Analiza textelor folosind serii de timp

Iulia BADEA

Universitatea "Politehnica" Bucureşti

Splaiul Independenţei, nr. 313, Bucureşti, 060032, Romania

juliebadea@gmail.com

Ştefan TRĂUŞAN-MATU

Universitatea "Politehnica" Bucureşti

Splaiul Independenţei, nr. 313, Bucureşti, 060032, Romania

Institutul de Cercetări în Inteligenţă artificială

Calea 13 Septembrie. nr. 13, Bucureşti, România

stefan.trausan@cs.pub.ro

REZUMAT

Articolul de faţă prezintă o aplicație de analiză a textelor în vederea căutării şi calculării corelațiilor dintre ritmicitățile unor termeni cu frecvență ridicată, folosind serii de timp. Lucrarea îşi propune să prezinte mai întâi unele baze teoretice ale modelului, inclusiv detalii despre preprocesarea textelor studiate folosind tehnici naturale de prelucrare a limbajului, pentru ca în final să fie evidențiate câteva detalii de implementare şi rezultate grafice.

Cuvinte cheie

Prelucrarea limbajului natural, mineritul textelor, serii de timp, detectarea ritmicității.

Clasificare ACM

Metodologii de calcul – Inteligenţă artificială –Metode de invăţare şi de clasificare

INTRODUCERE

Tehnologia limbajului, un domeniu al inteligenţei artificiale care se ocupă cu modul în care programele de calculator sau roboţii înţeleg şi procesează limbajul uman, cuprinde Prelucrarea Limbajului Natural (NLP) şi, ca un sub-domeniu al acestuia, Mineritul Textelor („Text Mining" - TM) [4] adică extragerea de cunoştinţe din texte. Principala întrebare care se impune în acest caz este cum pot fi transformate documentele nestructurate în cunoştinţe, cum pot fi procesate informaţiile pentru a obţine relaţii între concepte sau termeni. Tehnicile de prelucrare a limbajului natural (NLP) joacă un rol important în TM, pentru a descoperi principalele noţiuni şi relaţii între ele [4].

În timp ce diferite prelucrări sunt deja curent folosite în text mining, precum eliminarea cuvintelor nesemnificative („stop words"), găsirea rădăcinii unui cuvânt, identificarea părților de vorbire, găsirea entităţilor cu nume sau alte tehnici din NLP, analiza bazată pe serii de timp a fost folosită foarte rar [3]. Analiza seriilor de timp este un instrument foarte important deoarece aceasta poate prezice evoluţia unui sistem în funcţie de comportamentul său anterior, facilitează asociaţiile şi comparaţiile între seturi de date provenind din diferite perioade de timp şi poate oferi informaţii despre felul în care evoluţia unei variabile în timp poate afecta evoluţia întregului sistem.

Obiectivul general al lucrării este de a prezenta folosirea combinată a TM (a extragerii de cunoştinţe din texte) şi a analizei seriilor de timp pentru a crea noi instrumente de analiză a textelor. Obiectivul principal este de a identifica asemănări/deosebiri, ritmicităţi sau tipare secvenţiale în extragerea de cunoştinţe şi gruparea textelor.

In acest scop, lucrarea prezintă rezultatul unui experiment aplicat pe o serie de articole de ştiri despre alegerile prezidenţiale din SUA 2012, pentru stabilirea corelaţiilor dintre cuvintele cele mai frecvent utilizate, având în vedere apariţiile lor în timp. Analiza propusă urmează o secvență de paşi, începând cu preluarea articolelor de ştiri pe tema propusă şi datele la care au fost scrise, urmată de găsirea cuvintelor frecvente din articolele de ştiri selectate pe perioade de timp specificate.

Următorul pas al abordării este compararea numărului de apariţii al celor mai frecvente cuvinte în timp, folosind o analiză bazată pe serii de timp şi vizualizarea lor grafică. De exemplu, în articolele de ştiri cu privire la alegerile prezidențiale din SUA în 2012, două cuvinte evident foarte frecvent utilizate sunt: *Obama* şi *Romney*. În cele din urmă, corelațiile dintre cuvintele cele mai frecvente menționate şi apariția altor termeni frecvent utilizați în aceleași momente de timp sunt analizate.

Articolul continuă cu două secțiuni, care aduc în discuție câteva aspecte legate de mineritul textelor şi analiza seriilor de timp. A patra secțiune prezintă câteva posibilități de analiză de text folosind tehnici din domeniul seriilor de timp şi se continuă cu o secțiune care cuprinde informații cu privire la colectarea de texte utilizate pentru aplicație şi modul în care aceasta a fost preprocesată. Următoarea secțiune conține descrierea aplicației şi a rezultatelor sale. Lucrarea se încheie cu concluzii şi referințe.

LUCRĂRI CONEXE

Aplicațiile de extragere de cunoştințe din texte (TM) dezvoltate până în prezent se concentrează pe regăsirea de informații, extragerea temelor (conceptelor) principale, detectarea entităților din texte, folosind tehnici de statistică, lingvistică comparată şi de învățare automată („machine learning"). [3]

Prin urmare, accentul a fost pus în principal pe extragerea de informații mai degrabă decât pe găsirea de modele sau regularități între entități sau concepte.

In afara proceselor de derivare lingvistică, prin care iau naştere noi cuvinte, părți de vorbire şi relații gramaticale

noi între termeni, frecvența cuvintelor reprezintă un alt subdomeniu din lingvistică utilizat la găsirea unui set de termeni reprezentativ pentru un text și implicit găsirea ideii centrale a textului. Frecvența termenilor este reprezentată, de cele mai multe ori, prin puncte ordonate cronologic și inter-corelate, în sensul că valorile actuale influențează apariția valorilor viitoare din șir. Această reprezentare a datelor este cunoscută în literatura de specialitate drept **serie de timp**.[8]

Cuvintele cheie găsite în texte au fost reprezentate ca serii de timp, luând în considerare frecvența lor de apariție, care este o valoare ce se poate schimba cu ușurință în timp. Seriilor de timp construite li s-au aplicat tehnici specifice NLP, cu scopul de a găsi tendințe și de a genera previziuni bazate pe datele colectate și a evenimentelor (reprezentate de apariția anumitor termeni) raportate la momente de timp.

APARAT TEORETIC

Mineritul textelor numit și Analiza inteligentă a textelor sau Descoperirea de cunoștințe în texte (KDT) [4], este un domeniu nou, al cărui scop principal este de a extrage automat termenii cheie din documente, fără intervenția umană, urmând algoritmi generici. Cuvintele cheie extrase de algoritm sunt transmise către instrumente de data mining, care clasifică informațiile preluate și le ordonează în nivelul corespunzător într-un sistem complex de management al cunoștințelor, în scopul de a fi utilizate în dezvoltarea ulterioară [4].

O schemă a unui astfel de algoritm generic este descrisă în *Figura 1*, care marchează principalele etape urmate de o aplicație de analiză de text. Punctul de plecare al algoritmului constă într-un set de documente care cuprinde diverse texte, care sunt, într-o etapă ulterioară: prelucrate, analizate și descompuse folosind tehnici specifice, unele dintre ele fiind exemplificate în figură [4].

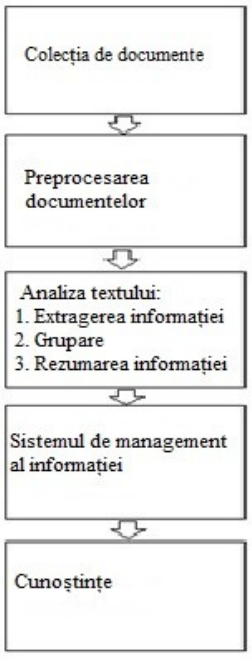


Figura 1. Exemplu de proces de analiză a textelor [4]

ANALIZA SERIILOR DE TIMP

O serie de timp este definită ca un set de valori fizice, măsurate la diferite momente într-un interval de timp [9]. Observarea acestor valori de-a lungul timpului se face folosind metoda seriilor de timp, care utilizează metode matematice și statistice. Cu toate acestea, tehnicile tradiționale sunt adesea limitate, din cauza faptului că pornesc de la ipoteza potrivit căreia valorile luate în considerare sunt identic și independent distribuite în timp, în realitate acestea putând varia în timp, uneori într-un mod care nu poate fi nici prognozat, nici estimat [7].

Descrierea unei serii de timp cuprinde identificarea componentelor sale principale, care pot apărea simultan sau unele dintre ele pot lipsi: tendința (ascendentă sau descendentă), componenta sezonieră sau periodică, analiza spectrală, de asemenea numită și componenta neregulată. Toate acestea pot manifesta unele fluctuații imprevizibile [10, 11]. Acest proces de identificare a componentelor principale este, de asemenea, cunoscut sub numele de descompunere a seriei de timp.

Există o gamă largă de domenii în care este utilizată analiza seriilor de timp, începând cu: previziuni, analiză financiară, biomedicină, motoare de căutare, programe de recunoaștere a vorbirii pentru aplicații Android [11].

Printre cele mai utilizate instrumente de analiză se numără: limbajul R (cu pachetele *ts* și *zoo* [17]), Matlab și Java (*timeseries public class*). Matlab furnizează două tipuri de obiecte de serii de timp: *timeseries*, care stochează o singură serie de timp, adică o asociere între valorile de date și momentele de timp corespunzătoare, și *tscollection*, pentru o colecție de serii de timp cu același interval de timp, facilitând procesarea lor sincronizată [16].

ANALIZA TEXTELOR FOLOSIND SERII DE TIMP

Ideea de a folosi seriile de timp este legată de evoluția numărului de concepte în secvențe de texte. În scopul de a face acest lucru, primul pas a fost identificarea principalelor concepte ("subiecte"), urmat de cea mai importantă etapă: recunoașterea entităților (termeni care apar cel mai frecvent în textele studiate), folosind tehnici NLP, pentru a genera serii de timp care corespund acestor concepte.

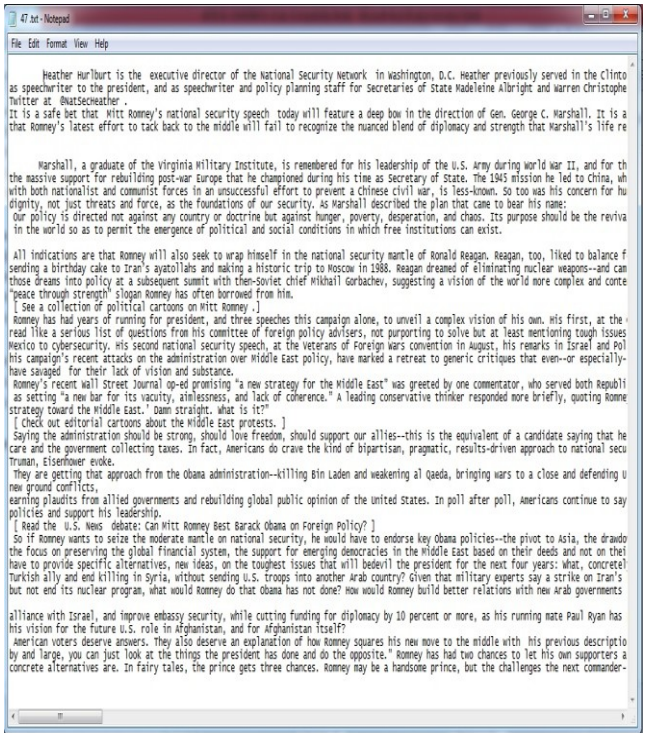
Seriile de timp generate inițial pentru numele candidaților Obama și Romney sunt apoi analizate prin diferite metode și rutine incluse în unele pachete din R, care se ocupă cu reprezentarea și analiza datelor și obiectelor de timp, cum ar fi *tseries* sau *zoo*.

Uneori, un anumit tip de date (în multe cazuri, date de înaltă frecvență) este neregulat, astfel că timpul dintre două observații nu este întotdeauna egal, cum ar fi distanța dintre două articole de știri, care poate varia de la câteva ore până la chiar câteva zile. Există mai multe pachete în R care tratează această problemă, unul dintre ele fiind pachetul numit zoo (Z[eilies] 's, după numele de familie al autorului original), prin intermediul căruia se creează un obiect de tip *zoo* în loc de un obiect *ts* (obiectele de tip *ts* se ocupă doar cu reprezentarea obiectelor cu frecvență regulată).

În plus faţă de pachetele deja menţionate, s-au dovedit a fi foarte utile unele funcţii incluse în: *Rcurl* (o interfaţă R, care oferă facilităţi HTTP, cum ar fi: descărcarea sau încărcarea fişierelor de pe servere Web, formularele poştale, utilizarea HTTPSecure) [12], *XML* (analizează fişiere XML sau HTML, URL-uri şi şiruri de caractere, folosind unul dintre cele două modele: Modelul Document - Obiect sau modelul arborescent, de tip *tree*) [14], *TM* (oferă metode de a preprocesa textele şi de a le transforma în documente, prin eliminarea spaţiilor multiple, a semnelor de punctuaţie, schimbarea capitalizării sau eliminarea cuvintelor de legătură) [15], care vor fi ilustrate cu exemple în acest articol.

ACHIZIŢIA DE DATE ŞI PREPROCESAREA LOR

Dintre posibilităţile de aplicare ale seriilor de timp în analiza limbajului am considerat problema detectării ritmicităţii ca fiind cea mai interesantă pentru studiile noastre. Prin urmare, punctul de plecare al analizei noastre a constat în crearea unei colecţii de texte (ştiri), scrisă între 1 octombrie 2012 şi 31 noiembrie 2012, ce formează un *corpus* (o colecţie de documente de tip text), care au fost salvate împreună cu datele lor de apariţie. Colecţia cuprinde 53 de articole, fiecare dintre acestea fiind salvate într-un fişier separat, în scopul facilitării aplicării tehnicilor NLP asupra textelor extrase. Cele mai multe dintre ele sunt scrise zilnic, dar există unele excepţii când mai multe articole au fost scrise în aceeaşi zi. Articolele selectate au, în medie, 500-600 de cuvinte. *Figura 2* prezintă un articol de ştiri inclus în acest *corpus*.


47.txt - Notepad

File Edit Format View Help

Heather Hurlburt is the executive director of the National Security Network in Washington, D.C. Heather previously served in the Clinto as speechwriter to the president, and as speechwriter and policy planning staff for Secretaries of State Madeleine Albright and Warren Christophe Twitter at @NatSecHeather .
It is a safe bet that Mitt Romney's national security speech today will feature a deep bow in the direction of Gen. George C. Marshall. It is a that Romney's latest effort to tack back to the middle will fail to recognize the nuanced blend of diplomacy and strength that Marshall's life re

Marshall, a graduate of the Virginia Military Institute, is remembered for his leadership of the U.S. Army during World War II, and for th the massive support for rebuilding post-war Europe that he championed during his tine as Secretary of State. The 1945 mission he led to China, wh with both nationalist and communist forces in an unsuccessful effort to prevent a Chinese civil war, is less-known. So too was his concern for hu dignity, not just threats and force, as the foundations of our security. As Marshall described the plan that cane to bear his name:
Our policy is directed not against any country or doctrine but against hunger, poverty, desperation, and chaos. Its purpose should be the reviva in the world so as to permit the emergence of political and social conditions in which free institutions can exist.

All indications are that Romney will also seek to wrap himself in the national security mantle of Ronald Reagan. Reagan, too, liked to balance f sending a birthday cake to Iran's ayatollahs and making a historic trip to Moscow in 1988. Reagan dreamed of eliminating nuclear weapons--and can those dreams into policy at a subsequent summit with then-Soviet chief Mikhail Gorbachev, suggesting a vision of the world more complex and conte "peace through strength" slogan Romney has often borrowed from him.
[See a collection of political cartoons on Mitt Romney .]
Romney has had years of running for president, and three speeches this campaign alone, to unveil a complex vision of his own. His first, at the read like a serious list of questions from his committee of foreign policy advisers, not purporting to solve but at least mentioning tough issues Mexico to cybersecurity. His second national security speech, at the Veterans of Foreign Wars convention in August, his remarks in Israel and Pol his campaign's recent attacks on the administration over Middle East policy, have marked a retreat to generic critiques that even--or especially- have savaged for their lack of vision and substance.
Romney's recent Wall Street Journal op-ed promising "a new strategy for the Middle East" was greeted by one commentator, who served both Republi as setting "a new bar for its vacuity, aimlessness, and lack of coherence." A leading conservative thinker responded more briefly, quoting Romne strategy toward the Middle East.' Damn straight. What is it?"
[Check out editorial cartoons about the Middle East protests.]
Saying the administration should be strong, should love freedom, should support our allies--this is the equivalent of a candidate saying that he care and the government collecting taxes. In fact, Americans do crave the kind of bipartisan, pragmatic, results-driven approach to national secu Truman, Eisenhower evoke.
They are getting that approach from the Obama administration--killing Bin Laden and weakening al Qaeda, bringing wars to a close and defending U new ground conflicts,
earning plaudits from allied governments and rebuilding global public opinion of the United States. In poll after poll, Americans continue to say policies and support his leadership.
[Read the U.S. News debate: Can Mitt Romney Best Barack Obama on Foreign Policy?]
So if Romney wants to seize the moderate mantle on national security, he would have to endorse key Obama policies--the pivot to Asia, the drawdo the focus on preserving the global financial system, the support for emerging democracies in the Middle East based on their deeds and not on thei have to provide specific alternatives, new ideas, on the toughest issues that will bedevil the president for the next four years: what, concrete? Turkish ally and end killing in Syria, without sending U.S. troops into another Arab country? Given that military experts say a strike on Iran's but not end its nuclear program, what would Romney do that Obama has not done? How would Romney build better relations with new Arab governments

alliance with Israel, and improve embassy security, while cutting funding for diplomacy by 10 percent or more, as his running mate Paul Ryan has his vision for the future U.S. role in Afghanistan, and for Afghanistan itself?
American voters deserve answers. They also deserve an explanation of how Romney squares his new move to the middle with his previous descriptio by and large, you can just look at the things the president has done and do the opposite." Romney has had two chances to let his own supporters a concrete alternatives are. In fairy tales, the prince gets three chances. Romney may be a handsome prince, but the challenges the next commander-

Figura 2. Articol rezultat în urma extragerii şi parsării textului de ştire din ziar

Prima etapă în preprocesarea datelor o reprezintă prelucrarea codului HTML al articolelor, urmat de identificarea adnotărilor HTML şi extragerea corpului articolului, respectiv data la care acesta a fost scris.

Al doilea pas în procesarea textelor cuprinde: înlocuirea spaţiilor duble cu un singur spaţiu, precum şi eliminarea adnotărilor HTML, semnelor de punctuaţie şi schimbarea majusculelor în minuscule. Acest lucru a fost realizat folosind funcţii din pachetul *text mining* (disponibil la [18]).

După ce articolele au trecut prin operaţii de preprocesare, a fost creată matricea de termeni asociată documentului, şi începând de aici, un set de date care conţine cuvintele cele mai frecvent utilizate, ordonate în funcţie de numărul lor de apariţii. Acest proces de selecţie reduce dimensiunea spaţiului de cuvinte, facilitând considerabil căutarea celor mai importanţi termeni. [9]

CÂTEVA REZULTATE

După detectarea celor mai frecvente cuvinte menţionate în timpul alegerilor, etapa următoare a constat în reprezentarea acestora într-o structură care afişează termenii în funcţie de frecvenţa lor în articole, Cloud Word, generat cu ajutorul pachetelor grafice *wordcloud* şi *RcolorBrewer*, ce permit alegerea mai multor opţiuni de afişare a imaginii, precum selectarea paletei de culori, fontului sau dimensiunii cuvintelor (a se vedea *Figura 3*).

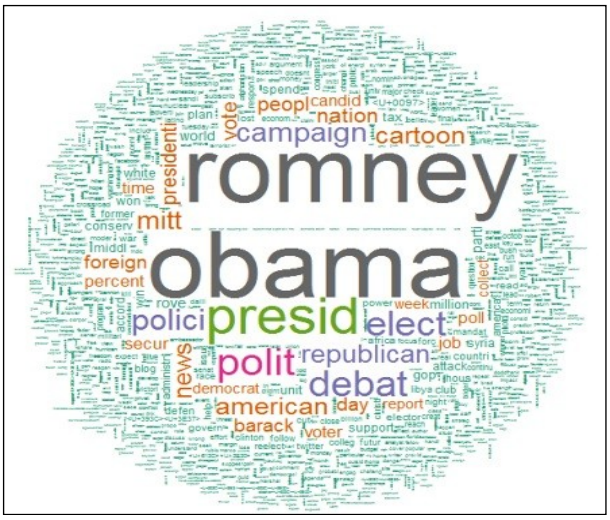


Figura 3. Word Cloud

Ritmicitatea în seria de articole (numărul de apariţii al unui cuvânt raportat la un interval de 24 de ore) este reprezentată în diagramele din variaţiile temporale, aşa cum se arată în *Figura 4*. *Aceasta* prezintă şi alte cuvinte utilizate frecvent în afară de primii doi termeni ca frecvenţă de utilizare (*Obama* şi *Romney*), corelate cu numărul lor total de apariţii în corpusul creat în prima parte a aplicaţiei.

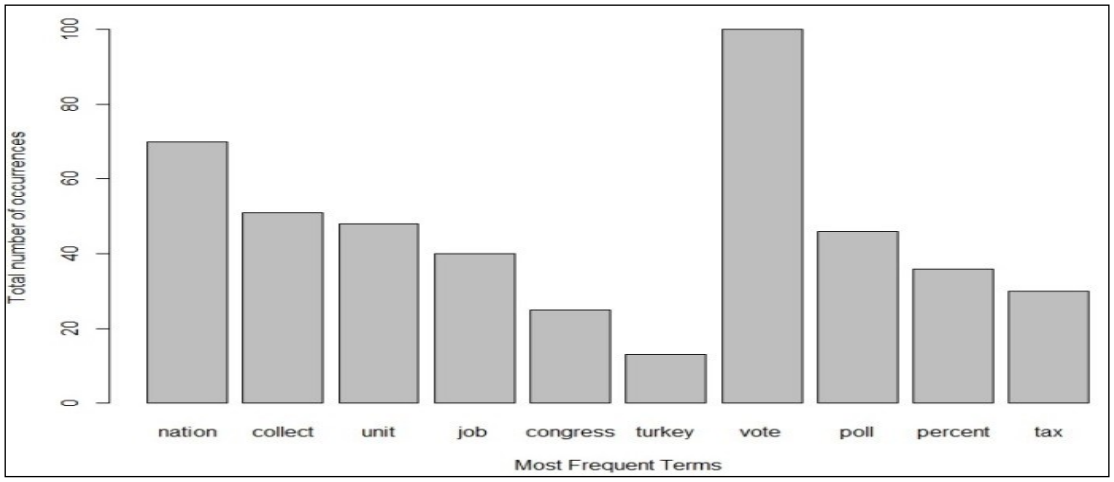


Figura 4. Cei mai frecvent utilizați termeni

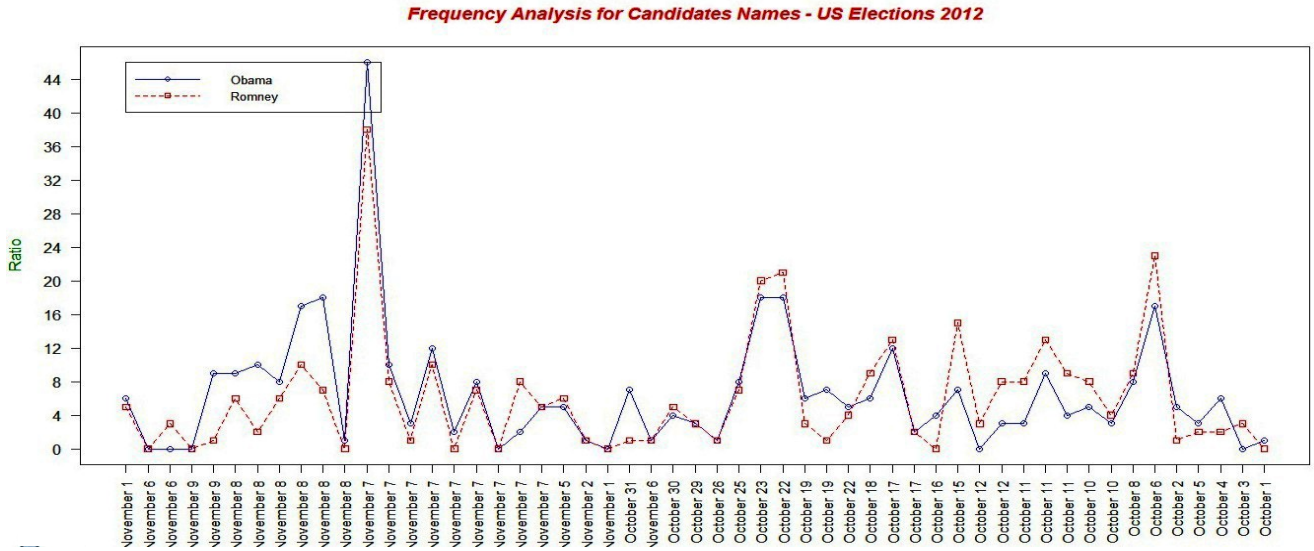


Figura 5. Analiza comparativă a frecvențelor termenilor Obama și Romney în articolele din corpus

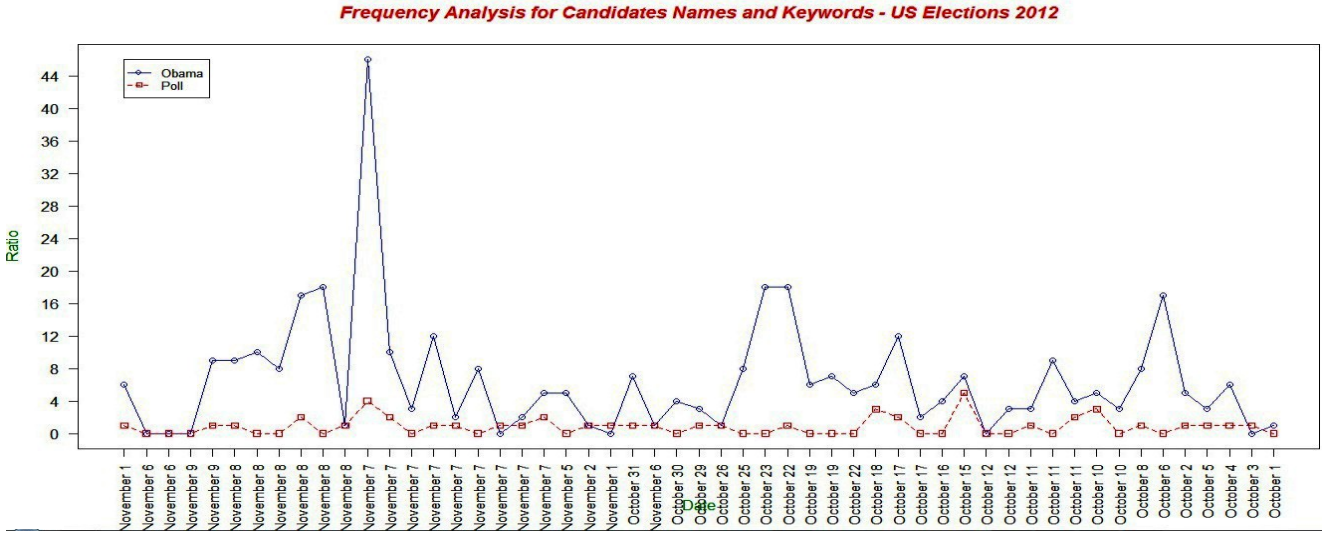


Figura 6. Serii de timp cu aparițiile lui Obama, respectiv Poll în articolele din perioada 1 octombrie 2012 - 31 noiembrie 2012

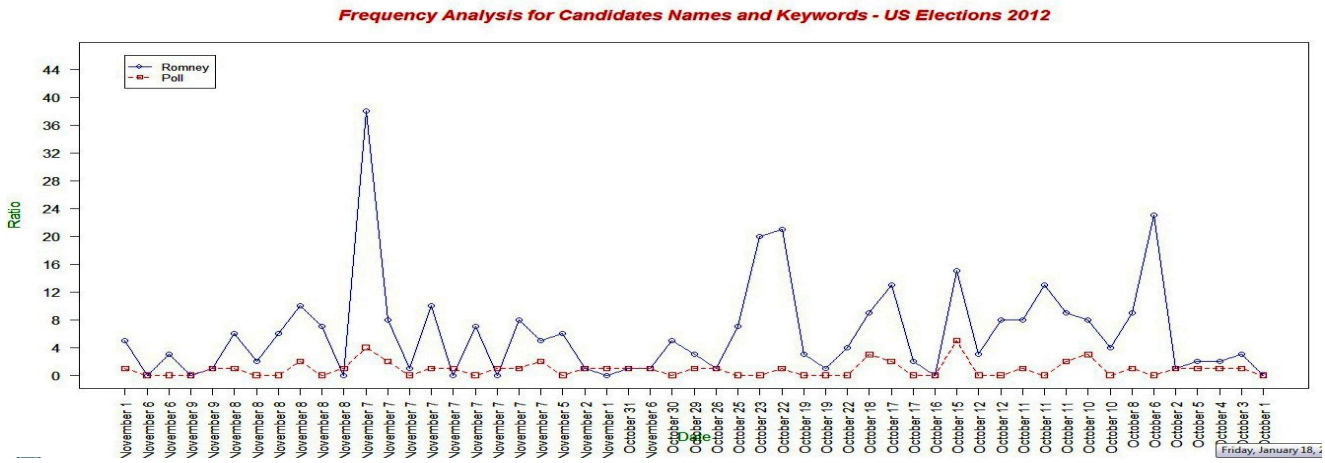


Figura 7. Analiza comparativă a frecvențelor termenilor Romney şi Poll în articolele din corpus

Figura 5 include două grafice ce reprezintă seriile de timp destinate să măsoare, în momente succesive în timp, în diferite articole, numărul de apariții ale cuvintelor *Obama*, respectiv *Romney*.

Ultimele două grafice (*Figura 6* și *Figura 7*) ilustrează, prin comparație, numărul de apariții al unui cuvânt aleator ales din corpus, cum ar fi *Poll*, împreună cu *Obama*, respectiv cu *Romney*. Prin observarea graficului, este uşor să concluzionăm că vârful pentru numărul de apariții ale lui *Obama* corespunde celui mai mare număr de apariții pentru *Poll*, acest lucru nefiind la fel de evident pentru perechea de cuvinte *Romney-Poll*. Aşadar, există o corelație directă între aparițiile în articole a termenilor *Obama* și *Poll*, cei doi termeni fiind prezenți aproape mereu simultan.

Partea finală a aplicației are ca scop principal detectarea corelației între ritmicitățiile unor serii de timp pentru cei mai importanți termeni: *Obama* și *Romney*.

Corelația dintre două serii poate fi determinată pe baza calculului unui coeficient obținut în mod explicit, folosind *cor,* o funcție din pachetul R. Coeficientul ce trebuie calculat este decis pe baza celui de-al patrulea parametru al funcției, care poate lua una dintre următoarele valori: Pearson (implicit), Kendall, sau Spearman [12].

Dacă metoda aleasa este "Kendall" sau "Spearman", corelațiile dintre cele două serii de timp construite pentru Obama, respectiv Romney, sunt de aproximativ 0,5, în timp ce dacă pe aceleași date se aplică funcția *cor* cu metoda Pearson, este returnată o valoare mai mare, egală cu 0,8. Valorile mai mari de 0,6 reprezintă corelații pozitive puternice. Ideea centrală din spatele calcului corelațiilor între date este că o valoare a coeficientului de corelație măsurată pe un eșantion de date se pastrează pentru întreaga populație.

În afară de cor, R oferă, de asemenea, alte seturi de teste pentru a verifica dacă două serii de timp sunt semnificativ diferite sau dacă există corelații între frecvențele de apariție pentru fiecare termen: testele Kolmogorov-Smirnov și Shapiro, care verifică dacă datele sunt distribuite normal.

CONCLUZII

Mineritul textelor nu furnizează întotdeauna rezultate satisfăcătoare folosind tehnicile de prelucrare a limbajului natural clasice, ca urmare a evoluției numărului de concepte în secvențe de texte, iar un model de serii de timp poate rezolva această problemă. Un astfel de model a fost conceput și folosit pentru punerea în aplicare a unui sistem care analizează articole de ştiri, în scopul de a evalua corelațiile dintre cuvintele cele mai frecvent utilizate, luând în considerare evoluția lor de timp.

Internetul reprezintă o sursă importantă pentru preluarea articolelor de ştiri în funcție de subiect și gruparea acestora într-o structură în scopul de a fi analizate, identificând principalele concepte și cele mai importante entități folosind elemente de procesare a limbajului natural (NLP) la nivelurile: lexical, semantic și al discursului.

Graficele și primele teste efectuate au demonstrat că punerea în aplicare a analizei seriilor de timp oferă o reprezentare sugestivă a corelaţiilor dintre cuvintele cele mai frecvent utilizate din seria de articole de ştiri selectate pentru experiment, având în vedere aparițiile lor în timp. Depistarea unor corelații între termenii cu cea mai mare frecvență de apariție din texte, având în vedere evoluția acestora în timp, reprezintă un prim pas spre realizarea unor algoritmi care prelucreaza automat textele, în scopul interpretării și analizei lingvistice sau statistice, cu mai multă rapiditate față de oameni.

REFERINŢE

1. M. Lundholm, “Introduction to R's time series facilities,” Stockholm University, Version 1.3, pp.1-31, September 22, 2011.
2. I. Feinerer, Introduction to the tm Package - Text Mining in R, 2012, pp.1-8.
3. A. Kao and S. Poteet, “Text Mining and Natural Language Processing-Introduction for the Special Issue” in SIGKDD Explorations, vol. VII, 2004, pp. 1-2.

4. V. Gupta and G. S. Lehal, "A survey of text Mining Techniques and Applications" in "Journal of Emerging Technologies in Web Intelligence", vol. I, no. I, pp. 60-61, August 2009.
5. Walter Zucchini, Oleg Nenadi, "Time Series Analysis with R-Part I," pp . 8-23.
6. V. Lavrenko, M. Schmill, D. Lawrie, P. Ogilvie, D. Jensen, J. Allan, "Mining of Concurrent Text and Time Series", Department of Computer Science, University of Massachusetts, Amherst, MA 01003
7. R. H. Shumway, David. S. Stoffer, "Time Series Analysis and Its Applications With R Examples", Third edition, Springer, 2011, pp. 1-2
8. A. Gerow, K. Ahmad, „*Proceedings of COLING 2012: Posters"*, COLING 2012, Mumbai, pages 381–390.
9. C. Chen - Drexel University, USA, F. Ibekwe - Université de Lyon 3, France, E. SanJuan -Université d'Avignon, France, C. Weaver - Penn State University, USA, "Visual Analysis of Conflicting Opinions"
10. http://www.abs.gov.au/websitedbs/D3310114.nsf/home/Time+Series+Analysis:+The+Basics, accesat pe 17 Martie 2013
11. http://www.slideshare.net/sachin.bme.abs/time-series-analysis, accesat pe 17 Martie 2013
12. http://stat.ethz.ch/R-manual/R-patched/library/stats/html/cor.html, accesat pe 17 Martie 2013
13. http://www.omegahat.org/RCurl/, accesat pe 17 Martie 2013
14. http://www.omegahat.org/RSXML/, accesat pe 17 Martie 2013
15. http://tm.r-forge.r-project.org/, accesat pe 17 Martie 2013
16. http://www.mathworks.com/help/matlab/data_analysis/time-series-objects.html, accesat pe 17 Martie 2013
17. http://www.r-project.org/, accesat pe 17 Martie 2013
18. http://cran.r-project.org/web/packages/tm, accesat pe 17 Martie 2013