 şi 200. Pe toate cele 4 tipuri de corpusuri s-au încercat două tipuri de abordări: în prima abordare fiecare cuvânt este un atribut iar în

[3] http://www.awordinyoureye.com/category%20jokes%20home%20page.html

Categoria	Rădăcina mai multor cuvinte un atribut	Orice cuvânt un atribut
Computers	5536	8252
Jewish	1561	1815
Dirty	988	920
Oneliners	4658	6113

Tabela 4.3: Numărul de componente al vectorilor

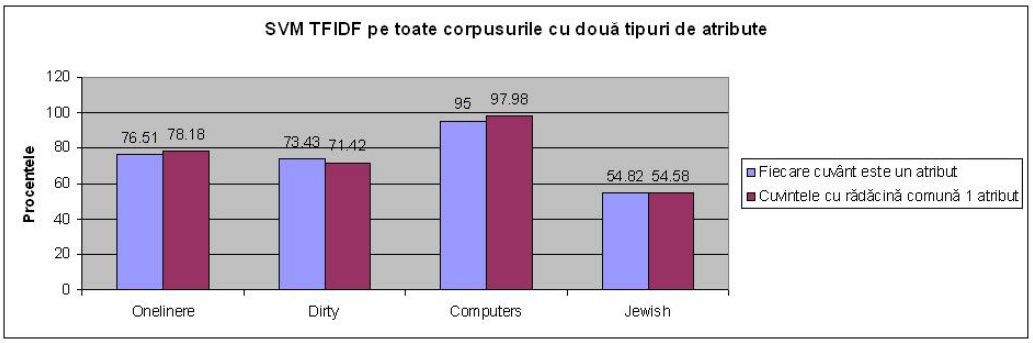


Figura 4.4: Experimetul I

cea de-a doua două cuvinte care provin din aceeaşi rădăcină sunt considerate cu un singur atribut(un exemplu de rădăcină ar fi *educ* care poate cuprinde printre altele *educate*, *education*,etc.). Pentru calculul valorii atributelor s-a folosit TFIDF.

S-a aplicat SVM cu factor de crossvalidare de 5%. Graficul 4.4 arată rezultatele obţinute (procentele reprezină acurateţea[4] experimentului).

1. Adăugarea de noi atribute euristice

Vrem să vedem dacă acurateţea clasificării creşte prin adăugarea de noi carcateristici la vectorul text. Pentru fiecare text se calculează următoarele atribute:

- **(a)Ambiguitatea medie**: Majoritatea glumelor tind să fie construcţii ambigue prin cuvintele costituite. Utilizând Wordnetul[5] am calculat media ambiguităţii medii a unui text (se adună numărul de synseturi pentru fiecare cuvânt/ numărul de cuvinte).

[4]raportul dintre numărul de texte corect clasificate/numărul total de texte*100
[5]http://www.rednoise.org/rita/wordnet/documentation/index.htm

- **(b)Antonimia medie**: Conform teoriei de nepotrivire unele construcţii vor să inducă opusul mesajului citit. Utilizand Wordnetul [6] am calculat media antonimiei medii a unui text (se adună numărul de antonime pentruu fiecare cuvânt/ numărul de cuvinte).

- **(c)Cuvinte murdare**: Folosind o bază de cuvinte murdare, [7] am numărat cuvintele "'murdare"' din text.

- **(d)Ambiguitatea maximă**: Utilizand Wordnetul [8] am calculat numărul de sysnseturi pentru fiecare cuvânt iar apoi am determinat maximul valorilor obţinute.

- **(e)Scorul maxim de antonimie**: Utilizând Wordnetul [9] am calculat valoarea maximă a numărului de antonime pe care le admit cuvintele din text.

- **(f)Lungimea maximă a proverbelor**: Folosind o bază de date de proverbe colectate de pe Internet [10] am determinat numărul maxim de cuvinte care se găsesc în proverbe şi în textul căruia îi calculăm atributul.

- **(g)Numărul de semne de punctuaţie**: Pentru a detrmina influenţa semnelor de punctuaţie pentru un text în clasificare s-a calculat numărul de semne de punctuaţie al textului.

- **(h)Numărul mic de proverbe** : Am calculat numărul de proverbe cu proprietatea că 3 cuvinte valide (cu lungimea mai mare decât 3) din acele proverbe se întâlnesc şi în textul analizat.

- **(i)Numărul mare de cuvinte din proverbe**: Am calculat numărul de proverbe cu proprietatea că 6 cuvinte valide (cu lungimea mai mare decât 3) din acele proverbe se întâlnesc şi în textul analizat.

- **(j)Numărul de repetiţii**:: numărul de repetiţii de cuvinte din textul studiat.

[6]http://www.rednoise.org/rita/wordnet/documentation/index.htm
[7]http://www.rimboy.com/words/
[8]http://www.rednoise.org/rita/wordnet/documentation/index.htm
[9]http://www.rednoise.org/rita/wordnet/documentation/index.htm
[10]http://en.wikiquote.org/wiki/

Atributul folosit	Rezultatul obţinut
Unigrame	76.51%
Bigrame	78.36%
Trigrame	79.01%

Tabela 4.4: Numărul de componente al vectorilor

Rezultate: Am folosit toate atributele anterior calculate în care fiecare radăcină este un atribut. Fiecare din atributele (a)...(j) au fost inserate câte unul. Rezultatele obţinute sunt cele din figura 4.5 (folosind SVM-uri): Se poate observa că toate atributele cresc acurateţea clasificării, cel mai

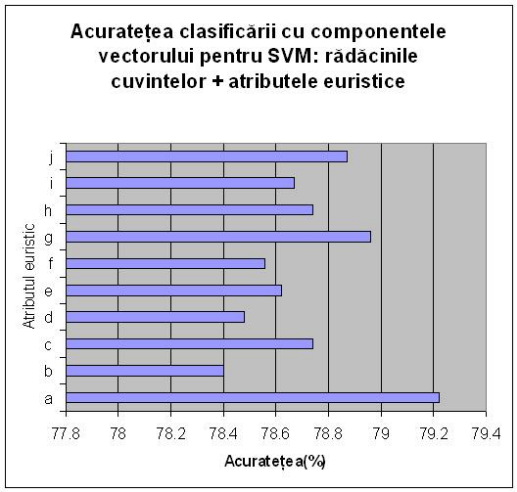


Figura 4.5: Experimentul II

important fiind atributul mediu de ambiguitate. Am creat vectori în care componente sunt rădăcinile cuvintelor textului şi toate atributele euristice. Rezultatele sunt optimiste, cu un procent de acurateţe de 80.08% şi rulând doar atributele euristice ca un vector obţinem o acurateţe de 66.05% rulat pe SVM-uri.

Influenţa n-gramelor în clasificarea de texte

Pentru primul corpus am păstrat doar cuvintele care au o frecvenţa de apariţie între 3 şi 200. Rulând pe SVM-uri cu o acurateţe de 5% obţinem: Se

poate observa că acurateţea creşte odată cu creţterea numărului de grupuri de cuvinte ale n-gramelor.

Bayes Naiv

Am rulat algoritmul Bayes naiv iar rezultatele au arătat un procent de 50% de acurateţe pentru primul corpus, şi de 45% pentru corpusul de glume evreieşti.

Arbori de decizie

Atributele stilistice au fost rulate şi cu arbori de decizie de unde s-a obţinut o acurateţe de 51% demonstrându-se astfel că SVM-urile sunt o soluţie mai bună pentru acest tip de clasificare.

Capitolul 5

Interpretarea rezultatelor

5.1 Observaţii privind rezultatele obţinute

Rezultatele obţinute sunt comparabile cu cele în domeniu chiar cu unele îmbunătăţiri(66.05%) faţ de 60% folosind atribute stilistice[Mihalcea and Strapparava, 2006]. Una din diferenţe ar putea fi faptul că experimentele noastre folosesc SVM-uri în locul arborilor de decizie şi de asemenea folosim mai multe tipuri de atribute.

Clasificarea de texte este posibilă atâta timp cât avem suficient de multe date din fiecare categorie care să fie specifice categoriei respective. După cum s-a putut observa pentru datele din corpusul cu glume evreieşti nu am reuşit să obţinem nişte rezultate prea bune deorece cele 14 categorii aveau prea puţine date de antrenament. Deşi am încercat un proces de clusterizare(K-means) pentru primul din 12000 de documente, câte 6000 din fiecare categorie, doar 200 au fost puse într-un cluster, iar restul în altul, rezultatele nu sunt prea optimiste (componente ale vectorilor sunt atributele structurale).

5.2 Posibilităţi de îmbunătăţii programul de recunoaştere al umorului în texte

- Crearea de corpusuri pentru alte limbi decât limba engleză.
- Crearea de corpusuri pe grupuri de utlizatori: clasificare de umor după vârstă, nivel de cultură, localizare.

- Testarea de noi atribute: de exemplu ar fi numărul de synseturi ale ultimului cuvânt din text, sau numărul maxim de de cuvinte din synseturi diferite.

- Majoritatea glumelor se bazează pe existenţ unui cuvât care face trecerea de la momente amuzante la cele neamuzante. Dacă s-ar găsi poziţia acelui cuvânt am putea exploata gluma mai bine. O modalitate ar fi să fie adnotate câteva glume şi să se încerce căutarea de atribute care ar putea contribui la determinarea locului cuvântului respectiv (câteva din atribute ar putea fi: pozitia verbelor, numărul de substantive).

- Încercarea de clusterizare a glumei: conform teoriei de nepotrivire există două cadre care sunt prezente într-o glumă de accea s-ar putea încerca împărţirea cuvintelor existente pe clustere în funcţie de distanţ dintre synseturi. Dacă un cuvânt apare în mai multe astfel de clustere se consideră ca este posibil ca textul respectiv să fie amuzant.

5.3 Concluzii

"Umorul se construieste pe aspiractiile si limitările omului. Nu există mai multa logică ca în umor. Pentru că umorul este adevăr." (Victor Borge) Odată cu apariţia calculatoarelor a apărut şi dorinţa de a face calculatorul să înţeleagă limbă vorbită şi scrisă. Primele cercetări au condus la rezultate promiţătoare şi era de aşteptat ca în anii cei mai apropiaţi calculatorul va fi în stare să prelucreze atât informaţia numerică, cât şi cea textuală. Totodată, pe parcursul cercetărilor, s-a conturat complexitatea limbajelor naturale, în special ambiguitatea elementelor lor, specificul utilizării în societate, etc. Umorul este una din cele mai dificile caracteristici de recunoscut nu doar pentru un calculator ci şi pentru oameni în general. Am încercat să arătăm că umorul poate fi recunoscut în texte, dar până să recunoască diferite tipuri de umor ale oamenilor este un drum lung.

Proiectul prezentat demonstrează că folosind învăţarea automată, maşina este capabilă să recunoască umorul individual.

Există anumite elemente care alcătuiesc registrul umoristic al individului care va fi valabil şi pentru:

- Nerecunoaşterea ca persoană cu umor. Dacă în experienţe anterioare o persoană a avut succese sau eşecuri în producerea umorului, acest

lucru îi va afecta motivaţia actuală în dezvoltarea abilităţii de generare a umorului. La nivelul individului dacă acesta va simţi că îi sunt recunoscute glumele va încerca să mai glumească.

- Recunoaşterea umorului altora. De asemenea, o persoană poate fi motivată să-şi dezvolte abilitatea de a înţelege gluma numai dacă a primit feed-back-uri pozitive în experienţa anterioară. Iar dacă calculatorul va primi feedback despre modul cum a dat răspunsul la o întrebare va putea să se reantreneze cu noile cunoştiinţe.

- Aprecierea umorului se referă la atitudini. S-a observat că atitudinile către oamenii cu umor sunt legate de atitudinile faţă de umor, astfel că, dacă cineva afirmă că "oamenii care glumesc chiar încearcă să mă manipuleze", acest lucru reflectă concepţia persoanei atât asupra utilizării umorului, cât şi asupra umorului însuşi. Astfel sunt oameni care resping umorul considerându-l o modalitate de manipulare.

- Râsul reprezintă un răspuns comportamental care poate fi legat sau nu de simţul umorului. Întrucât sunt persoane care râd şi fără să se amuze şi se amuză fără să râdă, râsul poate fi una din reacţiile posibile legate de simţul umorului. Cercetări suplimentare au relevat că umorul şi râsul pot să aibă nu numai efecte pozitive asupra sănătăţii psihologice, dar chiar pot avea efecte negative [Kelly, 2003].

- Concepţia persoanei, de asemenea, poate constitui un element al simţului umorului, în special atunci când acea concepţie cuprinde o apreciere a absurdităţilor vieţii.

- Utilizarea umorului ca un mecanism adaptativ reprezintă o componentă pozitivă a simţului umorului. A fi capabil de detaşare faţă de probleme şi de a le putea trata cu umor reprezintă un mijloc de protecţie împotriva evenimentelor nefavorabile ale vieţii. Umorul de autodepreciere, dar cu autoacceptare reprezintă unul din cele mai mature mecanisme adaptative, însă umorul pe seama altora (ostil, agresiv) reprezintă un mecanism nevrotic de adaptare cu efect negativ în plan interpersonal.

O modalitate prin care s-ar putea realiza acest lucru ar fi prin analiza-manifestării umorului unui individ faţa de persoane diferite. Un exemplu ar fi arhiva de mesaje (de exemplu de tip Messenger) ale individului. O analiză

mai profundă ar putea releva ce se întâmplă cu evoluţia umorului individual. Noutatea lucrării constă în primul rând în încercarea, pentru prima dată, a unei clasificări a umorului obţinând o acurateţe de 73% pentru categoria dirty/non dirty humor şi 54.8% pentru tipuri de bancuri evreieşti. De asemenea utilizarea doar de atribute stilistice a dus până la o acurateţe de 66.7%. Experimentele au demonstrat deasemnea că n-gramele cresc acurateţea clasificării de la 76.5 pentru unigrame până la 79.8 pentru trigrame. Cel mai bun rezultat este obţinut combinând atributele structurale cu cele stlistice, obţinând o acuratet]e de 80.08% prin aplicarea SVM-urilor cu factor de cross-validare de 5%. Conform informţiilor noastre este prim lucrare în limba română care abordează umorul computaţional, iar prin posibilităţile de extindere mai sus menţionate acurateţea algoritmului de clasificare va creşte.

Bibliografie

Savatore Attardo and Victor Raskin. Non-literalness and non-bona-fide in language: An approach to formal and computational treatments of humor. *Pragmatics and Cognition*, 1994.

Nancy Baym. The performance of humor in computer-mediated communication. *Journal of Computer-Mediated Communication, Volume 1 Issue 2*, 1995.

Henri Bergson. *Teoria rîsului*. Institutul European, 1992.

Kim Binsted and Graeme Ritchie. Computational rules for generating punning riddles. *HUMOR, the International Journal of Humor Research, Mouton de Gruyter, volume 10-1*, 1997.

Kim Binsted, Benjamin Bergen, Seana Coulson, Anton Nijholt, Oliviero Stock, Carlo Strapparava, Graeme Ritchie, Ruli Manurung, Helen Pain, Annalu Waller, and Dave O'Mara. Computational humor. *IEEE Intelligent Systems*, 21(2):59–69, 2006. ISSN 1541-1672.

Justin McKay Binsted-McKay. Generation of idiom-based witticisms to aid second language learning. *Proceedings of the April Fools Day Workshop on Computational Humour, Trento, Italy*, 2000.

Leah Black and Denise Forro. Humor in the academic library : You must be joking! or, how many academic librarians does it take to change a lightbulb? *Michigan State University Libraires at Michigan State University*, 1999.

Elmer Blistein. Theories of humour. *Encyclopedia americana*, 1964.

Susan Dumais. Using SVM for text categorisation. *Decision Theory and Adaptive Systems Group*, 1998.

Sigismund Freud. *Jokes and their Relation to the Unconscious*. 1957.

Ken Goldberg, T. Roeder, D. Gupta, and C. Perkins. Eigentaste: A constant time collaborative filtering algorithm. Technical Report UCB/ERL M00/41, EECS Department, University of California, Berkeley, 2000. URL `http://www.eecs.berkeley.edu/Pubs/TechRpts/2000/3869.html`.

William Hazlitt. *Lectures on the English Comic Writers*. Wiley and Putnam, 1963.

John Hewitt. *The Architecture of Thought: A New Look at Human Evolution*. Holmhurst House, Beds., 2002.

Martin Rod.Anton Puhlik-Doris Patricia Larsen Gwen Gray Jeanette Weir Kelly. Individual differences in uses of humor and their relation to psychological wellbeing. *Journal of Research in Personality*, 2003.

Craig McDonough. Mnemonic string generator:software to aid memory of random password. Technical report, CERIAS Technical Report. West Lafayette, IN: Purdue University., 2001.

Rada Mihalcea. Learning to laugh (automatically):computational models for humor recognition. *Computational Intelligence*, 22:126–142(17), May 2006.

Rada Mihalcea and Stephen G. Pulman. Characterizing humour: An exploration of features in humorous texts. In *CICLing*, pages 337–347, 2007.

Rada Mihalcea and Carlo Strapparava. Technologies that make you smile: Adding humor to text-based applications. *IEEE Intelligent Systems*, 21(5):33–39, 2006. ISSN 1541-1672. doi: http://dx.doi.org/10.1109/MIS.2006.104.

Marvin Minski. Jokes and the relation of the cognitive unconscious. 1981.

Tom M. Mitchell. *Machine Learning*. McGraw-Hill, New York, 1997.

Sayan Mukherjee and Vladimir Vapnik. Multivariate density estimation: an SVM approach. Technical Report AIM-1653, AI Memos (1959 - 2004), 1999.

M.P. Mulder and Nijholt Anton. Humour research: State of the art. Technical report, Deliverable IST Programme on Future and Emerging Technologies. University of Twente, 2001.

Anton Nijholt. Conversational agents a little humour too. *IEE IntelligentSystems*, pages 22–26, 2006.

Anton Nijholt. Conversational agents, humorous act construction, and social intelligence. *University of Hertfordshire*, pages 1–8, 2005. ISBN=1-902956-45-X.

Itvan Pilasyz. Text categorization and suport vector machine. *Departmenet of Measurment and Information System*, 2005.

Jason Rutter. Stand-up as interaction: Performance and audience in comedy venues. *University of Salford*, 1997.

Jonas Sjobergh and Kenji Araki. Recognizing humor without recognizing meaning. 2007.

Tesu Solomivici. *5000 de ani de umor evreiesc - o antologie subiectivă*. Editura Tesu, Bucureşti, 2002.

Stock and Carlo Strapparava. Computational humor and prospects for advertising. *Rob Milne: A Tribute to a Pioneering AI Scientist, Entrepreneur and Mountaineer. IOS Press.*, 2006.

Oliviero Stock and Carlo Strapparava. Getting serious about the development of computational humor. *IJCAI*, pages 59–64, 2003.

Thomas C. Veatch. A theory of humor. *International Journal of Humor Research*, 1998.

Ian H. Witten. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations (The Morgan Kaufmann Series in Data Management Systems)*. 2000.