**Analiza sentimentelor privind comentariile de pe Goodreads**

Amarghioalei Ioana – Cristina şi Roman Otilia - Izabela

**Scurtă descriere**

Proiectul vizează realizarea analizei sentimentelor privind comentariile la cărţi de pe site-ul “Goodreads”. În vederea realizării analizei sentimentelor am folosit programul Python, iar pentru extragerea comentariilor şi a titlului cărţilor am folosit programul ScrapeStorm.

**Arhitectura proiectului**

**Fig. 1 Arhitectura proiectului**

**Detalii proiect**

**Motivaţia**

Analiza sentimentelor se referă la detectarea atitudinii față de obiecte, persoane, fenomene, privită din perspectiva celui care emite acest sentiment, celui care este ținta lui, tipului de atitudine și a textului care conține atitudinea respectivă.

Se poate observa o creștere masivă a numărului de lucrări axate pe analiza sentimentelor și extragerea opiniei în ultimii ani. Peste 7.000 de lucrări pe această temă sunt prezente și aproximativ 99% din lucrări au apărut după 2004, făcând analiza sentimentelor unul dintre domeniile cu cea mai rapidă creștere în ariile de cercetare. Această tehnică are astăzi o utilitate în foarte multe domenii, de la prezicerea piețelor financiare la reacții față de atacuri teroriste. Atunci când a început să se dezvolte la jumătatea anilor 2000, se concentra pe recenziile de produse disponibile pe Web. Analiza sentimentelor contribuie chiar și la detectarea ironiei sau a emoțiilor negative precum frica sau mânia.

Luând în considerare utilitatea în diverse domenii cât şi importanţa din ce în ce mai ridicată în ultimii ani în ceea ce priveşte analiza sentimentelor, am ales şi noi realizarea unei astfel de analize în proiect.

**Provocări**

Un text poate fi definit de cinci factori în ceea ce privește analiza sentimentelor:

* Obiectul – produsul, serviciul, individul, organizația, evenimentul sau topic ce urmează să fie analizat;
* Atributele – componentele specifice și proprietățile obiectelor (de exemplu, pentru componente: baterie, touch screen, căști sau pentru proprietate: mărime, greutate, viteză de procesare);
* Cel care emite opinia – persoana sau organizația care exprimă sentimentele (spre exemplu un cititor de carte);
* Orientarea sau polaritatea opiniei – poziția general referitoare la opinie (pozitivă, negative, neutră);
* Intensitatea opiniei – nivelul, scala sau puterea părerii (entuziasmat, vessel, încântat sau nesatisfăcut).

Pentru a obține informații de valoare, corecte și utile de la o bucățică de text, este important nu doar să se identifice fiecare dintre aceste cinci elemente individual, dar și să se înțeleagă modul cum ele lucrează împreună pentru a obține întregul context. Deoarece procesarea cuvintelor cheie doar identifică sentimental unui singur cuvânt specific, în general eșuează în a oferi toate elementele necesare înțelegerii întregului sens.

NLP folosește tehnici de învățare automata și data mining pentru a oferi o imagine mai complete, dar complexitatea limbajului face dificilă asigurarea unor algoritmi eficienți care să analizeze tonul și contextul. Factorii limitatori sunt nuanțele gramatice, înțelesuri implicite ce provin din limbajul semnelor, greșeli gramaticale sau regionalisme. În timp ce oamenii sunt mai bine pregătiți pentru a identifica toate cele elemente, procesarea manuală prezintă propriile provocări, mai ales referitoare la viteză și cantitate.

**State of the art**

Cercetările și metodele au evoluat o dată cu canalele de social media. La început, era vorba doar despre analiza unor etichete pentru detectarea caracterului pozitiv, negativ sau uneori neutru pentru review-uri în engleză. Astăzi se referă la detectarea subiectivității, analiza sentimentelor pe limbaj figurat, clasificare bazată pe topicuri, polaritatea evenimentelor (plăcute sau neplăcute), clasificarea emoțiilor și cel mai recent detectarea argumentelor, într-o mare varietate de limbi și de surse de date.

Hu și Liu prezintă o abordare bazată pe limbajul natural pentru a extrage caracteristici ale review-urilor consumatorilor. Metoda lor a constat într-un tag cu părți de vorbire care să dividă cuvintele în categorii lexicale și doar semnificația semantică a adjectivelor este considerată în algoritm. Acesta poartă denumirea de ”FBS”.

Wilson, Wiebe and Hoffman au folosit o analiză la nivel de frază folosind un algoritm de învățare automată, care judeca dacă o expresie este polară sau neutră. Contribuția notabilă a acestei lucrări o reprezintă numeroasele caracteristici de îmbunătățire a clasificării polarității sentimentelor pornind la nivel de frază (de exemplu adverbe care negau sau schimbau expresia sentimentului).

Articolul lui Dave, Lawrence și Pennock prezintă o abordare a interogării opiniilor, pornind de la păreri asupra produselor obținute de pe internet și analizate folosind tehnici de programare lingvistică naturală. Opiniile erau divizate în sentimente pozitive sau negative de către algoritm, dar erau luate în considerare și opiniile personale și contextul.

Popescu și Etzioni au introdus un sistem de extragere a informațiilor nesupervizat, aducând îmbunătățiri modelului propus de Hu și Liu pentru review-urile consumatorilor. Contribuția majoră față de abordările anterioare a fost folosirea unui motor de căutare de pe internet pentru calcularea scorului PMI (Point Wise Mutual Information), care analiza dacă un substantiv putea fi considerat ca parte a unei caracteristici a produsului.

Una dintre lucrările cele mai timpurii a fost realizată în anul 2002 de către Pang, Lee și Vaithyanathan. Trei clasificatori de învățare automata diferiți sunt folosiți în analiza sentimentelor la nivel de documente, în mod particular pentru a analiza review-urile la filme și pentru a le clasifica în sentimente pozitive sau negative. Toți clasificatorii au bătut alegerile aleatoare și selecția umană în evaluarea experimentală.

Dacă la început, se punea accent mai ales pe analiza sentimentelor la nivel de opinii ale consumatorilor, în ultimii ani platformele de social media au căpătat tot mai mare popularitate. Pak și Paroubek au prezentat o metodă pentru colectarea automată a unui corpus de la microbloguri și le-au folosit pentru a crea un clasificator de sentimente. În acest caz, corpusul este realizat prin Twitter. Autorii susțin că abordarea lor poate fi adaptată în multe limbi, însă ei au realizat acest lucru doar în engleză.

Tumasjan și colaboratorii au analizat în jur de 100.000 de comentarii de pe Twitter ca un predictor pentru alegerile electorale. Ei au descoperit că proporția mențiunilor candidaților a oglindit rezultatele electorale destul de bine, cu toate acestea având o precizie mai scăzută decât sondajele de opinie. O altă descoperire interesantă a fost că 40% dintre mesaje erau postate de 4% dintre utilizatori.

Datele de pe Twitter au constituit sursă de analiză și pentru O’Connor, Balasubramayan, Routledge și Smith. În timp ce corelația dintre măsurarea sentimentelor și datele din sondaje varia în diverse seturi de date, trendurile cele mai acute erau capturate de analiza sentimentelor.

În ceea ce privește instrumentele folosite pentru analiză unul dintre ele este cel supranumit ”Semantic Orientation CALculator (SO-CAL) prin completarea regulilor precum intensificare și negare a metodelor deja existente. Noua abordare a fost apoi validate, evaluate și comparată cu câteva instrumente de analiză a sentimentelor deja existente, într-un experiment care analiza diverse seturi de date prin folosirea tuturor abordărilor. Seturile de date utilizate au fost diverse: review-uri de produse, comentarii de pe platforme sociale, articole de ziar și titluri de presă.

În 2012, Thelwall, Buckley and Paltoglou au evaluat un instrument numit ”SentiStrength 2”, ce era un clasificator bazat pe lexicon folosit împreună cu câțiva algoritmi de învățare automată. Acest instrument a fost evaluat cu o gamă largă de seturi de date care foloseau texte scurte, comentarii de pe Youtube și Twitter. Acest program reușește cu succes să analizeze textele scurte de pe internet și reușește chiar să le izoleze de context.

SentiWordNet este un lexicon generat automat de către Baccianella, Esuli and Sebastiani să fie folosite specific în analiza sentimentelor. Lexiconul a fost generat folosind un algoritm de învățare semi-automată și evaluat prin compararea valorii semantice generate a cuvintelor cu un set manual de sinonime.

**Scurtă descriere a arhitecturii**

Primul pas a fost extragerea comentariilor de pe site, cu ajutorul programului ScrapeStorm. Am reuşit să extragem 1000 comentarii. Am creat un corpus cu comentariile cititorilor, iar mai apoi, am început analiza acestora în Python.

În vederea realizării analizei, am început cu tokenizarea comentariilor, am obţinut scoruri pentru sentimente, am eliminat cuvintele de legătură, dar şi caracterele speciale, apoi am continuat cu lemmatizarea comentariilor. După toate acestea am clasificat comentariile pe baza scorurilor obţinute anterior în comentarii negative, neutre şi pozitive.

Următorul obiectiv pe care am propus a-l atinge a fost să verificăm acurateţea modelului folosind mai multe metode. Am folosit Naïve Bayes, regresia logistică, Random Forest şi arbori de decizie. Am constatat că folosind regresia logistică obţinem cea mai mare valoare a acurateţii, mai exact 74%.

Am testat mai apoi cu ajutorul modelului de regresie logistică noi comentarii pentru a vedea în ce categorii le va încadra. Am observat că face o clasificare bună în ceea ce priveşte noile comentarii.

**Managementul proiectului**

Amarghioalei Ioana – Cristina :

* Eliminare cuvinte de legătură
* Eliminare caractere speciale
* Lemmatizare comentarii
* Naive Bayes, Regresie logistică, Random Forest, Decision Tree
* Acurateţe model

Roman Otilia – Izabela:

* Extragere comentarii
* Tokenizare comentarii
* Scoruri pentru sentimente
* Clasificare comentarii
* Testare comentarii noi cu modelul de regresie logistică

**Termene limită**

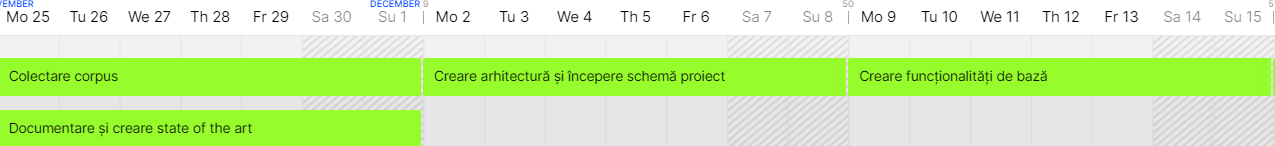
01.12.2019 – colectare corpus, documentare și creare state of the art

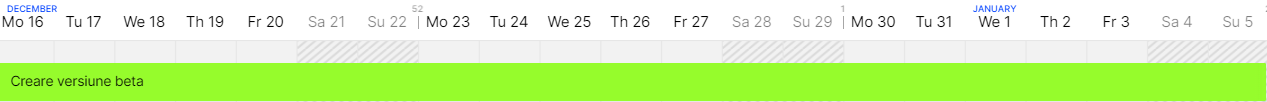
08.12.2019 – creare arhitectură și începere schemă proiect

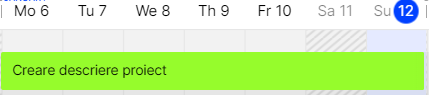
15.12.2019 - creare funcționalități de bază

05.01.2020 – creare versiune beta

12.01.2020 – definitivare document cu descriere proiect



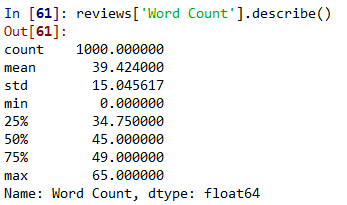




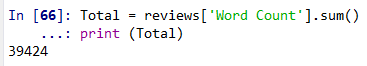
**Fig. 2 Milestones**

**Metrici ai proiectului**

Așa cum se poate observa din figurile de mai jos, în ceea ce privește numărul total de cuvinte procesate din 1000 de review-uri, acesta este de 39.424, cu un număr maxim de 65 de cuvinte pentru un review curățat. Media este de 39,24 de cuvinte, apropiată de mediana de 45 de cuvinte.

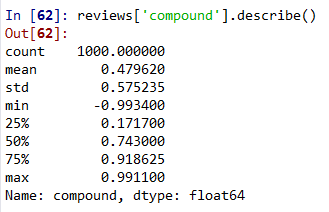


**Fig. 3 Analiza descriptivă pentru numărul de cuvinte**

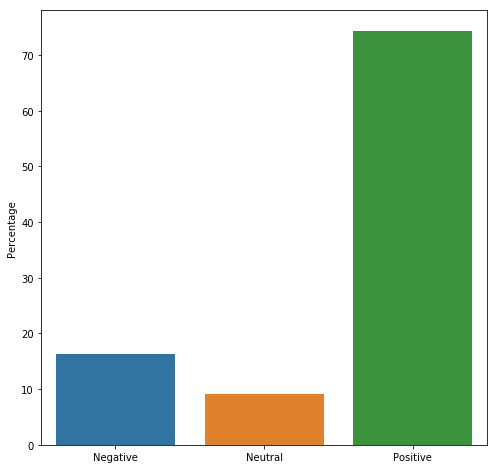


**Fig. 4 Totalul cuvintelor procesate**

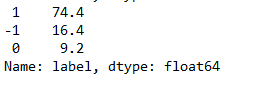
Referitor la scorul comentariilor procesate, se poate afirma că 9,2% au fost încadrate neutre, 16,4% pozitive, în timp ce marea majoritate 74,4% au fost comentarii negative. Media comentariilor este de 0,47 în timp ce mediana este 0,74.



**Fig. 5 Analiza descriptivă pentru scorurile comentariilor**



**Fig. 6 Graficul distribuției tipurilor de comentarii**



**Fig. 7 Procente distribuție comentarii**

**Referințe bibliografice**

A. Esuli and F. Sebastiani, “SENTIWORDNET: A high-coverage lexical resource for opinion mining,” Evaluation, pp. 1–26, 2007

A. Pak and P. Paroubek, “Twitter as a Corpus for Sentiment Analysis and Opinion Mining.,” in LREc, 2010, vol. 10.

A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, “Predicting elections with twitter: What 140 characters reveal about political sentiment.,” ICWSM, vol. 10, no. 1, pp. 178–185, 2010.

A.-M. Popescu and O. Etzioni, “Extracting product features and opinions from reviews,” in Natural language processing and text mining, Springer, 2007, pp. 9–28.

B. Liu, “Sentiment analysis and opinion mining,” Synthesis lectures on human language technologies, vol. 5, no. 1, pp. 1–167, 2012.

B. O’Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith, “From tweets to polls: Linking text sentiment to public opinion time series.,” ICWSM, vol. 11, no. 122–129, pp. 1–2, 2010.

B. Pang and L. Lee, “Opinion mining and sentiment analysis,” Foundations and trends in information retrieval, vol. 2, no. 1–2, pp. 1–135, 2008.

K. Dave, S. Lawrence, and D. M. Pennock, “Mining the peanut gallery: Opinion extraction and semantic classification of product reviews,” in Proceedings of the 12th international conference on World Wide Web, 2003, pp. 519–528

S. Baccianella, A. Esuli, and F. Sebastiani, “SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining.,” in LREC, 2010, vol. 10, pp. 2200–2204