

Extragerea opiniilor implicite din texte economice

Claudiu Cristian Mușat

Universitatea „Politehnica” din București

Splaiul Independenței 313, București

claudiu.musat@cs.pub.ro

Ștefan Traușan-Matu

Universitatea „Politehnica” din București

Splaiul Independenței 313, București

trausan@cs.pub.ro

REZUMAT

În această lucrare propunem o nouă metodă de extragere a opiniilor din texte economice, bazată nu pe apariții explicite de opinii în textele analizate, ci pe polaritatea intrinsecă a indicatorilor economici întâlniți. Metoda are două etape. În primă instanță am creat corpusul de texte pornind de la o colecție de articole ale unei publicații financiare extinsă apoi urmând legăturile care ies din siturile analizate și verificând dacă un subiect economic este discutat în cadrul acestora. În cea de a doua fază vom dovedi că majoritatea textelor economice conțin opiniile personale, chiar dacă opiniile explicite lipsesc, bazându-se pe apariții simultane ale indicatorilor economici și modificărilor de stare pentru a determina polaritatea fiecărui text. Rezultatele acestor cercetări pot ajuta utilizatorii din domeniul economic să navigheze cu mai mare ușurință prin vasta cantitate de informație disponibilă prin gruparea textelor după predicțiile conținute și includerea acestei informații 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 totuși domenii unde această abordare este deficitară, unde trebuie să lărgim perspectivele pentru a detecta și opiniile care nu sunt exprimate direct. Limităm această analiză la textele din domeniul economic.

Extragerea opiniilor

Una dintre lucrările de referință din domeniul extragerii opiniilor din texte este cea a lui Pang și Lee [13], în care documentele sunt clasificate în funcție de opinia prevalentă din întreg textul, obținută ca sumă a tuturor opiniilor întâlnite. Wiebe [15] împarte textele în propoziții subiective și obiective iar mai apoi accentul a fost pus în special pe determinarea la nivel individual a polarității opiniilor întâlnite. Această împărțire 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 fațete ale unor concepte [12], apărând astfel ideea de opinii fațetate. Scopul acestei fracționări 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 situații și exemple relativ simple ca recenziile de filme [15]. Simplitatea în acest caz mai survine și din faptul că obiectul discuției este cunoscut a priori, o largă majoritate a opiniilor prezente în text sunt legate de acel obiect iar părerile exprimate pot fi împărțite cu ușurință în pozitive și 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 și un clasificator Bayesian, au arătat că în mai mult de 60% dintre textele analizate au fost găsite opinii. Trebuie totuși menționat 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 între opiniile din presa financiară și variațiile prețurilor entităților 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 conținutului generat de utilizatori Amigo [2] a arătat că simpla menționare a unui produs într-o analiză poate exprima o atitudine pozitivă față de acel produs. Twitter este un foarte bun exemplu de mediu informal și foarte polarizat unde în general sentimentele nu sunt exprimate explicit dată fiind limitarea dimensiunii mesajelor. Cruz [6] a descoperit că în căutarea părerilor clienților unei companii de cele mai multe ori rezultatele relevante sunt menționi 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 își exprime îndoielile legate de finalitatea unor evenimente viitoare. Ceea ce stă la baza imixtiunii dintre fapte și opinii este empiricismul. Certitudinea rezultată din acumularea de exemple pozitive a fost numită încă din 1958 de către Galbraith [9] „*conventional wisdom*” - înțelepciunea convențională, și a ajuns să se confunde cu orice corp de cunoștințe universal

acceptate fără a avea în mod necesar și o certitudine științifică.

SISTEMUL PROPUȘ

Metoda prezentată în continuare are ca scop extragerea părerilor exprimate în texte economice care sunt prezentate ca și certitudini. Predicțiile economice sunt analizate sub forma combinațiilor între indicatori economici ca „șomajul”, „piața” sau „productivitatea” și termeni ce indică o variație a stărilor viitoare, pe care îi vom denumi modificatori.

Indicatorii sunt împărțiți în doua submulțimi – una care conține indecșii care cresc odată cu economia per ansamblu și una compusă din indecși care scad în perioade de prosperitate economică. Vom denumi în continuare indicatorii din prima mulțime „pozitivi”, iar pe cei ramași „negativi”.

Modificatorii sunt o colecție de n-grame care indică ori o creștere viitoare ori o descreștere a indicatorului de care sunt atașați. Ca și în cazul indicatorilor de mai sus, și modificatorii sunt împărțiți în două mulțimi – pozitivi și negativi, unde cei pozitivi semnaleză o creștere cantitativă iar cei negativi o descreștere.

Termenii relevanți

Termenii și frazele economice considerați cheie – indicatorii – au fost obținuți prin procesări succesive pornind de la dicționarul de termeni economici al *EconomyWatch* iar modificatorii au fost selectați pornind de la o colecție de termeni ce semnaleză o creștere sau descreștere.

Prima modalitate de extindere a mulțimilor inițiale de termeni relevanți se bazează pe *WordNet* [7] și constă în includerea celorlalți termeni din *synset*-urile în care se află termenii inițiali și totodată a termenilor care în descriere (*gloss*) au termeni din mulțimea inițială fără a fi negați.

A doua modalitate de extragere folosește o Rețea Conceptuală (*Conceptual Network Graph*) [4] care întoarce termenii ce tind să apară împreună cu cei inițiali. Am împărțit corpusul prezentat mai jos în două părți egale, iar una dintre jumătăți a fost folosită pentru antrenarea rețelei. Rezultatul fazei de antrenare este un graf bipartit care reprezintă întreaga colecție, în care una din părți este proiecția termenilor folosiți iar cealaltă o reprezentare a documentelor folosite în antrenare. Conexiunile dintre cele două părți ale grafului sunt direct proporționale ca forță cu numărul de apariții ale termenilor folosiți în documentele date. Din totalitatea termenilor aparținând tuturor documentelor din colecție doar un mic număr sunt adăugați mulțimilor inițiale de termeni relevanți. Aceștia sunt obținuți energizând graful din nodurile corespunzătoare termenilor inițiali și 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 funcție de caracteristicile legăturilor dintre ele.

Corpusul folosit

În experimentul prezentat am folosit articolele publice din secțiunea „Finanțe” ale *The Daily Telegraph* și articolele

din subsecțiunile acesteia, obținute pornind de la ultimele texte publicate și urmând apoi legăturile către articolele mai vechi. În total 21106 articole financiare publicate între 2007 și 2010 au fost procesate, iar dintre acestea 499 au fost marcate ca având o încărcătură pozitivă și 216 au fost considerate încărcate negativ. Cum și interpretarea umană a acestor articole este subiectivă, iar adnotarea ar putea avea de suferit din această cauză, ne-am bazat pe înțelegerea adnotatorilor [3]. Astfel pentru experimentele următoare am reținut numai acele articole unde părerile adnotatorilor au coincis. Aceștia au selectat textele cu încărcătură subiectivă – sau care conțin previziuni economice – din secțiunile *Finanțe* (rădăcina arborelui de categorii), *Piețe*, *Locuri de muncă* și *Comentarii*. Alte preprocesări ale corpusului folosit constau în eliminarea sufixelor (*stemming*) tuturor cuvintelor folosind algoritmul lui Porter [14] și eliminarea cuvintelor comune sau de legătură.

Combinarea indicatorilor și modificatorilor

Următoarea operație constă în găsirea polarității prevalente la nivel de document, care reiese din apariții simultane de indicatori economici și modificatori de stare.

Un bun exemplu este legat de discuțiile recente despre creșterea șomajului. Termenul de șomaj este el însuși negativ, iar când modificatorul care îi este atașat este unul pozitiv, de exemplu „va crește” atunci predicția per ansamblu va fi una negativă. În mod asemănător, dacă modificatorul atașat 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 construcției
- O este opinia rezultată
- I este indicatorul folosit
- M este modificatorul atașat

Constrângeri

Nu toate perechile indicator - modificador întâlnite în text sunt opinii exprimate implicit. De exemplu expresia „creșterea șomajului se poate dovedi dezastruoasă pentru economia locală” nu semnaleză o predicție a autorului despre evoluția viitoare a șomajului. Ținând 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 părțile de vorbire pe care le reprezintă cuvintele cheie, am folosit un algoritm de marcare a părții de vorbire pentru fiecare termen din textul analizat (*part of speech tagger*).

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

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

Tabelul 1. Combinații relevante de termeni cheie

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

Totodată, detectarea negațiilor aferente indicatorilor și modificatorilor este crucială pentru acuratețea sistemului întrucât folosirea termenilor negați se dovedește a fi mai întâlnită decât folosirea lor directă.

Un document în care majoritatea perechilor indică o creștere a unui indicator pozitiv sau o scădere a unui indicator negativ va fi marcat ca „pozitiv” și analog un document în care majoritatea perechilor arată o creștere 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 extrași folosind jumătate din articolele marcate ca pozitive sau negative din corpusul *Telegraph Finance*, iar cealaltă jumătate constituie corpusul de testare.

Rezultatele operațiilor de extragere de opinii precum și precizia și 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
Piețe	253	44	128	81
Comentarii	153	32	73	48
Pozitive				
Toate	216	87	50	79
Slujbe	14	4	3	7
Piețe	152	71	31	50
Comentarii	34	11	9	14

Recallul sistemului este proporțional cu dimensiunea corpusului de antrenare, cele mai slabe rezultate fiind obținute la secțiunile *slujbe* și *comentarii* din articolele pozitive, acelea fiind și secțiunile în care numărul de articole este cel mai mic.

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

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

Totodată este important de menționat faptul că pe secțiunile de *comentarii* atât din articolele pozitive cât și din cele negative s-a obținut o acuratețe mai mică decât în celelalte secțiuni, acest rezultat putând fi pus pe seama numărului comparativ mai mare de subiecte abordate în secțiunea de *comentarii*, de la macroeconomie până la finanțe personale.

CONCLUZII ȘI CERCETĂRI VIITOARE

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

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

Alte viitoare îmbunătățiri includ procesarea ironiilor și metaforelor din text, ambele fiind tehnici folosite în largă măsură în textele economice și care nu au fost incluse în prezentul experiment. În plus, sunt numeroase cazuri în care autorul folosește citate externe a căror validitate este apoi contestată, situație care duce la o inversare a polarității opiniilor extrase din acea parte de text. Astfel această metodă trebuie inclusă printre negațiile pe care sistemul trebuie să le poată trata.

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

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

documente care conțin predicții economice de o anumită polaritate sunt cu o probabilitate mai mare conectate cu documente având aceeași polaritate.

O astfel de concluzie ar fi una foarte importantă deoarece ar implica faptul că deși cititorii cred că au o privire de ansamblu nepărtinitoare asupra realității economice, fiind informați din surse diverse din Internet, dacă acele surse sunt interconectate atunci e o mare probabilitate să împartă de fapt și aceeași viziune economică, una părtinitoare. Finalitatea întregului sistem este deci de a ușura informarea utilizatorilor cu privire la temele economice ale zilei, de a facilita accesul la informația relevantă și de a evidenția opiniile distincte.

REFERINȚE

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