UNIVERSITATEA BABEŞ-BOLYAI, CLUJ-NAPOCA, ROMÂNIA, FACULTATEA DE MATEMATICĂ ŞI INFORMATICĂ

Tehnici de Modelare a Datelor şi Diverse Aplicaţii ale Analizei Sentimentelor

Sumarul tezei de doctorat

Student doctorand: Sergiu Limboi
Coordonator ştiinţific: Prof. dr. Laura Dioşan

2024

Cuvinte cheie: Analiza sentimentelor, Twitter, Extragere de caracteristici din text, Calcul afectiv, Preprocesarea datelor, Învăţare automată

Rezumat

În ultimii ani, social media a devenit un mediu popular în care toată lumea își poate exprima ideile despre evenimente, celebrități, politică sau educație. Pas cu pas, opiniile și știrile postate aici au devenit o sursă esențială de informații pentru mulți oameni. Analiza acestor date poate fi esențială pentru a prezice și defini strategii de marketing, pentru a influența evaluările politice sau opiniile sociale sau pentru a crea recomandări relevante pentru potențialii clienți.

Analiza sentimentelor este un domeniu ce implică mai multe activități care pot fi folosite pentru această analiză. Una dintre acestea se numește detectarea polarității, care se concentrează pe determinarea sentimentului dintr-un anumit tip de date. Această sarcină este potrivită pentru studierea și evaluarea datelor textuale furnizate de mediile sociale. Explorarea și proiectarea unor caracteristici relevante pe baza textelor online reprezintă un pas important în cadrul acestui proces.

Prin urmare, un obiectiv important al tezei este de a modela caracteristici interesante din datele Twitter, folosindu-ne și de dicționare predefinite de sentimente. În plus, fuziunea dintre aceste caracteristici este utilizată pentru detectarea sentimentului din tweet-uri.

Un alt obiectiv al cercetării este combinarea detectării polarității cu sisteme de recomandare. Îmbunătățirea unei recenzii despre un hotel sau un restaurant, prin atașarea sentimentului corespunzător, poate fi foarte utilă pentru a crea sugestii semnificative pentru utilizatori. În plus, este definită o nouă măsură de similaritate. Măsura *ARP (Atractivitate-Relevanță-Popularitate)* ia în considerare sentimentul determinat din recenzii.

Totodată, experimentele indică faptul că utilizarea caracteristicilor definite îmbunătățește sarcina de detectare a polarității pentru texte. De asemenea, rezultatele evidențiază faptul că explorarea polarității recenziilor pentru a face sugestii poate fi dificilă, dar foarte utilă. În concluzie, partea de modelare a datelor din procesul de analiză a sentimentelor, precum și aplicațiile aferente, aduc informații valoroase pentru a înțelege mai bine cantitatea imensă de date care ne înconjoară zilnic.

Lista Publicațiilor

- Petruşel, Mara-Renata and Limboi, Sergiu-George **A restaurants recommendation system: Improving rating predictions using sentiment analysis** In 2019 21st International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC) (pp. 190-197). IEEE. (Conferință categoria C-2 puncte)
- Limboi, Sergiu and Dioşan, Laura. **Hybrid Features for Twitter Sentiment Analysis.** In Artificial Intelligence and Soft Computing: 19th International Conference, ICAISC 2020, Zakopane, Poland, October 12-14, 2020, Proceedings, Part II 19 (pp. 210-219). Springer International Publishing. (Conferință categoria C-2 puncte)
- Deac-Petruşel, Mara and Limboi, Sergiu. **A sentiment-based similarity model for recommendation systems.** In 2020 22nd International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC) (pp. 224-230). IEEE. (Conferință categoria D-0 puncte)
- Limboi, Sergiu and Deac-Petruşel, Mara. **A Validation Framework for ARP Similarity Measure**. In 2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA) (pp. 1266-1271). IEEE.(Conferință categoria C-2 puncte)
- Limboi, Sergiu and Dioşan, Laura. **An unsupervised approach for Twitter Sentiment Analysis of USA 2020 Presidential Election.** In 2022 International Conference on INnovations in Intelligent SysTems and Applications (INISTA) (pp. 1-6). IEEE. (Conferință categoria C-2 puncte)
- Sergiu Limboi and Laura Dioşan. **The Twitter-Lex Sentiment Analysis System.** In Proceedings of the 14th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2022) - KDIR, pages 180-187. (Conferință categoria C-2 puncte)
- Limboi, Sergiu and Dioşan, Laura. **A Lexicon-based Feature for Twitter Sentiment Analysis.** In 2022 IEEE 18th International Conference on Intelligent Computer Communication and Processing (ICCP) (pp. 95-102). IEEE. (Conferință categoria D-1 punct)
- Sergiu Limboi. **Comparison of Data Models For Unsupervised Twitter Sentiment Analysis**. Studia Universitatis Babeş-Bolyai Informatica, 67(2), p. 65-80, May 2023. (Jurnal categoria D-1 punct)

Total puncte: 12

Cuprins

1 Introducere **1**
1.1 Motivație 1
1.2 Contribuții originale 2

2 Analiza Sentimentelor pentru Informații Textuale **4**

3 Atribute și Modele pentru Texte de Mici Dimensiuni **5**

4 Reprezentări de Date pentru Analiza Sentimentelor Nesupervizată **7**

5 Aplicații ale Analizei Sentimentelor **9**

6 Concluzii și Dezvoltări ulterioare **11**

Capitolul 1

Introducere

1.1 Motivație

În zilele noastre, supraîncărcarea cu informații [15] este o problemă reală pentru oameni, deoarece aceștia se confruntă cu o mulțime de știri, reclame și oferte. Dacă o persoană dorește să cumpere o mașină, poate alege dintre diverse modele cu multe opțiuni sau îmbunătățiri. Astfel, cantitatea mare de informații copleșește un potențial client, iar acesta tinde să amâne deciziile pentru a analiza mai în detaliu posibilitățile. De asemenea, din punct de vedere psihologic, cu cât oferta este mai mare, cu atât decizia este mai complexă.

Online-ul a câștigat teren în ultimii ani și a devenit o adevărată putere de influențare a oamenilor, înlocuind canalele tradiționale de comunicare precum radioul, ziarele sau televiziunea. Mediile sociale le permit utilizatorilor să își exprime opiniile despre politică, sănătate, vreme sau celebrități. În plus, utilizatorii pot avea o rețea socială, pot împărtăși opinii și pot urmări diferiți utilizatori. În comparație cu modul tradițional de scriere, textele de pe rețelele sociale sunt reprezentate de stilul liber de scriere, de utilizarea abrevierilor sau de limbajul colocvial.

Domeniul de analiză a sentimentelor [11] ar putea fi un instrument puternic de analiză a informațiilor bazate pe text. Unul dintre obiectivele sale principale este determinarea sentimentului sau a polarității din diferite surse de informații. În plus, sentimentul poate fi pozitiv, neutru sau negativ. Etichetarea unui text postat pe rețelele de socializare poate indica în mod corespunzător opiniile utilizatorilor cu privire la un eveniment. Înainte de partea de detectare a sentimentului, preprocesarea datelor este o provocare din cauza modului de scriere pe platformele sociale. De asemenea, este esențială modelarea datelor pentru a reprezenta o intrare adecvată pentru un algoritm de clasificare. În final, toate aceste acțiuni construiesc o sarcină de clasificare care determină dacă un mesaj are un anumit sentiment.

Analiza sentimentelor are multe aplicații în diferite domenii. Unul dintre acestea este procesul de realizare a recomandărilor pentru utilizatori. Dacă un utilizator dorește să facă cumpărături online, recenziile sunt esențiale pentru deciziile sale. Un produs cu multe recenzii proaste nu va fi atractiv pentru un potențial client, dar unul cu comentarii pozitive poate fi ales. Analiza recenziilor online ale altor utilizatori este o parte interesantă a analizei sentimentelor. Pe baza dificultăților de interpretare a unui text (negații, sarcasm sau scriere colocvială), o recenzie îmbunătățită a sentimentului poate fi foarte constructivă. Astfel, întocmirea unei liste de sugestii cu comentarii pozitive poate îmbunătăți activitatea sistemului de prezentare a recomandărilor către utilizatori.

Cu alte cuvinte, este necesară descrierea datelor postate pe rețelele de socializare pentru a defini

tendințele și opiniile actuale și pentru a prezice evenimentele viitoare. Prin urmare, teza prezentată va aplica tehnicile de analiză a sentimentelor pentru informații textuale (recenzii și tweet-uri) pentru a îmbunătăți calitatea detectării sentimentelor.

1.2 Contribuții originale

Contribuțiile originale ale acestei teze sunt evidențiate prin descrierea pas cu pas a abordărilor definite.

Prima direcție proiectată [7] se concentrează pe nivelul de extragere a caracteristicilor pentru îmbunătățirea clasificării sentimentelor pentru tweet-uri. Scopul acestor atribute este de a dezvălui informații ascunse din tweet-ul original. Având în vedere aceste aspecte, propunem combinarea caracteristicilor extrase din ambele părți ale unui mesaj Twitter (text și hashtag-uri) pentru a testa dacă fuziunea lor îmbunătățește procesul de clasificare. Astfel, sunt definite patru caracteristici: Baseline-based Sentiment Analysis (BSA), Hashtag-based Sentiment Analysis (HSA), Fused-based Sentiment Analysis (FSA) și Raw-based Sentiment Analysis (RSA).

Următoarea abordare [8] definește o caracteristică bazată pe lexicon pentru a verifica dacă polaritatea unui tweet poate fi consolidată cu un indiciu. Prin urmare, un anumit tweet este îmbunătățit cu un **indicator de sentiment**. Acest indicator de sentiment este determinat pe baza scorului de sentiment calculat de un lexicon. În cadrul experimentelor sunt utilizate patru lexicoane de sentiment: Text Blob [17], Vader [4], Senti WordNet [3] și aFINN [12]. Rezultatele arată valori similare în ceea ce privește lexicoanele utilizate, astfel încât, pentru experimentele viitoare, este suficient să se utilizeze doar unul dintre lexicoanele menționate anterior.

Următorul pas firesc este de a verifica dacă combinarea caracteristicilor Twitter și a indicatorului de sentiment poate îmbunătăți și mai mult problema clasificării sentimentului pentru informațiile despre tweet. Un sistem numit **Twitter-Lex Sentiment Analysis** [9] îmbină punctul de vedere bazat pe lexicon cu caracteristicile specifice Twitter. Astfel, sunt definite patru caracteristici: Baseline Sentiment Analysis-Lexicon (BSA_{lex}), Hashtag Sentiment Analysis-Lexicon (HSA_{lex}), Fused Sentiment Analysis-Lexicon (FSA_{lex}) și Raw Sentiment Analysis-Lexicon (RSA_{lex}). Fiecare conține caracteristica Twitter originală (BSA, HSA, FSA și RSA) și indicatorul de sentiment determinat pe baza lexiconului Vader.

În cazul abordărilor ulterioare, accentul se mută pe contextul nesupervizat pentru detectarea sentimentului în tweet-urile colectate. În primul rând, sunt luate în considerare pentru analiză tweet-urile prezidențiale de la alegerile din 2020 din Statele Unite ale Americii [10]. Obiectivul principal este de a construi clustere corespunzătoare pentru sentimente (pozitive sau negative) și de a verifica dacă mai multe reprezentări sau modele cresc calitatea grupării. Un nou model numit **hash index** este definit luând în considerare hashtag-urile extrase dintr-un tweet. Pe parcursul procesului sunt utilizate și alte reprezentări: TF-IDF și $TF - IDF_{hash}$.

Perspectiva din [5] introduce două reprezentări de date definite pe baza caracteristicilor text și hashtag extrase din tweet-uri. Reprezentările menționate sunt aplicate în context nesupervizat pentru detectarea a două grupuri de mesaje: unul cu tweet-uri pozitive și unul cu mesaje negative. Aceste modele corespund caracteristicilor **BSA** și **HSA** din contextul supervizat. Contribuția este dată de compararea celor două reprezentări pentru a determina care se potrivește cel mai bine în scenariul nesupervizat. Rezultatele arată că reprezentarea bazată pe hashtag-uri este mai bună decât cea bazată pe text, iar tehnicile de clusterizare nu influențează întregul proces.

Ultima parte a contribuțiilor evidențiază o lucrare de colaborare care prezintă utilizarea analizei sentimentelor pentru recenzii în contextul sistemelor de recomandare [14]. Lucrarea noastră implică clasificatori de învățare automată care sunt utilizați pentru etichetarea recenziilor colectate. Un set de date îmbunătățit cu polaritatea rezultată este transmis unei tehnici de filtrare colaborativă, iar în faza de recomandare sunt luate în considerare doar recenziile pozitive.

În continuare, măsura de similaritate **ARP (Atractivitate-Relevanță-Popularitate)** [2] este definită luând în considerare un modul numit **Sentiment Scoring**. Prin urmare, fiecărei recenzii îi este asociat un scor de sentiment calculat luând în considerare scorul determinat pe baza lexiconului Senti WordNet. Experimentele arată că sistemul de evaluare a sentimentului produce rezultate foarte bune și poate fi utilizat pentru procesul de recomandare. Întreaga propunere pentru această similitudine este prezentată în [2], descriind și partea de recomandare. Subiectul acestei teze este doar contribuția noastră, care este modulul de clasificare a sentimentelor.

O altă etapă interesantă este cadrul de validare definit pentru similitudinea **ARP** [6] pentru a verifica dacă poate fi utilizată la fel ca cele clasice (de exemplu, coeficientul de corelație Pearson, distanța euclidiană, metrica Jaccard etc.). Validarea este realizată pe baza a cinci criterii, dar în această teză se explică criteriul de robustețe la zgomot, deoarece aceasta este contribuția noastră. Această condiție verifică modul în care se comportă sistemul cu date ce conțin zgomot. În cele din urmă, pe baza rezultatelor menționate în articolul anterior, toate aceste cerințe validează similitudinea.

Capitolul 2

Analiza Sentimentelor pentru Informații Textuale

Analiza sentimentelor poate fi modelată ca o problemă de clasificare care implică clasificarea subiectivității (clasificarea opiniilor în subiective și obiective), clasificarea polarității (clasificarea expresiilor în negative, pozitive și neutre) sau detectarea spam-ului de opinie.

Analiza sentimentelor, vizualizată ca o sarcină de clasificare a polarității, implică trei etape principale: cea de inițializare, cea de învățare și cea de evaluare. Etapa de inițializare pregătește datele pentru algoritmul de clasificare. Colectarea datelor înseamnă alegerea datelor și analizarea conținutului acestora. Este necesară o adnotare manuală pentru a verifica dacă datele mai trebuie etichetate. Faza de preprocesare înseamnă transformarea informațiilor nestructurate într-una clară, fără greșeli de ortografie, abrevieri sau cuvinte de argou. Apoi, o etapă de extragere a caracteristicilor determină modul în care sunt modelate datele în ceea ce privește caracteristicile relevante. Etapa de reprezentare a datelor utilizează o tehnică de încorporare a cuvintelor pentru a converti modelele în reprezentări numerice. Apoi, etapa de învățare descrie momentul în care un model antrenat este transmis unui algoritm de învățare automată. Ultima etapă este evaluarea, când se calculează măsurile de performanță pentru a reflecta cât de bună este metodologia.

Twitter (rebranduit la **X** începând cu iulie 2023) este o platformă populară de socializare pentru comunicarea cu alte persoane, exprimarea sentimentelor și opiniilor și difuzarea de știri. Avantajele unui instrument atât de puternic sunt disponibilitatea diferitelor dispozitive electronice, posibilitatea de a avea un număr mare de prieteni și faptul că puteți trimite mesaje mici și concise altor prieteni pe diferite subiecte [16].

Principalele concepte utilizate pe platforma Twitter sunt URL, mențiune, utilizator, prieten, urmăritor, tweet, recență, hashtag, emoticon, re-tweet și singleton [13]. Un tweet este un mesaj text simplu de maximum 280 de caractere și poate conține hashtag-uri și elemente bazate pe meta-text: medii conexe (hărți, fotografii, videoclipuri) și site-uri web. Hashtag-urile [16] sunt cuvinte-cheie prefixate cu simbolul „#” care pot apărea într-un tweet. Utilizatorii Twitter folosesc această notație pentru a-și categorisi mesajele sau pentru a le evidenția mai usor la o căutare ulterioară.

Capitolul 3

Atribute și Modele pentru Texte de Mici Dimensiuni

Chiar dacă un tweet are text, hashtag-uri și elemente de meta-text, doar primele două sunt supuse analizei, deoarece meta-textul oferă informații diverse din punct de vedere tipologic (link-uri, videoclipuri etc.) și semantic (semnificație). În plus, conceptele multimedia sunt prezente doar uneori în mesaje, astfel încât setul de date ar fi diminuat considerabil pentru partea de analiză. Patru perspective sunt prezentate și aplicate în cadrul procesului de detectare a polarității tweeturilor.

Atributul **Baseline-Based Sentiment Analysis (BSA)** presupune că datele de intrare sunt constituite din informații textuale fără cuvintele-cheie (hashtaguri) care definesc un mesaj Twitter. Atributul **Hashtag-Based Sentiment Analysis (HSA)** este construit prin extragerea hashtag-urilor dintr-un tweet. Acesta păstrează o listă de hashtag-uri (indicatori) pentru fiecare mesaj. Abordarea **Fused-Based Sentiment Analysis (FSA)** combină cele anterioare. Intrarea pentru un algoritm de clasificare va fi reprezentată de text (fără hashtag-uri) concatenat cu lista de hashtag-uri. Caracteristica **Raw-Based Sentiment Analysis (RSA)** preia datele de intrare ca un text brut în care semnul # pentru hashtag-uri este eliminat. Dacă semnul # este eliminat, atunci cuvântul devine unul obișnuit și va fi procesat ca și celelalte în etapa de preprocesare.

Un alt pas esențial în experimentele noastre legate de caracteristici este proiectarea unei caracteristici bazate pe lexicon pentru o problemă de clasificare a polarității tweet-urilor, luând în considerare mesajele în limba engleză. În încercările anterioare, am detectat tweet-uri pozitive și negative, dar acum am inclus și sentimentul neutru. În plus, un simplu mesaj tweet poate exprima o anumită polaritate, dar numai câteva cuvinte rămân relevante din cauza preprocesărilor și a numărului mic de cuvinte. Chiar dacă hashtag-urile indică sentimentul mesajului, am decis să îmbunătățim mesajul prin oferirea unui indiciu. Acest indiciu este o tehnică bazată pe reguli care determină un scor pentru sentiment pe baza unui lexicon. În ceea ce privește acest scor, putem extrage un sentiment (de exemplu, pozitiv, negativ sau neutru). Cu alte cuvinte, pornim de la datele de intrare originale, le transmitem unui lexicon și îmbunătățim textul cu un indicator, numit **indicator de sentiment**. În final, informațiile actualizate sunt transmise unui clasificator care poate decide polaritatea reală a tweet-ului.

În continuare, definim un sistem care combină caracteristicile hibride cu cele bazate pe lexicon. Sistemul conceput, **Twitter-Lex SA**, are ca scop explorarea informațiilor oferite de platforma Twitter în combinație cu utilizarea unui lexicon bazat pe sentiment. Analizarea unui singur set de caracteris-

tici (de exemplu, cele lexicale) este insuficientă pentru o clasificare bună. În cele mai multe cazuri, informația textuală nu este suficientă atunci când se discută despre platforma Twitter, deoarece caracteristicile specifice pot evidenția mai bine mesajul (de exemplu, hashtag-uri, mențiuni). În plus, contextul poate fi esențial, iar combinarea cuvintelor din cadrul propoziției și a altor caracteristici poate schimba polaritatea generală a textului. Ținând cont de toate aceste lucruri și pornind de la caracteristicile anterioare (hibride și bazate pe lexicon), sunt definite patru caracteristici:

- Baseline Sentiment Analysis-Lexicon (BSA_{lex})
- Hashtag Sentiment Analysis-Lexicon (HSA_{lex})
- Fused Sentiment Analysis-Lexicon (FSA_{lex})
- Raw Sentiment Analysis-Lexicon (RSA_{lex})

Sinteză a experimentelor

O prezentare generală a abordărilor definite este prezentată în **Tabelul 3.1**, luând în considerare seturile de date utilizate, clasificatorii de învățare automată, măsura de evaluare și utilizarea unui lexicon pentru sentimente.

Tabela 3.1: Experimente ale Analizei Sentimentelor Aplicate pe Twitter.

Abordare	**Set de date**	**Clasificator**	**Lexicon**	**Măsură de evaluare**
Caracteristici hibride (BSA, HSA, FSA, RSA)	Sanders	SVM, LR și NB	Neutilizat	Acuratețe, precizie
Caracteristica bazată pe lexicon	Apple Twitter Sentiment, Sanders, Twitter US Airline	SVM, LR și NB	Text Blob, aFINN, Vader și Senti WordNet	Acuratețe, precizie
Caracteristicile Twitter-Lex	Apple Twitter Sentiment, Sanders, Twitter US Airline și Twitter Climate Change	SVM, LR și NB	Vader	Acuratețe, precizie

Capitolul 4

Reprezentări de Date pentru Analiza Sentimentelor Nesupervizată

Alegerile prezidențiale din 2020 din Statele Unite au fost urmărite în întreaga lume datorită polarităților mari dintre fanii candidaților finaliști, Joe Biden și Donald Trump. Social media a fost mediul cel mai utilizat pentru a-și exprima sentimentele și opiniile despre favoritul lor sau pentru a-l jigni și ataca pe candidatul opus. Platforma Twitter a încapsulat diferitele idei sau gânduri care au construit o imagine de ansamblu a întregii perioade de campanie și a dezbaterilor prezidențiale.

Pe baza acestor aspecte, analiza sentimentelor poate fi aplicată pentru a explora lumea Twitter pentru a defini tendințele și direcțiile privind alegerile din 2020, pre și post electorale.

În cadrul abordării prezentate, se proiectează și implementează un sistem care oferă o perspectivă nesupervizată pentru recunoașterea sentimentului dominant în tweet-urile postate despre ultimii doi candidați. Scopul este de a defini două grupuri de mesaje (pozitive și negative) care ne pot arăta opinia utilizatorilor despre Donald Trump și Joe Biden.

Prin urmare, rezumatul lucrării este reprezentat de următoarele etape:

- se definește un nou model de tweet numit **hash index**, o reprezentare care ia în considerare hashtag-urile unui mesaj;
- analizăm impactul reprezentării datelor pentru întregul proces. Aplicăm modelul TF-IDF pe o reprezentare hashtag pentru un tweet, mai precis pe lista extrasă de hashtag-uri;
- utilizăm algoritmi de clusterizare pe seturi de date neetichetate pentru a detecta grupurile relevante de sentimente pentru fiecare candidat la președinție;
- extindem procesul de validare a întregii abordări prin adăugarea etapei de validare externă prin utilizarea unui proces automat de etichetare bazat pe lexiconul Vader.

Experimentele ulterioare verifică dacă modelele sau reprezentările propuse sunt adecvate pentru a identifica sentimentul tweet-urilor în context nesupervizat. Aplicând diferiți algoritmi de grupare, dorim să definim două grupuri de mesaje (unul pozitiv și unul negativ) prin utilizarea a două modele noi. Ținând cont de faptul că, în lumea Twitter, hashtag-urile reprezintă o caracteristică esențială, deoarece sunt indicatori ai mesajului, definim o reprezentare bazată pe hashtaguri. Pe de altă parte, o reprezentare bazată pe text este construită pe baza ideii că textul (fără hashtag-uri) constituie o

informație relevantă pentru problema detectării sentimentului. În plus, în cadrul experimentelor numerice, determinăm care model este mai bun pentru clasificarea sentimentului în context nesupervizat.

Sinteză a experimentelor

O imagine de ansamblu a abordărilor descrise anterior este prezentată în **Tabelul 4.1**, luând în considerare seturile de date utilizate, algoritmii de clusterizare și măsura de evaluare.

Tabela 4.1: Sinteză a Analizei Sentimentelor Nesupervizată

Abordare	**Set de date**	**Algoritm de clusterizare**	**Masură de evaluare internă**	**Masură de evaluare externă**
Codificare bazată pe hash index	Tweeturi de la alegerile din 2020 din SUA	K-Means, Aglomerativ	Silhouette, Davies-Bouldin	Precizie
Reprezentări bazate pe text și pe hashtag-uri	Joe Biden, COVID-19 și setul de date privind dezbaterile prezidențiale republicane	K-Means, Aglomerativ, Spectral	Silhouette, Davies-Bouldin	Acuratețe

Capitolul 5

Aplicații ale Analizei Sentimentelor

Acest capitol se ocupă de mesajele de dimensiuni mai lungi, recenziile, și explică modul în care analiza sentimentelor poate fi combinată cu sistemele de recomandare pentru a îmbunătăți calitatea unui proces de recomandare. Scopul principal este de a combina procesul de analiză a sentimentelor cu o tehnică de filtrare colaborativă pentru a analiza dacă luarea în considerare a sentimentului unei recenzii poate crește calitatea recomandărilor. *Yelp Restaurant Reviews* [1] este utilizat ca informație pentru sistem. Apoi, datele sunt transmise către modulul de analiză a sentimentului, care are ca scop clasificarea recenziilor colectate în comentarii pozitive și negative. Aceste date etichetate vor reprezenta datele de intrare pentru modulul de recomandare care se ocupă de partea de filtrare colaborativă care va genera, în final, o listă de recomandări pentru utilizatori.

Următoarea etapă interesantă în procesul nostru de combinare a analizei de sentimente cu sistemele de recomandare pentru recenzii implică definirea unei abordări bazată pe scorul polarității. Aceste scoruri sunt utilizate pentru o similaritate nou definită, numită **ARP (Atractivitate-Relevanță-Popularitate)** care este integrată în tehnica de filtrare colaborativă **k-Nearest-Neighbors (kNN)** [1] pentru procesul de recomandare. După colectarea datelor (în acest caz, recenzii), textele sunt transmise către modulul **Sentiment Scoring**. Această componentă calculează valoarea sentimentului pentru informațiile textuale furnizate. Datele colectate sunt preprocesate și transmise către o funcție de calculare a scorului. Această funcție utilizează instrumentul *Senti WordNet* pentru a calcula sentimentul fiecărui cuvânt. Scorul sentimentului este apoi transformat într-un rating de la unu la cinci.

Necesitatea unei măsuri de similaritate este foarte importantă în mai multe contexte: clustering, sisteme de recomandare, clasificări etc. În literatura de specialitate există o mulțime de măsuri cum ar fi distanța euclideană, similaritate cosinus, Pearson sau Jaccard. Deoarece măsura de similaritate **ARP (Atractivitate-Relevanță-Popularitate)** este una nouă, trebuie să definim câteva etape de validare care să dovedească faptul că această măsură poate fi inclusă în setul de măsuri deja cunoscute. Dacă setul de date conține informații textuale (de exemplu, descrieri de produse, recenzii bazate pe text), putem utiliza măsura ARP. În caz contrar, măsurile clasice pot fi aplicate într-un context specific (de exemplu, clustering, sisteme de recomandare etc.). Pentru validarea măsurii ARP se vor utiliza următoarele faze: verificarea îndeplinirii condițiilor de a fi metrică, utilitatea, expresivitatea, corelația cu alte măsuri și condiția de robustețe a zgomotului. Cu alte cuvinte, prima etapă este de a verifica dacă ARP este deja o măsură sau poate fi generată ca măsură. Condiția de utilitate verifică

[1] https://www.yelp.com/dataset

dacă măsura poate fi aplicată la diferite seturi de date. Etapa de expresivitate determină dacă ARP este mai potrivită pentru colecțiile bazate pe text decât alte măsuri. Următoarea etapă calculează corelația dintre ARP și celelalte măsuri. Ultima etapă presupune verificarea modului în care ARP se comportă în cazul datelor cu zgomot (zgomotul poate proveni de la clasificatori/lexicoane de sentiment sau tehnici de preprocesare).

Sinteză a experimentelor

O prezentare generală a abordărilor prezentate este reflectată în **Tabelul 5.1**, luând în considerare setul de date, clasificatorii de sentiment și măsurile de evaluare.

Tabela 5.1: Sinteză a aplicării analizei sentimentelor în diverse domenii.

Abordare	**Seturi de date**	**Clasificator**	**Măsură de evaluare**
Analiza sentimentelor pentru determinarea polarităților recenziilor	Recenzii de restaurante Yelp	SVM, LR, NB	Acuratețe, precizie, rapel, măsura f
Modul de evaluare a sentimentelor pentru măsura de similaritate ARP	Yelp Restaurants Reviews, Datafiniti Hotel Reviews	Senti WordNet lexicon	MAE, RMSE
Validarea măsurii ARP	Yelp Restaurants Reviews, Datafiniti Hotel Reviews	aFINN, Text Blob, Vader, Senti WordNet	MAE
Aylien API	Sanders	învățare profundă	acuratețe, precizie, rapel, măsura f

Capitolul 6

Concluzii și Dezvoltări ulterioare

Analiza sentimentelor este un domeniu provocator care poate fi aplicat pe o gamă largă de date de intrare. Această teză se concentrează pe clasificarea polarității pentru informații textuale. În zilele noastre, oamenii își exprimă ideile cu privire la un eveniment, un politician, un film etc., pe rețelele de socializare. Inconvenientul unor astfel de mesaje este reprezentat de modul în care ele sunt scrise (folosirea argoului, utilizarea de expresii colocviale etc.). Prin urmare, provocarea principală constă în preprocesarea și curățarea datelor care vor fi transmise unui sistem care se ocupă de detectarea sentimentului pentru astfel de mesaje. De asemenea, extragerea caracteristicilor și modelarea datelor reprezintă etape critice datorită diversității atributelor care pot fi extrase din texte. Chiar dacă folosim aceiași clasificatori de învățare automată, caracteristici și modele diferite pot produce rezultate diferite.

Prin urmare, primul obiectiv al experimentelor noastre se concentrează pe aplicarea tehnicilor de analiză a sentimentelor la mesajele Twitter, numite tweets. Activitatea principală este reprezentată de construirea de noi caracteristici și reprezentări pentru a crește calitatea clasificării. Astfel, sunt proiectate patru caracteristici bazate pe elemente specifice platformei Twitter.

Următorul pas al abordării noastre este definirea unei caracteristici bazate pe lexicon pentru detectarea sentimentului din tweet-uri. Prin urmare, mai multe lexicoane de sentiment (aFINN, Vader, Senti WordNet și Text Blob) sunt utilizate pe parcursul procesului. Tweetul îmbunătățit este tweetul preprocesat combinat cu un **indicator de sentiment**, care reprezintă sentimentul determinat pe baza scorului de sentiment calculat de către un lexicon. Ultima fază a experimentelor noastre constă în fuzionarea caracteristicilor **BSA, HSA, RSA** și **FSA** cu cea bazată pe lexicon. Această fuziune este reprezentată de sistemul **Twitter-Lex Sentiment Analysis**. Ideea principală este că fiecare caracteristică Twitter (BSA, FSA, HSA și RSA) este concatenată cu indicatorul de sentiment al lexiconului Vader.

În cadrul următoarelor experimente se realizează trecerea de la învățarea supervizată la cea nesupervizată. Astfel, mai mulți algoritmi de grupare a datelor sunt aplicați pe tweet-urile prezidențiale din SUA 2020 pentru a determina grupuri de mesaje care vor indica sentimentul corespunzător. În plus, se definește un nou model de încorporare a cuvintelor, numit **hash index**, o reprezentare care ia în considerare hashtag-urile dintr-un tweet. Întregul proces de grupare este evaluat cu ajutorul unor măsuri interne și externe. În continuare, se realizează o comparație detaliată între cele două reprezentări pentru învățarea nesupervizată.

Al doilea scop important al tezei de doctorat este de a evidenția aplicațiile analizei de sentimente în alte domenii, în special în sistemele de recomandare. În consecință, procesul de analiză a sentimentelor

este aplicat la texte de dimensiuni mai mari decât tweet-urile, recenziile postate de utilizatori. Rating-ul din sistemele de recomandare (de obicei, un rating de la unu la cinci acordat de utilizator pentru un anumit articol/produs) va fi înlocuit cu sentimentul derivat din aplicarea clasificatorilor de învățare automată. Apoi se definește o nouă măsură de similaritate numită **ARP (Atractivitate-Relevanță-Popularitate)**. În final, măsura ARP este validată pe baza a cinci condiții.

Pentru viitor avem în vedere extinderea experimentelor existente prin aplicarea caracteristicilor și reprezentărilor proiectate pe mai multe seturi de date, ținând cont de mărimea și zgomotul acestora. În plus, este necesară o comparație detaliată cu abordări existente în literatura de specialitate (de exemplu, modelele BERT) și combinarea caracteristicilor cu acestea. Pe de altă parte, dorim să explorăm un clasificator bazat pe o rețea de tip **transformer** pentru determinarea polarității unui tweet. De asemenea, caracteristicile oferite de Twitter nu sunt suficient exploatate astfel încât putem lua în considerare mai multe aspecte, cum ar fi retweeturile, mențiunile sau răspunsurile. În ceea ce privește tweet-urile prezidențiale, dorim să generalizăm procesul pentru a utiliza metodologia pentru viitoarele alegeri și să realizăm o analiză bazată pe un anumit interval de timp (de exemplu, impactul alegerilor din SUA pe parcursul a zece ani) prin definirea unor modele matematice pentru întregul sistem. Un cadru general pentru aceste caracteristici poate fi conceput și implementat pentru a determina tendințele generale și pentru a îmbunătăți calitatea sistemelor de recomandare, a tehnicilor de grupare sau a clasificatorilor de învățare automată. În cele din urmă, domeniul analizei sentimentelor pentru informații textuale încă reprezintă o provocare, cu multe oportunități de explorat.

Bibliografie

[1] R Baeza-Yates. Modern information retrieval. *Addison Wesley google schola*, 2:127–136, 1999.

[2] Mara Deac-Petrusel and Sergiu Limboi. A sentiment-based similarity model for recommendation systems. In *2020 22nd International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*, pages 224–230. IEEE, 2020.

[3] Hussam Hamdan, Frederic Béchet, and Patrice Bellot. Experiments with dbpedia, wordnet and sentiwordnet as resources for sentiment analysis in micro-blogging. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 455–459, 2013.

[4] Clayton Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225, 2014.

[5] Sergiu Limboi. Comparison of data models for unsupervised twitter sentiment analysis. *Studia Universitatis Babeş-Bolyai Informatica*, 67(2):65–80, 2023.

[6] Sergiu Limboi and Mara Deac-Petrusel. A validation framework for arp similarity measure. In *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1266–1271. IEEE, 2021.

[7] Sergiu Limboi and Laura Dioşan. Hybrid features for twitter sentiment analysis. In *Artificial Intelligence and Soft Computing: 19th International Conference, ICAISC 2020, Zakopane, Poland, October 12-14, 2020, Proceedings, Part II 19*, pages 210–219. Springer, 2020.

[8] Sergiu Limboi and Laura Dioşan. A lexicon-based feature for twitter sentiment analysis. In *2022 IEEE 18th International Conference on Intelligent Computer Communication and Processing (ICCP)*, pages 95–102. IEEE, 2022.

[9] Sergiu Limboi and Laura Dioşan. The twitter-lex sentiment analysis system. In *Proceedings of the 14th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, IC3K 2022, Volume 1: KDIR, Valletta, Malta, October 24-26, 2022*, pages 180–187. SCITEPRESS, 2022.

[10] Sergiu Limboi and Laura Dioşan. An unsupervised approach for twitter sentiment analysis of usa 2020 presidential election. In *2022 International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*, pages 1–6. IEEE, 2022.

[11] Walaa Medhat, Ahmed Hassan, and Hoda Korashy. Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 5(4):1093–1113, 2014.

[12] Finn Årup Nielsen. A new anew: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903*, 2011.

[13] Kishori K Pawar, Pukhraj P Shrishrimal, and RR Deshmukh. Twitter sentiment analysis: A review. *International Journal of Scientific & Engineering Research*, 6(4):9, 2015.

[14] Mara-Renata Petrusel and Sergiu-George Limboi. A restaurants recommendation system: Improving rating predictions using sentiment analysis. In *2019 21st International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*, pages 190–197. IEEE, 2019.

[15] Gloria Phillips-Wren and Monica Adya. Decision making under stress: The role of information overload, time pressure, complexity, and uncertainty. *Journal of Decision Systems*, 29(sup1):213–225, 2020.

[16] Jagan Sankaranarayanan and et al. Twitterstand: News in tweets. In *Proceedings of the 17th ACM SIGSPATIAL*, GIS '09, pages 42–51. ACM, 2009.

[17] Bhupender Singh Shekhawat. *Sentiment classification of current public opinion on brexit: Naïve Bayes classifier model vs Python's Textblob approach.* Phd thesis, Dublin, National College of Ireland, 2019.