Extragerea opiniilor implicite din texte economice

Claudiu Cristian Muşat
Universitatea „Politehnica" din Bucureşti
Splaiul Independentei 313, Bucureşti
claudiu.musat@cs.pub.ro

Ştefan Trauşan-Matu
Universitatea „Politehnica" din Bucureşti
Splaiul Independentei 313, Bucureşti
trausan@cs.pub.ro

REZUMAT

În această lucrare propunem o nouă metodă de extragere a opiniilor din texte economice, bazată nu pe apariţii explicite de opinii în textele analizate, ci pe polaritatea intrinsecă a indicatorilor economici întâlniţi. Metoda are două etape. În primă instanţă am creat corpusul de texte pornind de la o colecţie de articole ale unei publicaţii financiare extinsă apoi urmând legăturile care ies din siturile analizate şi verificând dacă un subiect economic este discutat în cadrul acestora. În cea de a doua fază vom dovedi că majoritatea textelor economice conţin opiniile personale, chiar dacă opiniile explicite lipsesc, bazându-se pe apariţii simultane ale indicatorilor economici şi modificatorilor de stare pentru a determina polaritatea fiecărui text. Rezultatele acestor cercetări pot ajuta utilizatorii din domeniul economic să navigheze cu mai mare uşurinţă prin vasta cantitate de informaţie disponibilă prin gruparea textelor după predicţiile conţinute şi includerea acestei informaţii în prezentarea lor.

Cuvinte cheie

Economie, Extragerea opiniilor, Corpus financiar, Indicatori economici.

Clasificare ACM

H5.2. User Interfaces: Natural Language.

INTRODUCERE

În general cercetările legate de extragerea opiniilor din texte s-au axat pe găsirea opiniilor exprimate direct în textele analizate [11]. Există totuşi domenii unde această abordare este deficitară, unde trebuie sa lărgim perspectivele pentru a detecta şi opiniile care nu sunt exprimate direct. Limităm această analiză la textele din domeniul economic.

Extragerea opiniilor

Una dintre lucrările de referinţă din domeniul extragerii opiniilor din texte este cea a lui Pang şi Lee [13], în care documentele sunt clasificate în funcţie de opinia prevalentă din întreg textul, obţinută ca sumă a tuturor opiniilor întâlnite. Wiebe [15] împarte textele în propoziţii subiective şi obiective iar mai apoi accentul a fost pus în special pe determinarea la nivel individual a polarităţii opiniilor întâlnite. Această împărţire variază de la o dihotomie pozitiv-negativ până la o clasificare mai complexă bazată pe intensitatea părerilor din text [5].

Un alt scop conex extragerii de opinii este acela de a lega opiniile extrase de concepte. Ulterior s-a mers mai departe, legând opiniile de anumite faţete ale unor concepte [12], apărând astfel ideea de opinii faţetate. Scopul acestei fracţionări este de a determina exact care sunt subcomponentele obiectului sau ideii analizate la care se referă opinia exprimată.

Toate metodele descrise mai sus se bazează pe întâlnirea opiniilor exprimate direct în textul analizat, acesta putând fi unul dintre motivele pentru care pentru mult timp analiza opiniilor s-a rezumat la situaţii şi exemple relativ simple ca recenziile de filme [15]. Simplitatea în acest caz mai survine şi din faptul că obiectul discuţiei este cunoscut a priori, o largă majoritate a opiniilor prezente în text sunt legate de acel obiect iar părerile exprimate pot fi împărţite cu uşurinţă în pozitive şi negative.

În domeniul economic, într-un studiu legat de analiza sentimentelor în *blog*-uri financiare, Ferguson et al. [8], folosind adnotări la nivel de paragraf şi un clasificator Bayesian, au arătat că în mai mult de 60% dintre textele analizate au fost găsite opinii. Trebuie totuşi menţionat că studiul lor s-a rezumat la analiza textelor în care se găseau referiri explicite la companii listate în indexul S&P500. Presa economică în ansamblu a fost analizată de către Ahmad [1], în studiile sale despre structura opiniilor în domeniul financiar. O posibilă legătură intre opiniile din presa financiară şi variaţiile preţurilor entităţilor vizate a fost investigată de Ku [10].

Opinii implicite

Lipsa opiniilor exprimate în mod explicit nu înseamnă întotdeauna lipsa totală a părerilor din textul analizat. În lucrarea despre gestionarea conţinutului generat de utilizatori Amigo [2] a arătat că simpla menţionare a unui produs într-o analiză poate exprima o atitudine pozitivă faţă de acel produs. Twitter este un foarte bun exemplu de mediu informal şi foarte polarizat unde în general sentimentele nu sunt exprimate explicit dată fiind limitarea dimensiunii mesajelor. Cruz [6] a descoperit că în căutarea părerilor clienţilor unei companii de cele mai multe ori rezultatele relevante sunt menţiuni obiective ale produselor companiei în cauză.

Opiniile exprimate în mediul economic fac în general parte din această ultimă categorie, rareori fiind un economist gata să îşi exprime îndoielile legate de finalitatea unor evenimente viitoare. Ceea ce stă la baza imixtiunii dintre fapte şi opinii este empiricismul. Certitudinea rezultată din acumularea de exemple pozitive a fost numită încă din 1958 de către Galbraith [9] „*conventional wisdom"* - înţelepciunea convenţională, şi a ajuns să se confunde cu orice corp de cunoştinţe universal

acceptate fără a avea in mod necesar şi o certitudine ştiinţifică.

SISTEMUL PROPUS

Metoda prezentată în continuare are ca scop extragerea părerilor exprimate în texte economice care sunt prezentate ca şi certitudini. Predicţiile economice sunt analizate sub forma combinaţiilor între indicatori economici ca „şomajul", „piaţa" sau „productivitatea" şi termeni ce indică o variaţie a stărilor viitoare, pe care îi vom denumi modificatori.

Indicatorii sunt împărţiţi în doua submulţimi – una care conţine indecşii care cresc odată cu economia per ansamblu şi una compusă din indecşi care scad în perioade de prosperitate economică. Vom denumi în continuare indicatorii din prima mulţime „pozitivi", iar pe cei ramaşi „negativi".

Modificatorii sunt o colecţie de n-grame care indică ori o creştere viitoare ori o descreştere a indicatorului de care sunt ataşaţi. Ca şi în cazul indicatorilor de mai sus, şi modificatorii sunt împărţiţi în două mulţimi – pozitivi şi negativi, unde cei pozitivi semnalează o creştere cantitativă iar cei negativi o descreştere.

Termenii relevanţi

Termenii şi frazele economice consideraţi cheie – indicatorii – au fost obţinuţi prin procesări succesive pornind de la dicţionarul de termeni economici al *EconomyWatch* iar modificatorii au fost selectaţi pornind de la o colecţie de termeni ce semnalează o creştere sau descreştere.

Prima modalitate de extindere a mulţimilor iniţiale de termeni relevanţi se bazează pe *WordNet* [7] şi constă în includerea celorlalţi termeni din *synset*-urile în care se află termenii iniţiali şi totodată a termenilor care în descriere (*gloss*) au termeni din mulţimea iniţială fără a fi negaţi.

A doua modalitate de extragere foloseşte o Reţea Conceptuală (*Conceptual Network Graph)* [4] care întoarce termenii ce tind să apară împreună cu cei iniţiali. Am împărţit corpusul prezentat mai jos în două părţi egale, iar una dintre jumătăţi a fost folosită pentru antrenarea reţelei. Rezultatul fazei de antrenare este un graf bipartit care reprezintă întreaga colecţie, în care una din părţi este proiecţia termenilor folosiţi iar cealaltă o reprezentare a documentelor folosite în antrenare. Conexiunile dintre cele două părţi ale grafului sunt direct proporţionale ca forţă cu numărul de apariţii ale termenilor folosiţi în documentele date. Din totalitatea termenilor aparţinând tuturor documentelor din colecţie doar un mic număr sunt adăugaţi mulţimilor iniţiale de termeni relevanţi. Aceştia sunt obţinuţi energizând graful din nodurile corespunzătoare termenilor iniţiali şi selectând termenii noi ale căror noduri primesc energie peste un prag dinainte stabilit. Energizarea grafului presupune trimiterea unui puls de energie egal cu o unitate din nodul sursă, energie care se distribuie vecinilor acelui nod în funcţie de caracteristicile legăturilor dintre ele.

Corpusul folosit

În experimentul prezentat am folosit articolele publice din secţiunea „Finanţe" ale *The Daily Telegraph* şi articolele din subsecţiunile acesteia, obţinute pornind de la ultimele texte publicate şi urmând apoi legăturile către articolele mai vechi. În total 21106 articole financiare publicate între 2007 şi 2010 au fost procesate, iar dintre acestea 499 au fost marcate ca având o încărcătura pozitivă şi 216 au fost considerate încărcate negativ. Cum şi interpretarea umană a acestor articole este subiectivă, iar adnotarea ar putea avea de suferit din această cauză, ne-am bazat pe înţelegerea adnotatorilor [3]. Astfel pentru experimentele următoare am reţinut numai acele articole unde părerile adnotatorilor au coincis. Aceştia au selectat textele cu încărcătură subiectivă – sau care conţin previziuni economice – din secţiunile *Finanţe* (rădăcina arborelui de categorii), *Piete*, *Locuri de muncă* şi *Comentarii.* Alte preprocesări ale corpusului folosit constau în eliminarea sufixelor (*stemming)* tuturor cuvintelor folosind algoritmul lui Porter [14] şi eliminarea cuvintelor comune sau de legătură.

Combinarea indicatorilor şi modificatorilor

Următoarea operaţie constă în găsirea polarităţii prevalente la nivel de document, care reiese din apariţii simultane de indicatori economici şi modificatori de stare.

Un bun exemplu este legat de discuţiile recente despre creşterea şomajului. Termenul de şomaj este el însuşi negativ, iar când modificatorul care îi este ataşat este unul pozitiv, de exemplu „va creşte" atunci predicţia per ansamblu va fi una negativă. În mod asemănător, daca modificatorul ataşat este negativ ca „a scăzut", rezultatul este pozitiv. O formulă simplă sumarizează cele de mai sus:

$$P(O) = P(I) * P(M)$$

Unde

- P este polaritatea construcţiei
- O este opinia rezultată
- I este indicatorul folosit
- M este modificatorul ataşat

Constrângeri

Nu toate perechile indicator - modificator întâlnite în text sunt opinii exprimate implicit. De exemplu expresia „creşterea şomajului se poate dovedi dezastruoasă pentru economia locală" nu semnalează o predicţie a autorului despre evoluţia viitoare a şomajului. Ţinând seama de faptul ca o supra simplificare poate reduce în mod semnificativ acoperirea (*recall)* sistemului, am folosit două euristici pentru a evita incertitudini de tipul celei prezentate mai sus. Cum limitările impuse sunt legate de părţile de vorbire pe care le reprezintă cuvintele cheie, am folosit un algoritm de marcare a părţii de vorbire pentru fiecare termen din textul analizat (*part of speech tagger).*

Prima limitare impusă este ca indicatorii trebuie sa fie ori substantive (sau o expresie care să conţină un substantiv) ori un adjectiv ataşat unui substantiv cu rol neutru. Substantivele neutre sunt foarte des întâlnite în textele economice din limba engleză, formând combinaţii ca „date economice" şi altele (*economic data, mortality figures, unemployment numbers*).

Totodată pentru modificatorii folosiţi ne limităm la verbele asociate indicatorilor sau, atunci când verbul asociat substantivului indicator este unul care sugerează o stare de continuitate, modificatorii pot fi şi adverbele ataşate acelor verbe. Un exemplu de modificator adverb este expresia „revenirea rămâne incertă", în care verbul „a rămâne" este cel care sugerează continuitatea şi pe care îi vom denumi în restul lucrării „continuatori".

Tabelul 1. Combinaţii relevante de termeni cheie

Termen	Parte de vorbire	Cerinţe
Indicator	Substantiv / n-gram	Nici una
Indicator	Adjectiv	Substantiv neutru
Modificator	Verb	Nici una
Modificator	Adverb	Verb continuator

Totodată, detectarea negaţiilor aferente indicatorilor şi modificatorilor este crucială pentru acurateţea sistemului întrucât folosirea termenilor negaţi se dovedeşte a fi mai întâlnită decât folosirea lor directă.

Un document în care majoritatea perechilor indică o creştere a unui indicator pozitiv sau o scădere a unui indicator negativ va fi marcat ca „pozitiv" şi analog un document în care majoritatea perechilor arată o creştere a unui indicator negativ sau o scădere a unui indicator pozitiv va avea o polaritate negativă.

REZULTATE

Precum am precizat mai sus, termenii cheie au fost extraşi folosind jumătate din articolele marcate ca pozitive sau negative din corpusul *Telegraph Finance*, iar cealaltă jumătate constituie corpusul de testare.

Rezultatele operaţiilor de extragere de opinii precum şi precizia şi acoperirea testelor sunt prezentate în tabelele 2 respectiv 3.

Tabelul 2. Rezultatele testelor de extragere de opinii implicite

Corpus	Total	Pozitive	Negative	Neutre
Negative				
Toate	499	86	252	161
Slujbe	54	1	35	18
Pieţe	253	44	128	81
Comentarii	153	32	73	48
Pozitive				
Toate	216	87	50	79
Slujbe	14	4	3	7
Pieţe	152	71	31	50
Comentarii	34	11	9	14

Recallul sistemului este proporţional cu dimensiunea corpusului de antrenare, cele mai slabe rezultate fiind obţinute la secţiunile *slujbe* şi *comentarii* din articolele pozitive, acelea fiind şi secţiunile în care numărul de articole este cel mai mic.

Tabelul 3. Precizia şi RECALLUL testelor de extragere de opinii implicite

Corpus	**Precizie (%)**	**RECALL (%)**
Negative		
Toate	74.55	67.73
Slujbe	97.22	66.67
Pieţe	74.41	67.98
Comentarii	69.52	68.62
Pozitive		
Toate	63.50	63.42
Slujbe	57.14	50
Pieţe	69.60	67.10
Comentarii	55	58.82

Totodată este important de menţionat faptul că pe secţiunile de *comentarii* atât din articolele pozitive cât şi din cele negative s-a obţinut o acurateţe mai mică decât în celelalte secţiuni, acest rezultat putând fi pus pe seama numărului comparativ mai mare de subiecte abordate în secţiunea de *comentarii,* de la macroeconomie până la finanţe personale.

CONCLUZII ŞI CERCETĂRI VIITOARE

Schimbând accentul de la opinii exprimate direct în texte economice la cuantificarea predicţiilor despre finalitatea unor procese economice viitoare sperăm sa fi creat o unealtă nouă în analiza opiniilor, complementară celor deja existente.

Rezultatele de mai sus indică faptul că prin identificarea cuvintelor cheie şi urmărirea apariţiilor simultane ale indicatorilor economici şi modificatorilor lor, clasificarea documentelor în pozitive şi negative are o precizie asemănătoare şi în unele cazuri superioară altor experimente din domeniu[8]**.** Din aceleaşi rezultate deducem totodată nevoia unor îmbunătăţiri. Una din acestea este legată de corelaţia dintre dimensiunea corpusului şi acoperirea experimentului, care indică necesitatea creşterii numărului de texte analizate.

Alte viitoare îmbunătăţiri includ procesarea ironiilor şi metaforelor din text, ambele fiind tehnici folosite în largă măsura în textele economice şi care nu au fost incluse în prezentul experiment. În plus, sunt numeroase cazuri în care autorul foloseşte citate externe a căror validitate este apoi contestată, situaţie care duce la o inversare a polarităţii opiniilor extrase din acea parte de text. Astfel această metodă trebuie inclusă printre negaţiile pe care sistemul trebuie sa le poată trata.

Faptul că un număr mic de cuvinte cheie poate duce la rezultate similare sau chiar superioare celor obţinute de alte experimente în care sunt implicate toate cuvintele din text sugerează că predicţiile sunt de fapt exprimate în general într-un mod succint şi mai ales relativ uşor identificabil.

Extragerea opiniilor emise în articole publicate pe Internet are şi avantajul de a permite suprapunerea cu analiza interconectivităţii textelor în cauză. Putem astfel afla dacă

documente care conţin predicţii economice de o anumită polaritate sunt cu o probabilitate mai mare conectate cu documente având aceeaşi polaritate.

O astfel de concluzie ar fi una foarte importantă deoarece ar implica faptul că deşi cititorii cred că au o privire de ansamblu nepărtinitoare asupra realităţii economice, fiind informaţi din surse diverse din Internet, dacă acele surse sunt interconectate atunci e o mare probabilitate să împartă de fapt şi aceeaşi viziune economică, una părtinitoare. Finalitatea întregului sistem este deci de a uşura informarea utilizatorilor cu privire la temele economice ale zilei, de a facilita accesul la informaţia relevantă şi de a evidenţia opiniile distincte.

REFERINŢE

1. Ahmad K. Cheng D. and Almas Y. Multi-lingual sentiment analysis of financial news streams. First International Workshop on Grid Technology for Financial Modeling and Simulation. Palermo, 2006.
2. Amigo E., Spina D. and Bernardino B. User Generated Content Monitoring System Evaluation. First workshop on Opinion Mining and Sentiment Analysis, Sevilla, 2009
3. Balahur A, Steinberger R. 2009. Rethinking Sentiment Analysis in the News. Proceedings of WOMSA. 79-89. 2009.
4. Ceglowski M., Coburn A. and Cuadrado J. Semantic Search of Unstructured Data using Contextual Network Graphs. 2003
5. Chen L.S., Chiu H.J.: Developing a Neural Network based Index For Sentiment Classification. IAENG Hong Kong 2009.
6. Cruz F., Troyano A., Ortega F. and Enriquez F. Domain Oriented Opinion Extraction Metodology. First workshop on Opinion Mining and Sentiment Analysis, Sevilla, 2009
7. Felbaum C. WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press, 1998
8. Ferguson P., O'Hare M. and Bermingham A. Exploring the use of Paragraph-level Annotations for Sentiment Analysis of Financial Blogs. First workshop on Opinion Mining and Sentiment Analysis, Sevilla, 2009
9. Galbraith J. The Affluent Society. Chapter 2. 1958
10. Ku L, Lee L, Wu T , Chen H. Novel Relationship Discovery Using Opinions Mined from the Web. AAAI, 213-221, 2006.
11. Liu B. Opinion Mining. In Proceedings of WWW-2008, Beijing, 2008.
12. Mei Q., Ling X., Wondra M., Su H., Zhai C.: Topic Sentiment Mixture: Modeling Facets and Opinions in Weblogs. WWW Calgary 2007.
13. Pang, B., Lee L. : "Opinion Mining and Sentiment Analysis", FTIR, 2008
14. Porter, M., An Algorithm for Suffix Stripping. New models in probabilistic information retrieval. London: British Library (1980)
15. Wiebe J., Wilson T. and Cardie C.: Annotating Expressions of Opinions and Emotions in Language. Language Resources and Evaluation, volume 39, issue 2-3, pp. 165-210. 2005