UNIVERSITATEA BABEŞ-BOLYAI
FACULTATEA DE MATEMATICĂ ȘI INFORMATICĂ

Dezvoltarea de resurse și abordări de învățare automată pentru task-uri de procesare a limbajului natural

Rezumatul tezei de doctorat

Student-doctorand: Anamaria Briciu
Conducător științific: Prof. dr. Czibula Gabriela

2022

Cuvinte cheie: procesarea limbajului natural, detecția emoțiilor în text, identificarea autorului unui text, învățare automată, învățare profundă, autoencoderi

Cuprins

Lista publicațiilor

Clasamentul publicațiilor a fost realizat conform standardelor CNATDCU (Consiliul Naţional de Atestare a Titlurilor, Diplomelor si Certificatelor Universitare) aplicabile pentru studenții doctoranzi înscriși după 1 octombrie 2018. Toate clasamentele sunt listate conform clasificării jurnalelor [1] și a conferințelor [2] în Informatică.

Publicații indexate în Web of Science - Science Citation Index Expanded

[CLB22] Gabriela Czibula, Mihaiela Lupea and **Anamaria Briciu**. *Enhancing the performance of software authorship attribution using an ensemble of deep autoencoders*. Mathematics, Special issue on Recent Advances in Artificial Intelligence and Machine Learning, 2022, under review (**2021 IF=2.592**, Journal IF Quartile Q1)

Rank A, 0 points.

[LBB21] Mihaiela Lupea, **Anamaria Briciu** and Elena Bostenaru. *Emotion-based Hierarchical Clustering of Romanian Poetry*. Studies in Informatics and Control, v. 30, n. 1, pp. 109–118, 2021 (**2021 IF=1.826**, Journal IF Quartile Q3)

Rank C, 2 points.

[LB19] Mihaiela Lupea and **Anamaria Briciu**. *Studying emotions in Romanian words using formal concept analysis*. Computer Speech & Language, v. 57, pp. 128–145, 2019 (**2019 AIS=0.117**, Journal AIS Quartile Q2)

Rank B, 4 points.

Publicații indexate în Web of Science, Conference Proceedings Citation Index

[LBCC22] Mihaiela Lupea, **Anamaria Briciu**, Istvan Gergely Czibula and Gabriela Czibula. *SoftId: An autoencoder-based one-class classification model for software authorship identification*. 26th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2022), accepted.

Rank B - CORE2021, 2 points.

[1] https://uefiscdi.ro/premierea-rezultatelor-cercetarii-articole
[2] http://portal.core.edu.au/conf-ranks/

[BCL21] **Anamaria Briciu**, Gabriela Czibula and Mihaiela Lupea. *AutoAt: A deep autoencoder-based classification model for supervised authorship attribution*. 25th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2021), September 8-10, 2021, Procedia Computer Science 192, pp. 397-406.

Rank B - CORE2021, 4 points.

[LB17] Mihaiela Lupea and **Anamaria Briciu**. *Formal Concept Analysis of a Romanian emotion lexicon*. 2017 13th IEEE International Conference on Intelligent Computer Communication and Processing (ICCP), pp. 111–118, 2017.

Rank C - CORE2017, 2 points.

[BL18] **Anamaria Briciu** and Mihaiela Lupea. *Studying the language of mental illness in Romanian social media*. 2018 IEEE 14th International Conference on Intelligent Computer Communication and Processing (ICCP), pp. 21–28, 2018 .

Rank C - CORE2018, 2 points.

Publicații în jurnale și volume ale conferințelor

[BL17] **Anamaria Briciu** and Mihaiela Lupea. *RoEmoLex - A Romanian Emotion Lexicon*. Studia Universitatis Babeș-Bolyai Informatica, v. 62, n. 2, pp. 45-56, 2017 (**indexed Mathematical Reviews**).

Rank D, 1 point.

[Bri19] **Anamaria Briciu**. *Quantitative Analysis of Style in Mihai Eminescu's Poetry*. Studia Universitatis Babeș-Bolyai Informatica, v. 64, n. 2, pp. 80–95, 2019 (**indexed Mathematical Reviews**).

Rank D, 1 point.

Scorul publicațiilor: 18 points.

Introducere

Domeniul principal de cercetare al prezentei teze de doctorat se referă la dezvoltarea resurselor și aplicațiilor pentru task-uri din domeniul procesării limbajului natural. Teza noastră de doctorat se intitulează „Dezvoltarea resurselor și abordări de învățare automată pentru task-uri de procesare a limbajului natural" și își propune să dezvolte resurse și modele de învățare automată și de învățare profundă pentru o serie de task-uri de prelucrare a limbajului.

Termenul de Natural Language Processing (NLP) se referă la domeniul de cercetare preocupat de explorarea modului în care calculatoarele pot fi folosite pentru a înțelege și manipula limbajul natural în contexte de comunicare diverse. Având la bază lingvistica tradițională, domeniul NLP a evoluat odată cu progresele în domeniul informaticii care au permis efectuarea unui număr tot mai mare de operații în mod automat, și creșterea complexității acestora. De asemenea, disponibilitatea crescândă a unor volume mari de text în limbaj natural stocate electronic a fost un factor important în popularizarea domeniului. Lingvistica clasică a urmărit elaborarea regulilor limbajului prin intermediul formalizării matematice. Aceasta a fost, de asemenea, abordarea primelor sisteme NLP, care au folosit reguli definite manual pentru a procesa limbajul natural. Cu toate acestea, astfel de abordări au limitări semnificative în a capta sensul unui cuvânt și în abordarea varietății în exprimare. Astfel, o schimbare fundamentală caracterizată prin aproximări simple și robuste în loc de analize profunde, un accent pe modele probabilistice de limbaj și crearea de corpusuri adnotate de care să fie utilizate împreună cu modele de învățare automată bazate pe statistici [Kle] a avut ca rezultat apariția subdomeniului de *procesare statistică a limbajului natural*, care definește domeniul actual NLP într-o mare măsură.

Importanța acestui domeniu se datorează numeroaselor aplicații pe care le are într-o varietate de contexte. În general, metodele NLP au fost concepute pentru indexarea și căutarea informațiilor în texte foarte lungi (*Information Retrieval (IR)* și *Information Extraction (IE)*), clasificarea textului în categorii, traducerea automată, sumarizarea automată, răspunsul automat la întrebări date, achiziția de cunoștințe și generarea de text [Cho03]. Mai concret, relevanța unor astfel de tehnici constă în automatizarea anumitor sarcini care permit oamenilor mai mult timp pentru a se dedica muncii de creație. În plus, astfel de abordări sunt caracterizate de un alt tip de învățare decât cel uman, bazat pe cantități mari de date care pentru un om ar fi imposibil de prelucrat în limite de timp rezonabile. De exemplu, tehnicile de clasificare a textelor și de achiziție a cunoștințelor sunt implicate într-un număr de domenii: *domeniul financiar* (de exemplu, detectarea cazurilor de spălăre a banilor, managementul riscului, managementul relațiilor cu clienții, detectarea fraudei, analiza automată a rapoartelor financiare [GDKS20]), *științe politice* (de exemplu, analiza comportamentului unei populații în cadrului unui vot și elaborarea politicilor sau urmărirea conflictelor internaționale [CJ20]), *studii clinice* (de exemplu, sprijin în diagnostic, potrivirea studiilor clinice cu potențiali participanți, fenotiparea computațională [FRC13, FH+99]) sau *educație* (de exemplu, notarea automată, sisteme inteligente de îndrumare și tutoriat, evaluarea eseurilor scrise de studenți și a cunoștințelor acestora [Bur09]).

Dintre aceste aplicații de procesare a limbajului natural, studiile din această teză se concentrează pe două task-uri de clasificare a textului: *detecția emoțiilor* și *identificarea autorului unui text*. Pentru

prima dintre acestea, *detecția emoțiilor din text*, au fost dezvoltate o serie de resurse pentru limba română, și anume un lexicon [BL17, LB17, LB19] și două corpusuri. Aceste resurse au fost folosite în studiul exprimării emoționale în texte literare în limba română [Bri19, LBB21]. Pentru cel de-al doilea task, *identificarea autorului unui text*, au fost propuse o serie de modele de *învățare profundă* pentru a identifica autorii unor poezii în limba română [BCL21]. De asemenea, a fost evaluată capacitatea unor sisteme similare de procesare a limbajului natural de a rezolva problema identificării programatorului care a scris un anumit fragment de cod sursă [CLB22, LBCC22]. În continuare, vom detalia motivația abordării fiecăreia dintre aceste task-uri.

Problemele abordate

Detecția emoțiilor în text

Recunoașterea emoțiilor din text este un domeniu extrem de relevant în contextul curent, în principal datorită gamei variate de cazuri de utilizare, de la analiza opiniilor în contexte comerciale până la studii psihologice în medii clinice. Detectarea emoțiilor din text poate juca un rol semnificativ și în educație, prin proiectarea de software educațional în care componentele emoționale ale interacțiunii se adaptează contextului și, practic, în orice sistem sau proces care implică conținut emoțional. Deoarece emoțiile sunt o parte vitală în alcătuirea unui individ și, în consecință, un fir țesut fără echivoc în existența umană, gama de aplicații ale detectării și analizei automate a emoțiilor este destul de largă.

Cu toate acestea, în ciuda popularității domeniului, există anumite dificultăți în a detecta emoțiile exprimate în text. Câteva dintre aceste provocări sunt: complexitatea inerentă a exprimării emoțiilor, lipsa seturilor de date de calitate și limitările modelelor computaționale actuale [STZ18]. Prima dintre aceste provocări se referă la o serie de aspecte pe care le prespune exprimarea emoțiilor în text, dintre care unele prezentând dificultăți în a fi surprinse chiar și pentru oameni, nu doar pentru calculatoare. Spre exemplu, aceeași emoție poate fi transmisă în mai multe moduri, fiecare cu vocabular diferit și uneori constructe semantice contrastante, inclusiv ironie și sarcasm, limbaj figurativ sau referințe dependente de context. În mod similar, numărul de semnificații și fațete emoționale posibile pe care o singură propoziție le poate încorpora sunt aspecte problematice atât în adnotarea de exemple, cât și în identificarea propriu-zisă a emoțiilor din text. Finalmente, o provocare fundamentală într-un astfel de task este complexitatea inerentă a taxonomiilor de emoții - deși există mai multe moduri de a modela stările emoționale, niciunul nu reușește să organizeze și să clasifice definitiv toate emoțiile și relațiile dintre acestea. În ceea ce privește lipsa datelor de calitate, acest lucru este de o importanță deosebită în modelele bazate pe învățarea din exemple, cum sunt majoritatea metodelor de învățare profundă utilizate în prezent în domeniu. Adnotarea conținutului emoțional este o sarcină recunoscută ca fiind dificilă [Öhm20], din următoarele motive: perspectiva adnotatorului și influența acesteia în etichetarea conținutului emoțional, deoarece experiența personală dictează înțelegerea diferitelor contexte și a răspunsurilor emoționale pe care le provoacă; dinamica emoțională dintr-o conversație dată și diferitele fațete ale emoțiilor din text - de exemplu, emoția *transmisă de autorul textului* și emoția *pe care o resimte cititorul*; și, nu în ultimul rând, întrebările care rămân cu privire la cât de adecvată este aplicarea fără modificări a teoriilor psihologice ale interacțiunii umane ca modele emoționale în ceea ce este, fără îndoială, o modalitate limitată de comunicare. În cele din urmă, ineficiența relativă a modelelor actuale poate fi explicată pe baza ultimei observații. Modelarea task-ului de detectare a emoțiilor este în sine o sarcină dificilă, deoarece se poate defini ca o problemă de clasificare binară, o problemă de clasificare cu mai multe clase, cu una sau mai multe etichete pentru un fragment de text sau o problemă de regresie (predicția intensității emoțiilor). Mai mult decât atât, domeniul prelucrării limbajului natural are încă un drum lung de parcurs în rezolvarea problemei înțelegerii limbajului

natural, întrucât abordările existente reușesc să surprindă doar într-o oarecare măsură aspectele semantice ale limbajului. Aspectele emoționale și semantice ale limbajului sunt strâns legate, ceea ce face ca dezvoltarea reprezentărilor de text îmbogățite cu emoții să fie un aspect crucial al cercetării.

Abordările timpurii, bazate pe reguli, în cadrul task-ului de detectare a emoțiilor din text nu au reușit să abordeze complexitățile problemei într-un mod satisfăcător. Cu toate acestea, progresele recente bazate pe modele de învățare profundă încurajează explorarea continuă a acestui domeniu, cele mai recente studii rezolvând, fie parțial, fie complet, multe dintre aceste dificultăți. Cercetarea noastră se concentrează pe *recunoașterea emoțiilor din texte în limba română*, și are obiectivul de a contribui la dezvoltarea și alinierea domeniului național la standardele internaționale. Alegerea subiectului este motivată de natura sarcinii, care este extrem de interesantă, atât din perspectivă psihologică, prin studiului modului de exprimare al emoțiilor în texte în limba română, cât și din perspectivă computațională, datorită implicațiilor sale în ceea ce privește inteligența artificială.

Un prim pas în atingerea acestui obiectiv a fost dezvoltarea unui lexicon, denumit **RoEmoLex**, care înregistrează asocierile dintre o serie de cuvinte și opt emoții de bază (*Anticipare, Bucurie, Dezgust, Frică, Furie, Încredere, Tristețe, Surpriză*) și două polarități (*Pozitivitate* și *Negativitate*). Următorea etapă a fost reprezentată de colectarea datelor sub formă de texte în limba română. Astfel, a fost creat corpusul **RoWikiLit**. Pentru anumite submulțimi de texte din acest corpus, am examinat conținutul emoțional cu ajutorul lexiconului RoEmoLex. Categoria operelor literare a fost aleasă din mai multe motive, printre care și faptul că unele sunt disponibile online, în mod gratuit. În plus, pentru lucrările binecunoscute, există un număr considerabil de analize literare și de recenzii ale criticilor și ale publicului. Prin urmare, performanța modelelor computaționale poate fi evaluată folosind aceste surse suplimentare, care nu sunt lipsite de importanță în sarcini complexe, cum ar fi studiul emoțiilor în text. În continuare, am folosit modele de *învățare nesupervizată* pentru a investiga relația dintre aspectele lingvistice ale textelor literare și conținutul emoțional. În cele din urmă, am dezvoltat un set de date pentru task-ul de detectare a emoțiilor în text. Acesta conține propoziții și fraze în limba română adnotate cu 14 emoții și 2 valențe, și a fost denumit **RoEmoData**. Setul de date RoEmoData va fi elementul de bază al experimentelor ulterioare de detectare a emoțiilor în texte în limba română.

Identificarea autorului unui text

Analiza proveniență documentelor și a autorilor acestora (*Authorship Attribution (AA)*) este un domeniu al NLP care se referă la extragerea de informații despre un autor pe baza textului și clasificarea textelor în clase de autori. Astfel, informații despre stilul autorului și caracteristicile sociolingvistice ale acestuia sunt determinate pe baza unei analize, care poate conduce chiar la identificarea autorului textului. Există mai multe task-uri înrudite în acest domeniu: (1) *Identificarea autorului (AA)* - identificarea autorului (autorilor) unui set de texte, (2) *Crearea profilurilor de autori (Author Profiling (AP))* - descoperirea datelor demografice ale autorului, cum ar fi vârsta, sexul, ocupația și nivelul educațional, (3) *Verificarea calității de autor (Authorship Verification (AV))* - confirmarea sau infirmarea faptului că un text a fost scris de un anumit autor, și (4) *Detectarea plagiatului (Plagiarism detection (PD))* - căutarea de paragrafe reproduse din textele altor autori. De exemplu, *plagiatul în domeniul dezvoltării de software* este o problemă majoră în mediile educaționale și corporatiste, care poate fi abordată printr-o astfel de analiză de proveniență prin identificarea asemănărilor textuale între fragmente de cod sursă [BT07].

Descoperirea trăsăturilor stilistice latente ale autorilor dintr-un corpus de date sub formă de text are multe aplicații în diverse domenii [DFIC19, SMS17]: *literatură și istorie* (pentru a determina paternitatea literară a unor documente literare sau istorice anonime sau în litigiu; detectarea de pastișe; compararea stilurilor diferiților autori); *educație* (pentru a înțelege personalitatea studenților; pen-

tru a detecta plagiatul în activitatea academică [HLLA14]); *analiza conținutului de pe rețelele sociale* (construirea de profiluri de utilizator care includ identitatea, caracteristicile sociolingvistice și opiniile acestora [MGAPD+17]); *investigarea infracțiunilor cibernetice* (pentru a identifica activitățile periculoase, cum ar fi spam-ul, mesajele de răscumpărare, hărțuirea, spălarea banilor, distribuirea ilegală de materiale în e-mailuri, bloguri sociale sau mesaje SMS-text [STASH19]; pentru a furniza dovezi în instanțele judecătorești); *ingineria software și securitatea cibernetică* (pentru a identifica autorul unui anumit fragment de cod sursă [KKG+19]; pentru a detecta plagiatul în cadrul unui produs software [BT07] și codul periculos; pentru a preveni atacurile cibernetice).

Provocările întâlnite în task-ul de determinare a autorului unui text pot fi sumarizate după cum urmează. În primul rând, există îndoieli cu privire la fiabilitatea tehnicilor de atribuire automată a autorului în situații în care acuratețea este primordială (de exemplu, în domeniul legal). În acest sens, se pune problema existenței unui stil caracteristic de scriere specific fiecărui individ [VHBT+05]. În al doilea rând, nu există o soluție universal acceptată pentru task-ul de identificare a autorului unui text dat. În consecință, multe tehnici prezentate în domeniu se dovedesc a funcționa în experimente la scară mică, dar rareori există validare pe seturi de date din alte domenii sau o comparație riguroasă cu lucrări similare. Chiar și la momentul actual, genul textelor și reprezentativitatea eșantionului sunt probleme cruciale [Juo06]. De exemplu, un model de învățare automată care este antrenat pe lucrările poetice ale unui autor ar putea să nu recunoască lucrările sale în proză. Prin urmare, pentru utilizare în contexte din lumea reală, poate fi utilă dezvoltarea modelelor care permit determinarea identității autorului pentru texte din mai multe genuri. De asemenea, modificările aduse manuscrisului unui scriitor de către editori sau tipografi ar putea face dificilă separarea adevăratului stil al autorului în lucrările publicate. În mare măsură, în studiile de atribuire a autorului pentru un text dat, există o convenție conform căreia forma finală, publicată, a textului aparține în întregime autorului.

Identificarea autorului unui fragment de cod, sau *Software Authorship Attribution (SAA)*, poate fi definită ca procesul de identificare a programatorilor care au scris anumite fragmente de cod sursă date. Această identificare se realizează în urma unei determinări prealabile cu privire la caracteristicile lor distinctive în ceea ce privește *stilul de programare*. Stilul de programare al unui dezvoltator de software poate fi caracterizat printr-un număr de preferințe și alegeri făcute în procesul de scriere a codului, inclusiv, dar fără a se limita la denumirea variabilelor, bibliotecile, structurile de date și secvențele de control utilizate. Stilul de programare are, de asemenea, o legătură directă cu experiența și competența programatorului și, desigur, cu modul particular de gândire logică și creativă folosite pentru a rezolva o problemă de programare [ARA+19].

Deși este adevărat că limbajele de programare au gramatici mult mai puțin flexibile decât limbajele naturale, *stilul de programare al unui programator* poate fi totuși identificat. Stilul de programare al unui programator se referă la preferințele acestuia în exprimarea constructelor logice, definirea structurilor de date și utilizarea ulterioară a acestora, precum și la denumirea variabilelor și constantelor și la apelurile efectuate către seturi de date fixe și temporare [SAM96].

Domeniul care cuprinde aceste task-uri de atribuire de autor unui fragment de cod sursă este unul care doar recent a atras interes la scară largă. Acest fapt se datorează nevoilor practice tot mai complexe din mediul academic și cel economic.

Detectarea plagiatului, și a *ghostwriting*-ului (detecția sarcinilor de programare externalizate ale studenților) sunt probleme care implică identificarea programatorului care a scris un fragment de cod, pentru care soluțiile AA astfel construite ar fi extrem de utile în mediul academic. În domeniul securității cibernetice, în care atât indivizii, cât și organizațiile pot fi ținte, atacurile cibernetice bazate pe software periculos (*adware*, *spyware*, viruși și multe altele) sunt probleme importante care pot fi prevenite cu ajutorul unor sisteme de analiză a proveniențeiei codului sursă. Domeniul ingineriei software beneficiază de progresele din domeniul SAA în rezolvarea diferitelor sarcini precum

întreținerea software-ului, analiza calității software-ului, managementul proiectelor, detectarea plagiatului cu efecte importante în problemele de copyright și licențiere [BKR+21].

În această teză sunt prezentate abordări de *învățare profundă* pentru a rezolva task-uri de *identificare a autorului unui text literar* și, respectiv, *atribuirea de autor pentru un fragment de cod sursă*. În acest sens, au fost dezvoltate trei modele: $AutoAt$, pentru task-ul de identificare a autorului unui text literar, $AutoSoft$ cu extensia sa $AuoSoft^{ext}$, pentru un task de clasificare SAA cu mai multe clase și $SoftId$, pentru o sarcină de clasificare SAA cu o singură clasă. Abordările propuse rezolvă cu succes task-urile date, comparația cu alți algoritmi de învățare automată existenți în contexte cu mai multe clase fiind favorabilă. În plus, modelele propuse oferă avantaje precum calculul probabilităților de apartenență la o clasă de autor și posibilitatea de a reîncadra sistemele pentru a rezolva sarcinile de atribuire a autorului în varianta cu o singură clasă prin recunoașterea autorilor „*cunoscuți*" și „*necunoscuți*" ($AutoSoft$ și $AutoSoft^{ext}$).

Contribuții originale

Cercetarea noastră s-a axat pe două direcții principale: (1) dezvoltarea de *resurse pentru detectarea emoțiilor din text* pentru limba română și studiul relațiilor dintre diversele trăsături lingvistice ale textelor și conținutul emoțional cu ajutorul tehnicilor de *învățare nesupervizată* și (2) investigarea modelelor *învățare profundă* pentru sarcina de atribuire a autorului, atât pentru texte în limbaj natural, cât și pentru fragmente de cod software. Astfel, rezultatele și principalele noastre contribuții sunt separate pe baza acestor direcții, care sunt prezentate în detaliu în Capitolele 2 și 3:

1. **Detecția emoțiilor**

   Numeroasele studii interesante privind detectarea emoțiilor în texte în limba engleză au declanșat un interes pentru dezvoltarea domeniului corespondent pentru limba română. În acest scop, am creat o serie de resurse pentru limba română, și am propus trei tipuri de analize ale conținutului semantico-emoțional: (1) folosind măsuri cantitative și statistici simple, (2) folosind o teorie matematică a analizei datelor numită *Formal Concept Analysis (FCA)* și (3) folosind o metodă de învățare nesupervizată, și anume *Hierarchical Agglomerative Clustering (HAC)*. Scopul principal al acestor analize a fost de a descoperi tiparele emoționale și relația lor cu conținutul semantic corespunzător. Rezultatele noastre privind această linie de cercetare au fost următoarele:

   (a) Primul pas a fost dezvoltarea **RoEmoLex**, un lexicon care cuprinde o serie de cuvinte adnotate cu conținut emoțional. RoEmoLex a fost dezvoltat în trei etape, documentate în trei articole științifice publicate: prima versiune a lexiconului ([LB17]) a constat într-un număr mic de termeni și a inclus corecturi față de traducerea automată a resursei originale (EmoLex [MT10]). În RoEmoLex v.2. ([BL17]) au fost adăugați termeni noi. A treia versiune a lexiconului [LB19] a fost supusă unei verificări riguroase și unei analize aprofundate a conținutului său pentru a prezenta o resursă gata de a fi utilizată. Detalii despre a treia versiune a RoEmoLex sunt incluse în secțiunea 2.1.1. RoEmoLex este disponibil public la `https://www.cs.ubbcluj.ro/~ica/romanian-nlp-resources/index.html`, și a fost deja folosit într-o serie de studii în psihologie [HO20] și în domeniul științelor politice [Bir20].

   (b) În continuare, s-a propus o explorare a conținutului lexiconului RoEmoLex prin FCA, o metodă care permite o reprezentare intuitivă a relațiilor semantice și lexicale dintre termeni și valențele și emoțiile asociate acestora. Rezultatele au fost publicate în [LB19]. În

această lucrare au fost evidențiate echivalențele dintre ordonările formale ale conceptelor bazate pe emoții și reprezentările ierarhice ale termenilor pe bază semantică (de exemplu, relația de *hypernymy/hyponymy*). O astfel de distribuție a etichetelor emoționale în seturi de cuvinte semnificative din punct de vedere semantic ar putea fi exploatată în aplicații ulterioare care integrează resursa RoEmoLex. Detalii despre acest studiu pot fi găsite în Secțiunea 2.2.1.

(c) În următorul nostru studiu, am folosit RoEmoLex pentru a investiga opera poetică a unui autor român (Mihai Eminescu) cu privire la mai multe aspecte lexicale, semantice și emoționale. Rezultatele au fost publicate în [Bri19]. Detalii despre această analiză sunt incluse în Secțiunea 2.2.2. Contribuția originală a acestei lucrări este reprezentată de o analiză cantitativă aprofundată a operelor poetice ale lui Mihai Eminescu și o examinare atentă a relației dintre etapele expresiei artistice ale autorului și aspectele cuantificabile ale limbajului. În special, trăsăturile emoționale și de valență care au fost definite au fost unice și, deși rezultatele analizei statistice simple în acest sens nu au permis identificarea unor concluzii definitive, ele au informat studiile noastre ulterioare.

(d) Deoarece studiul anterior a cuprins doar o analiză cantitativă a unui set limitat de documente, am creat un set mai mare de date de texte literare, **RoWikiLit**, din surse online, proces descris în Secțiunea 2.1.2. Acest set de date a fost folosit în două dintre studiile noastre, dintre care unul a implicat studiul tiparelor semantice și emoționale prin tehnici de *învățare nesupervizată* într-un subset adnotat cu teme literare al poeziei lui Mihai Eminescu. Rezultatele au fost publicate în [LBB21], iar contribuția originală a acestei lucrări se referă la aplicarea unei tehnici de învățare automată nesupervizată pe un corpus de poezie pentru a explora asocierile dintre conținutul tematic și cel emoțional. În urma acestui studiu, s-a observat că analiza computațională a operei susține analize literare. Detalii despre acest studiu se regăsesc în Secțiunea 2.2.3.

(e) Obiectivul final al studiilor de pe această direcție de cercetare este dezvoltarea de sisteme de detecție a emoțiilor din texte în limba română. Pentru a atinge acest obiectiv, un set de date care include propoziții și paragrafe în limba română din 4 categorii semantice diferite (fragmente sau replici din conversații sau intrări de jurnal folosind limbaj literal, text care implică limbaj figurativ și expresii specific românești, fragmente literare și fragmente de știri și din articole) adnotate cu 14 emoții și 2 valențe prin *crowdsourcing*. Setul de date construit a fost denumit **RoEmoData**, iar statistici despre forma sa curentă pot fi găsite în Secțiunea 2.1.3.

2. **Identificarea autorului unui text**

   Întrebarea inițială care a condus la studii ulterioare pe această direcție de cercetare a fost următoarea: pot fi autorii operelor literare identificați automat pe baza anumitor reprezentări de text? Astfel, scopul nostru a fost să definim un nou model de *învățare profundă* pentru a răspunde la această întrebare. S-a realizat un experiment relativ reușit în definirea modelului ***AutoAt***, un model de clasificare din categoria de învățare profundă bazat pe un ansamblu de *autoencoderi* pentru *identificarea autorului unui text literar în mod supervizat*. În mod specific, au fost considerate opere poetice aparținând unor autori români. În continuare, s-a investigat capacitatea modelelor bazate pe tehnica de învățare profundă *autoencoder* de a face distincția între stilul programatorilor în fragmente de cod sursă. Au fost dezvoltate următoarele trei modele bazate pe învățarea profundă:

   (a) ***AutoAt***. Primul model creat a fost bazat pe un ansamblu de *autoencoderi* cu mai multe straturi ascunse. Modelul a fost evaluat pe un subset al setului de date RoWikiLit. Ideea principală a acestui model a fost de a antrena câte un autoencoder pentru fiecare autor în setul de date dat, cu scopul de a codifica caracteristicile particulare (atât structurale, cât și conceptuale) ale documentelor aparținând aceluiași autor. Rezultatele acestui studiu au fost publicate în [BCL21] și sunt descrise și în secțiunea 3.1. S-a obținut o valoare de de 0,81 $\pm 0,02$ pentru metrica F-score. Comparația cu alți clasificatori existenți a fost favorabilă în 5 din 6 cazuri.

   (b) ***AutoSoft*** and ***AutoSoft^{ext}***. Modelul de bază, $AutoSoft$, este similar cu $AutoAt$, conceput cu scopul de a investiga performanța unui model de clasificare supervizat bazat pe *autoencoderi* pentru identificarea programatorului care a scris un fragment de cod sursă. Reprezentarea codului sursă a fost inspirată din domeniul procesării limbajului natural, *embedding*-urile `doc2vec` dovedindu-se a fi capabile să codifice stiluri de programare pentru dezvoltatorii software luați în considerare. Mai mult, s-a propus o extensie a clasificatorului $AutoSoft$, denumită $AutoSoft^{ext}$, prin care s-a propus recunoașterea unei clase de „autor necunoscut" în plus față de setul dat de clase de programatori. Rezultatele obținute în acest studiu au fost detaliate în [CLB22] și sunt incluse și în Secțiunea 3.2. Cele două modele au fost introduse ca dovadă a fezabilității conceptului și evaluate pe subseturi ale setului de date Google Code Jam, pe care s-au obținut valori $F\text{-}score$ între 0,902 și 0,986 pentru clasificatorul $AutoSoft$. Modelul $AutoSoft$ depășește alți clasificatori din literatură în **69%** dintre cazuri. Pentru modelul $AutoSoft^{ext}$, cele mai bune rezultate au fost obținute într-o configurație de testare folosind secvențe de *tokeni* de lungime N, în particular $N = 5$. În 7 task-uri de clasificare binară diferite cu autori din clasele *original* și *autor necunoscut*, s-au obținut valori $F\text{-}score$ între 0,983 și 0,996. Comparația cu *One Class Support Vector Machines*, un alt clasificator din literatură de acest tip, a fost favorabilă. Acesta din urmă a depășit $AutoSoft^{ext}$ doar în termeni de $Specificitate$.

   (c) ***SoftId***. Acest model a fost dezvoltat ca un model de clasificare cu o singură clasă și a fost antrenat pentru a recunoaște stilul de programare al unui anumit grup de dezvoltatori de software. $SoftId$ a fost conceput cu scopul de a detecta dacă un anumit fragment de cod sursă a fost scris de un dezvoltator din grupul dat sau de un programator „necunoscut". Spre deosebire de cele două modele anterioare, $SoftId$ se bazează pe un singur *autoencoder* antrenat pe reprezentări textuale ale fragmentelor de cod sursă. Rezultatele acestui studiu au fost detaliate în [LBCC22]. În 24 de experimente pe subseturi ale setului de

date Google Code Jam, pentru care s-a variat tipul de reprezentare și valoarea lui N pentru secvențe de *tokeni* (cuvinte-cheie, operatori, nume de variabile, etc), $SoftId$ depășește clasificatorul *One-Class Support Vector Machines (OSVM)* utilizat pentru comparație în 95% din cazuri. O descriere detaliată a experimentelor noastre și discuția bazată pe rezultatele obținute este disponibilă în Secțiunea 3.3.

Structura tezei

Structura tezei este următoarea.

Primul capitol descrie contextul teoretic pentru task-urile luate în considerare în teza noastră și o prezentare generală a domeniilor abordate, precizând studiile deja existente și modul în care abordările noastre se raportează la acestea. Acest capitol este împărțit în trei secțiuni.

În Capitolul 2 sunt prezentate studiile noastre în detectarea emoțiilor în texte în limba română. Acest capitol are două secțiuni principale, Secțiunea 2.1, care prezintă resursele dezvoltate în vederea rezolvării aceastui task, și Secțiunea 2.2, care prezintă analizele nesupervizate ale conținutului semantic și emoțional în texte în limba română folosind resursele dezvoltate.

În cel de-al treilea capitol, sunt prezentate experimentele realizate folosind tehnici de învățare profundă pentru task-ul de identificare a autorului unor texte, mai specific folosirea modelelor de tip *autoencoderi* pentru date literare (Secțiunea 3.1) și cod sursă (Secțiunile 3.2 și 3.3) .

Capitolul 1

Fundamente teoretice

Acest capitol oferă o scurtă prezentare a lucrărilor conexe referitoare la tipul de resurse și task-urile propuse în această teză. De asemenea, sunt incluse scurte descrieri ale algoritmilor utilizați.

Capitolul este împărțit în trei secțiuni: în primul rând, o secțiune care detaliază studiile existente cu privire la resursele și task-urile prezentate; în al doilea rând, o secțiune în care sunt prezentate fundamentele teoretice specifice analizei conținutului emoțional în text și perspectivele literare asupra operei lui Mihai Eminescu și, în final, o secțiune în care sunt descrise metodele și tehnicile folosite în abordările prezentate.

Prima dintre aceste secțiuni reprezintă o privire de ansamblu asupra studiilor existente în cele două direcții principale de cercetare pe care le abordează teza: *detectarea emoțiilor din text* și *identificarea autorului unui text*. Structura internă a acestei secțiuni o oglindește pe cea a tezei în ansamblu, primele subsecțiuni concentrându-se pe lucrările conexe privind resursele pentru analiza conținutului emoțional dintr-un text, analiza nesupervizată a asocierilor între latura semantică și cea emoțională, atât în ceea ce privește termeni, cât și texte, analiza cantitativă a stilului literar, și, respectiv, gruparea unor opere literare pe bază de conținut emoțional. Aceste subsecțiuni acoperă prima direcție de cercetare abordată, detectarea și analiza emoțiilor.

În ceea ce privește existența resurselor de *analiză a emoțiilor pentru texte în limba română*, atât în ceea ce privește lexicoanele, cât și corpusurile, o analiză a literaturii de specialitate indică faptul că sunt puține lucrări care vizează limba română, în comparație cu multitudinea de resurse care există pentru limba engleză. În consecință, există o tendință similară în ceea ce privește aplicațiile bazate pe emoții care iau în considerare limba română. Cu toate acestea, există câteva grupuri de cercetare care au dezvoltat și continuă să dezvolte resurse cuprinzătoare și fiabile [TBM14, PT18, CD21] pentru o varietate de task-uri. Studiile prezentate în această teză vizează analiza emoțiilor în texte în limba română, subdomeniu pentru care există mai puține studii decât, de exemplu, pentru analiza atitudinilor în texte și pentru care doar recent au fost propuse modele supervizate [CD21].

Următoarele două subsecțiuni includ descrieri ale celor două probleme de identificare a autorului unui text care au fost abordate - *identificarea autorului unor texte literare* și *identificarea autorului unui fragment de cod sursă*. Sunt prezentate lucrări conexe pentru ambele task-uri, furnizând astfel baza teoretică pentru a doua direcție de cercetare prezentată în teză, *identificarea autorului unui text*. Principalele contribuții ale cercetării noastre în cadrul acestei direcții de cercetare sunt: (1) dezvoltarea unui model general bazat pe *autoencoderi* pentru identificarea autorului, atât în texte literare [BCL21], cât și în software [CLB22] și (2) flexibilitatea modelelor dezvoltate, care pot fi adaptate ușor pentru a aborda diferite formulări ale problemei de atribuire a autorului unui text.

În a doua secțiune a capitolului sunt prezentate fundamente teoretice specifice task-urilor abordate. În mod particular, pentru studiile care implică analiza conținutului emoțional al unui text, este

descris modelul emoțional al lui Robert Plutchik [Plu80], care este folosit în crearea resurselor pentru acest task. De asemenea, sunt prezentate aspecte ale operei lui Mihai Eminescu, care este autorul pe care sunt concentrate studiile nesupervizate din Capitolul 2 - analiza cantitativă a stilului său poetic și gruparea poeziilor sale pe baza conținutului emoțional. Stilul lingvistic, etapele de creație și temele literare din opera lui Mihai Eminescu sunt descrise în această secțiune.

În final, sunt prezentate metodele și tehnicile folosite. În primul rând, ca bază teoretică pentru analiza nesupervizată a conținutului semantico-emoțional a termenilor din RoEmoLex, este descrisă metoda *FCA*. *FCA* reprezintă o formalizare matematică a înțelegerii filozofice a conceptelor, unde *conceptele* sunt definite prin seturi de *obiecte*, *atribute* și *relații* între ele ([GW97], [Wil05]). Această tehnică este folosită pentru a ordona și organiza conceptele formate din termeni RoEmoLex ca obiecte și etichetele emoționale aferente ca atribute.

În al doilea rând, este prezentat algoritmul de clusterizare aglomerativă ierarhică. HAC este o tehnică nesupervizată care produce o ierarhie de clustere, oferind informații despre modul în care clusterele de pe fiecare nivel din ierarhie sunt formate din clusterele de pe nivelul anterior. Acest algoritm a fost folosit în studiul conținutului emoțional al poeziei lui Mihai Eminescu.

Apoi, sunt discutate reprezentările distribuite de cuvinte și text, cu accent pe modelele `word2vec` și `doc2vec`. În acest context, termenul *embeddings* se referă la metodele și tehnicile concepute pentru a obține vectori cu valori reale ca reprezentări pentru cuvinte sau documente individuale. Aceste reprezentări vectoriale distribuite sunt învățate pe baza utilizării cuvintelor într-un corpus dat, cel mai adesea cu ajutorul unei rețele neuronale. În acest mod, reprezentări ale unităților cu semnificații similare sunt mai apropiate în spațiul vectorial dat. Reprezentările `doc2vec` au fost folosite în majoritatea studiilor noastre de identificare a autorului unui text dat.

La final, sunt descrise modelele de învățare profundă denumite *autoencoderi*. Un autoencoder este compus din două rețele neuronale (denumite *encoder* și *decoder*) care sunt antrenate pentru a aproxima funcția identică (adică pentru a reconstrui ceea ce este dat ca input), fiind astfel cunoscute ca tehnici de *învățare auto-supervizată*. Toate modelele de identificare a autorului propuse - $AutoAt$, $AutoSoft$ și $AutoSoft^{ext}$ și $SoftId$ se bazează pe autoencoderi, atât sub forma unui ansamblu de autoencoderi ($AutoAt$ și $AutoSoft$), cât și a unui singur autoencoder antrenat pentru a răspunde problemei date ($SoftId$).

Capitolul 2

Dezvoltarea de resurse și aplicații pentru detectarea emoțiilor în texte în limba română

Importanța detectării emoțiilor din text se datorează multitudinii de aplicații care se pot construi pe baza unei soluții pentru acest task în domenii precum psihologie, sociologie sau științe politice. Provocările în detectarea emoțiilor se datorează în principal faptului că teoriile și modelele emoționale existente se referă o înțelegere mai largă a conceptului de afect, care include procese cognitive, comportament și schimbări psihofiziologice, iar comunicarea textuală, în mod evident, nu le poate încorpora pe toate. Mai exact, ambiguitatea textului scris (de exemplu, folosirea sarcasmului, a ironiei, a limbajului figurat etc.), varietatea de moduri în care poate fi exprimată o anumită emoție și numărul de interpretări care pot fi atribuite unui text de către un cititor reprezintă câteva dintre aceste provocări. În consecință, dezvoltarea resurselor de calitate pentru analiza computațională a emoțiilor exprimate în text este crucială, la fel cum este și studiul diferitelor tipuri de text cu ajutorul acestora. Astfel de studii ar putea oferi noi perspective cu privire la analizele existente (de exemplu, în studiile literare) sau pot oferi puncte de plecare în recunoașterea automată a emoțiilor în texte (de exemplu, în conținutul de pe rețelele sociale).

În acest capitol sunt descrise eforturile noastre de a contribui la dezvoltarea domeniului de detectare a emoțiilor din texte în limba română. În acest sens, sunt urmărite două direcții de cercetare: în primul rând, crearea de resurse pentru limba română, și în al doilea rând, explorarea asocierilor dintre conținutul semantic și cel emoțional al cuvintelor și textelor într-o manieră *nesupervizată*, cu ajutorul resurselor dezvoltate.

Studiile incluse în acest capitol au fost publicate și în cinci lucrări originale [[BL17, BL18, LB17, LB19, LBB21]]. Un scurt rezumat privind principalele contribuții originale prezentate în acest capitol este inclus în continuare.

1. Secțiunea 2.1 detaliază procesul de dezvoltare a trei resurse de detectare a emoțiilor pentru texte în limba română: **RoEmoLex**, **RoWikiLit** and **RoEmoData**.

    - Procesul de dezvoltare în trei etape al **RoEmoLex** este descris în Secțiunea 2.1.1. RoEmoLex este un lexicon care include asocieri cuvânt-emoție. RoEmoLex conține 9177 de termeni adnotați cu opt emoții (*Anticipare, Bucurie, Dezgust, Frică, Furie, Încredere, Surpriză* și *Tristețe*) și două valențe (*Pozitivitate, Negativitate*). Primele două etape ale dezvoltării acestui lexicon au fost detaliate în două articole științifice [LB17, BL17], în timp ce într-o a treia lucrare a fost prezentată versiunea disponibilă public a lexiconului

(RoEmoLex v.3.), și o analiză a conținutului semantico-emoțional din cadrul anumitor termeni din lexicon [LB19]. Scopul construirii acestui lexicon a fost de a oferi un punct de plecare în task-ul analizei emoțiilor în texte în limba română, obiectiv care a fost atins, întrucât există cercetători care au început deja să folosească RoEmoLex în studii cu o componentă emoțională în diferite domenii.

- Secțiunea 2.1.2 prezintă **RoWikiLit**, un corpus de opere literare românești colectat în principal de pe site-ul Wikisource (varianta în limba română) și adnotat cu diverse informații pentru a facilita analiza computațională a unor aspecte diverse din cadrul operelor. În studiile noastre au fost folosite trei subseturi ale corpusului RoWikiLit: un subset de lucrări poetice ale autorului Mihai Eminescu adnotate cu data publicării și etapa de creație din care fac parte [Bri19], un subset al poeziei lui Mihai Eminescu etichetat cu teme literare [LBB21], și un subset de lucrări poetice a 8 autori români [BCL21]. Datele din acest corpus au fost colectate cu scopul de a construi un set de texte în limba română gata de utilizare în modele de învățare, care evită problemele legate de confidențialitate și drepturi de autor. Toate operele preluate de pe site-ul WikiSource sunt texte libere de dreptul de autor.
- Secțiunea 2.1.3 prezintă **RoEmoData**, un set de date care cuprinde texte în limba română creat în vederea experimentelor de recunoaștere a emoțiilor din text bazate pe modele de *învățare automată* și de *învățare profundă*. Această secțiune oferă detalii despre conținutul setului de date, procesul aferent de adnotare și o scurtă analiză a porțiunii din setul de date care este complet adnotată. Acest studiu reprezintă direcția noastră actuală de cercetare.

2. Secțiunea 2.2 prezintă studiile realizate în manieră nesupervizată cu privire la conținutul semantic-emoțional al cuvintelor și textelor.

   - Secțiunea 2.2.1 include o descriere detaliată a lucrării [LB19], care are ca scop examinarea structurii și conținutului RoEmoLex prin FCA, o metodă de explorare a datelor. Utilizând această tehnică, s-au descoperit dependențe între cuvinte la nivelul conținutului emoțional, s-au construit ierarhii conceptuale și s-au realizat comparații cu informații din ontologia lexicală RoWordNet. Rezultatele încurajează cercetările ulterioare și posibilitatea de a integra RoEmoLex ca subcomponentă a sistemelor de analiză a emoțiilor pentru texte în limba română.
   - Secțiunea 2.2.2 prezintă o analiză computațională a unui corpus de poezie românească. În mod specific, este vorba despre opera poetică a lui Mihai Eminescu, un poet reprezentativ pentru literatura română. Au fost examinate o serie de caracteristici care se referă la vocabularul folosit, complexitatea limbajului și conținutul emoțional specific fiecărei etape a creației artistice. Rezultatele arată că, în privința anumitor aspecte, observațiile teoretice din domeniul criticii literare găsesc corespondență în analiza cantitativă, indicând faptul că această abordare și, în special, vizualizarea rezultatelor, ar putea fi folosite ca instrument de sprijin pentru interpretarea operelor literare. Acest studiu este publicat în [Bri19].
   - Secțiunea 2.2.3 prezintă o analiză bazată pe conținut emoțional a operei lui Mihai Eminescu. În acest studiu, este luat în considerare un corpus de 131 de poezii, pe care este aplicat algoritmul de *clusterizare ierarhică aglomerativă (HAC)*. Reprezentările acestor poezii sunt obținute cu ajutorul lexiconului RoEmoLex, și se referă la procentul de cuvinte care au asociat conținut emoțional din fiecare poezie (și pentru fiecare emoție). Au fost

explorate conexiunile între conținutul tematic și tiparele emoționale, rezultatele arătând că unele dintre aspectele descrise în lucrările de analiză literară pot fi reproduse folosind modele computaționale. Acest studiu este publicat într-un articol științific original [LBB21].

2.1 Resurse pentru limba română

2.1.1 RoEmoLex

RoEmoLex (Romanian Emotion Lexicon) [LB17, LB19, BL17] este o resursă dezvoltată pentru a fi folosită în sisteme de detectare a emoțiilor în texte în limba română. RoEmoLex a pornit de la o traducere automată a unui lexicon pentru limba engleză (EmoLex [MT10]). Lexiconul a trecut prin trei etape de dezvoltare care au inclus corectarea traducerilor inițiale, re-adnotarea datelor, integrarea altor resurse și adăugarea de noi termeni și alinierea la ontologia lexicală RoWordNet.

Versiunea actuală a RoEmoLex conține 9177 de termeni adnotați cu opt emoții primare (după modelul lui Plutchik [Plu80]) (*Anticipare, Bucurie, Dezgust, Frică, Furie, Încredere, Surpriză, Tristețe*) și două etichete de polaritate (*Pozitivitate, Negativitate*).

2.1.2 RoWikiLit - Corpus de texte literare în limba română

În cartea sa, "Macroanaliza - Metode digitale & Istorie literară" [Joc13], Matthew Jockers explică beneficiile abordărilor computaționale în cercetări de natură literară. Autorul prezintă avantajele unei abordări combinate, bazată atât pe lectura tradițională, realizată de oameni, care citesc fiecare text în parte și în întregime (*"close reading"*) și cea bazată pe metode computaționale, în care este vorba de statistici sau reprezentări care surprind doar anumite aspecte ale textului, și nu includ o înțelegere a sa (*"distant reading"*). Altfel spus, în contextul analizelor literare tradiționale și posibilitatea realizării unor studii la scară largă pentru corpusuri digitale, autorul privește metodele computaționale ca o metodă nouă, mai potrivită pentru prelucrarea textelor și descoperirea aspectelor importante în cercetarea literară în epoca *"Big Data"*, care presupune un volum foarte mare de date, intractabil în limite rezonabile de timp din perspectivă umană. El afirmă că în timp ce lectura tradițională nu trebuie desconsiderată, fiind singura modalitate de a dezvălui „pepite" de cunoștințe, explorarea „venelor mai adânci îngropate sub masa de pietriș" poate fi la fel de relevantă. Concret, argumentul se referă la necesitatea metodelor de agregare și înțelegere a ambelor tipuri de informații - care sunt furnizate de lecturi tradiționale și, respectiv, de analize computaționale - pentru a obține o viziune cuprinzătoare și echilibrată asupra operelor luate în considerare.

Corpusul actual conține **5861** opere a **57** de autori români și include opere lirice (*poezie*), opere în *proză* din diferite specii literare (basme, povestiri, texte istorice, romane, nuvele), *piese de teatru* și *scrisori și articole de ziar* preluate dintr-o sursă online[1]. În plus, **147** de cărți de sine stătătoare și **65** volume cu poezii, eseuri, povești pentru copii sau piese de teatru au fost descărcate dintr-o librărie online care oferă o serie de cărți electronice gratuite[2]. Tabela 2.1 cuprinde o prezentare generală a acestui set de date.

Trei subseturi ale acestui set de date au fost utilizate în studii publicate sau care urmează a fi publicate, toate trei conținând opere poetice. Primul a fost un subset care a cuprins întreaga operă poetică a lui Mihai Eminescu, care a fost supusă unei analize cantitative din punct de vedere lexical, semantic și al conținutului emoțional [Bri19]. Cel de-al doilea a fost un subset de 131 de poezii ale

[1] https://ro.wikisource.org/wiki/Pagina_principală
[2] https://www.bestseller.md/ebooks.html?limba=130

Categorie	Număr de texte	Specii literare	Exemple de autori
poezie	2157	balade, meditații, ode pasteluri, elegii, epigrame, satire	M. Eminescu, V. Alecsandri G. Coșbuc
fabule	307	-	G. Asachi, G. Alexandrescu
articole	2130	articole din ziare cu tematică politică, recenzii și critici literare	M. Eminescu, I.L. Caragiale, G. Ibrăileanu
povești pentru copii	162	basme, povești moralizatoare	P Ispirescu, I. Creangă
povestiri	417	-	I. Slavici, L. Rebreanu, A. Holban B. Ștefănescu-Delavrancea, N. Gane
piese de teatru	57	comedii, tragedii	I.L Caragiale
romane	62	-	I. Slavici, A. Odobescu, L. Rebreanu, G. Mihăescu
alte tipuri de proză	569	eseuri, fragmente istorice	N. Iorga

Tabela 2.1: Statistici privind corpusul RoWikiLit (iulie 2021)

unui singur autor (Mihai Eminescu) pentru care au fost obținute manual informații suplimentare sub forma unor teme literare [LBB21]. În ultimul rând, un subset cu texte lirice ale celor mai prolifici 8 poeți români din setul de date a fost folosit pentru un studiu de identificare automată a autorului unui text [BCL21].

	Autori							
	Alexandru Macedonski	George Coșbuc	George Topîrceanu	Ion Minulescu	Mihai Eminescu	Octavian Goga	Vasile Alecsandri	Ștefan O. Iosif
Nr. de poezii	190	212	113	159	366	181	186	164
Nr. tokeni	39 403	124 809	31 525	35 380	182 270	37 761	72 025	30 870

Tabela 2.2: Descrierea setului de date folosit în experimentul de identificare a autorului unor texte literare [BCL21]

Utilitatea corpusului literar RoWikiLit a fost dovedită în cele două aplicații discutate mai sus, atât în sensul contribuțiilor la studiul computațional al operelor literare, cât și ca punct de plecare în experimente pentru task-uri mai generale de procesare a limbajului natural precum identificarea autorului unui text. Cu toate acestea, aceste task-uri s-au concentrat pe subseturi mici din RoWikiLit. Studiile la scară mai largă și utilizarea unor perspective diferite, de exemplu utilizarea traiectoriilor de evoluție a emoțiilor într-un text (în locul reprezentării bazate pe conținut emoțional global) sunt încă domenii de cercetare neexplorate pentru limba română care s-ar putea dovedi extrem de interesante.

2.1.3 RoEmoData

RoEmoData (*Romanian Emotion Detection Data Set*) este un set de date dezvoltat pentru detectarea emoțiilor în texte în limba română. Acest set de date conține aproximativ 6.000 de propoziții și fraze similare cu fragmente din intrări de jurnal și replici din conversații (subset **CDIA** și, respectiv, subset **EXPR**, care conține același tip de intrări, însă include limbaj figurativ și expresii), scurte fragmente din opere literare ale unor autori români (subset **LIT_FRAGMENTS**) și o serie de titluri de știri și fragmente din articole (subset **NEWS&ARTICLES**). Tabela 2.3 descrie forma curentă a setului de date.

Set	Numărul total de instanțe	Numărul de instanțe adnotate	% Adnotat
CDIA	2187	2187	100%
EXPR	2193	2193	100%
NEWS&ARTICLES	815	165	20.25%
LIT_FRAGMENTS	1340	165	12.31%
Total	**6535**	**4710**	**72.07%**

Tabela 2.3: Descrierea setului de date RoEmoData (aprilie 2022)

2.2 Analiza nesupervizată a conținutului semantico-emotional

Prezentarea din această secțiune se bazează pe lucrările noastre originale, publicate în [LB19] (Secțiunea 2.2.1), [Bri19] (Secțiunea 2.2.2) și [LBB21] (Secțiunea 2.2.3).

2.2.1 Analiza conceptuală formală a conținutului RoEmoLex

Această secțiune detaliază explorarea conținutului RoEmoLex prin FCA, o metodă care permite o reprezentare intuitivă a relațiilor semantice și lexicale dintre termeni și valențele și emoțiile asociate acestora. Rezultatele au fost publicate în [LB19]. Studiul propus a arătat că echivalențele dintre ordonările formale ale conceptelor bazate pe emoții și reprezentările semantice ierarhice ale termenilor (de exemplu, pe baza relațiilor de *hypernymy/hyponymy* între termeni) și distribuția etichetelor emoționale în seturi de cuvinte semnificative din punct de vedere semantic ar putea fi exploatate în aplicații ulterioare care integrează resursa RoEmoLex.

Principalele întrebări la care studiul și-a propus să răspundă sunt: (1) care este conexiunea între o ierarhie de substantive RoWordNet bazată pe relația *hypernymy/hyponymy* și ierarhia conceptuală aferentă generată de conținutul emoțional al termenilor (așa cum este identificat acesta de RoEmoLex)? (2) poate conținutul emoțional să delimiteze grupuri semnificative de termeni în RoEmoLex în contextul unor categorii conceptuale specifice? și (3) cum se poate introduce o metodă de calculare a unor scoruri emoționale relative pentru termeni bazată pe ierarhiile de concepte obținute?

2.2.2 Analiza cantitativă a stilului în poezia lui Mihai Eminescu

Această secțiune detaliază lucrarea noastră originală [Bri19]. Metodele stilistice cantitative urmăresc exprimarea anumitor aspecte ale unui text sub formă numerică, permițând astfel introducerea unor abordări computaționale rapide, puternice și precise pentru analiză. În timp ce în cazul literaturii de specialitate, validitatea și utilitatea unor astfel de studii este foarte controversată, nu se pot nega oportunitățile pe care le introduc metodele computaționale: în primul rând, explorarea unor seturi mari de documente în căutarea unor modele altfel greu de descoperit de către cititorii umani; în al doilea rând, posibilitatea descoperirii unor perspective noi prin examinarea trăsăturilor latente ale textelor. În acest studiu, este investigată opera poetică a unuia dintre cei mai importanți poeți români, Mihai Eminescu, printr-o varietate de metode cantitative care abordează aspecte lexicale, morfologice, semantice și emoționale ale textului. În această secțiune, propunem o comparație între rezultatele abordării computaționale și interpretările consacrate ale operei poetice a lui Mihai Eminescu pentru a evalua viabilitatea metodelor computaționale în studiile de stil poetic.

Pentru datele luate în considerare, am colectat 339 de poezii dintr-o sursă disponibilă online[3], subset care face parte și din setul RoWikiLit. Dintre acestea, au fost excluse numai lucrările care nu puteau fi asociate definitiv cu un an de publicare.

2.2.3 Clusterizarea poeziilor pe baza conținutului emoțional

Această secțiune prezintă detalii despre studiul [LBB21], care se concentrează pe un subset de 131 de poezii ale lui Mihai Eminescu din setul RoWikiLit (Secțiunea 2.1.2). Acest set de poezii este compus din poezii pentru care conținutul tematic a fost identificat manual cu ajuorul unor cărți de analiză literară și al unor antologii. Temele literare luate în considerare sunt: *timpul* (45 de poezii), *cosmogonia* (13 poezii), *condiția geniului* (9 poezii), *natura* (68 de poezii), *moartea* (32 de poezii), *viziunea referitoare la creația poetică* (17 poezii).

Scopul acestui studiu a fost să investigheze o abordare nesupervizată numită HAC în vederea identificării tiparelor emoționale din poezia lui Mihai Eminescu și asocierea acestora cu temele literare prezente în text. Caracteristicile emoționale definite pentru algoritmul de clusterizare au fost obținute cu ajutorul lexiconului RoEmoLex.

Abordarea folosită în acest studiu a constat în crearea de reprezentări vectoriale ale textelor poetice pe baza conținutului emoțional prin calculul unui scor pentru fiecare valență și emoție, în modul descris în ecuația 2.1, folosind RoEmoLex (vezi Secțiunea 2.1.1).

$$score(e, poem) = \frac{1}{l_{poem}} \cdot \sum_{t_i \in poem} 1_{t_i \in RoEmoLex_e} \tag{2.1}$$

[3] https://ro.wikisource.org/wiki/Autor:Mihai_Eminescu

Capitolul 3

Îmbunătățirea sistemelor de identificare a autorului unui text folosind autoencoderi

Identificarea autorului unui text este un task important într-o varietate de domenii, cum ar fi literatură, educație, inginerie software sau securitate cibernetică. În ceea ce privește atribuirea în mod automat a unui autor pentru texte în limba română, există puține studii care abordează această problemă dintr-o perspectivă computațională. În mod similar, există un număr relativ mic de lucrări care au exploatat caracteristici inspirate de domeniul de procesare a limbajului natural în task-ul de identificare a autorului unui program software. Acest capitol prezintă trei abordări bazate pe *autoencoderi* pentru task-ul de identificare a autorului: în primul rând, pentru texte literare (Secțiunea 3.1) și în al doilea rând, pentru programe software (Secțiunea 3.2 și Secțiunea 3.3). Modelele propuse reprezintă contribuții originale în domeniu și pot fi rezumate după cum urmează:

1. Secțiunea 3.1 detaliază lucrarea originală [BCL21]. Studiul propune sistemul $AutoAt$, un model de clasificare bazat pe tehnicile de învățare profundă numite *autoencoderi* pentru a *determina autorul unui text* literar. Experimentele sunt realizate folosind un subset al setului de date RoWikiLit de 1571 de poezii scrise de 8 poeți români. În cadrul experimentelor, se folosește o reprezentare distribuită a documentelor, numită `doc2vec`. Comparația modelului $AutoAt$ cu alți clasificatori de învățare automată este favorabilă. În plus, formularea modelului $AutoAt$ permite calcularea probabilității apartenenței unei instanțe de test la o anumită clasă de autor, ceea ce poate fi o proprietate utilă într-o varietate de aplicații de identificare a autorului unui text.

2. Secțiunea 3.2 prezintă un model de clasificare supervizat, denumit $AutoSoft$, pentru determinarea programatorului care a scris un fragment de cod. Acest model a fost introdus ca dovadă a fezabiltății conceptului și este compus dintr-un ansamblu de autoencoderi care sunt antrenați să codifice și să recunoască stilul de programare al dezvoltatorilor de software luați în considerare. Reprezentarea vectorială a documentelor utilizate este `doc2vec`, iar experimentele sunt efectuate pe subseturi ale setului de date Google Code Jam. Rezultatele obținute demonstrează în mod empiric că *autoencoderii* reprezintă o tehnică care reușește să surprindă elemente relevante referitoare la modul în care dezvoltatorii software își scriu codul. De asemenea, în acest studiu este propusă o extensie a lui $AutoSoft$, denumită $AutoSoft^{ext}$, pentru a recunoaște nu numai clasele autorilor (programatorilor) originali, pe care a fost antrenat modelul $AutoSoft$, ci și o clasă de „autor necunoscut”. Acest studiu a fost rezumat într-o lucrare originală care este

în prezent în curs de recenzare [CLB22].

3. Secțiunea 3.3 prezintă $SoftId$, un model de clasificare cu o singură clasă bazat pe un *autoencoder* care este antrenat să recunoască stilul de programare al unui anumit set de dezvoltatori software. În acest studiu, se folosesc reprezentări textuale pentru codul sursă Python, iar experimentele sunt efectuate pe subseturi de 3, 5 și 12 programatori din setul de date Google Code Jam. Performanța clasificatorului propus a fost evaluată folosind o serie de metrici care au demonstrat că $SoftId$ rezolvă cu succes task-ul de identificare a autorului, depășind un clasificator cu o singură clasă din literatură într-o majoritate covârșitoare a configurațiilor de testare, cu privire la toate metricile de performanță, și în special la $Specificitate$. Acest studiu a fost rezumat într-o lucrare originală care a fost acceptată spre publicare [LBCC22].

3.1 Identificarea autorului unor texte literare

Acest studiu propune $AutoAt$, un model de clasificare profund bazat pe *autoencoder*, care exploatează capacitatea *autoencoderilor* de a codifica tipare importante, pentru a rezolva un task de *identificare a autorului unui text*. Experimentele sunt efectuate pe un set de date de 1571 de poezii scrise de 8 poeți români folosind o reprezentare distribuită a documentelor, numită `doc2vec`. Abordarea propusă obține rezultate comparabile sau mai bune în raport cu alți clasificatori de învățare automată. În plus, formularea modelului $AutoAt$ permite calcularea probabilității ca o instanță de test să aparțină unei anumite clase de autor, ceea ce poate fi o proprietate utilă într-o varietate de aplicații care vizează identificarea autorului unui text. Acest aspect și faptul că pentru $AutoAt$ se obțin rezultate bune în task-ul dificil de determinare a autorului unui text poetic fără să fie introduse informații de natură literară demonstrează faptul că modelul propus este unul general, cu potențial de a fi utilizat cu succes în alte domenii.

Cele trei întrebări de cercetare la care își propune să răspundă acest studiu sunt:

RQ1 Cum se poate introduce un model de clasificare cu mai multe clase bazat pe un ansamblu de *autoencoderi* profunzi pentru a identifica în mod supervizat autorul unui text dat, pe baza relațiilor structurale și conceptuale codificate din documentele scrise de același autor?

RQ2 Care este performanța abordării introduse pentru a răspunde la RQ1 pentru identificarea unor poeți români din texte lirice date și cum se compară aceasta cu performanța modelelor similare de clasificare?

RQ3 Care este relevanța reprezentării distribuite a textelor poetice în identificarea diferiților autori?

3.2 Identificarea autorului unor fragmente de cod sursă. Modelul $AutoSoft$

Acest studiu introduce un model de clasificare supervizat, denumit $AutoSoft$, pentru identificarea programatorului care a scris un fragment de cod sursă. Modelul este introdus ca dovadă a fezabilității conceptului și este compus dintr-un ansamblu de autoencoderi care sunt antrenați să codifice și să recunoască stilul de programare al dezvoltatorilor de software. Ulterior, $AutoSoft$ este capabil să prezică autorul unui anumit fragment de cod sursă pe baza asemănării dintre codul dat și informațiile codificate (prin *autoencoderi*) despre fiecare dezvoltator de software. În acest mod, este exploatată capacitatea *autoencoderilor* de a codifica, prin reprezentările latente pe care le generează, tipare referitoare la stilul de programare al unor dezvoltatori de software specifici. Reprezentarea

programelor software este inspirată din domeniul procesării limbajului natural (NLP) [TCMM09]. Un program, procesat ca text (o secvență de tokeni specifici), este reprezentat ca un vector distribuit, vector care este furnizat de un model `doc2vec` [LM14]. Experimentele sunt efectuate pe programe software colectate de la un concurs internațional de programare organizat de Google și utilizate anterior în literatura de specialitate cu privire la identificarea autorului programelor software. Rezultatele obținute demonstrează în mod empiric faptul că tehnicile care implică *autoencoderi* au capacitatea de a surprinde, din perspectivă computațională, informații relevante despre modul în care programatorii își scriu codul. Un punct forte al modelului $AutoSoft$ este faptul că poate fi extins pentru a recunoaște nu numai clasele care corespund autorilor (programatorilor) originali pe care a fost antrenat, ci și o clasă de „autor necunoscut". Examinarea literaturii existente indică faptul că abordarea propusă în acest studiu este nouă în literatura de specialitate în ceea ce privește identificarea programatorului care a scris un fragment de cod sursă.

Pentru a face un sumar, studiul se concentrează pe a răspunde la următoarele întrebări:

RQ1 Cum se proiectează un clasificator supervizat bazat pe un ansamblu de *autoencoderi* pentru predicția dezvoltatorului de software care este probabil să fi scris un anumit fragment de cod sursă, ținând cont de stilul de programare specific pentru fiecare din programatorii luați în considerare?

RQ2 Îmbunătățește clasificatorul propus performanța sistemelor de identificare a autorului unui fragment de cod sursă în comparație cu clasificatorii convenționali din literatura de învățare automată?

RQ3 Ar putea un astfel de model de clasificare să fie folosit nu numai pentru a recunoaște clasele de dezvoltatori pe care a fost antrenat, ci și pentru a detecta o clasă de "autor necunoscut"?

3.3 Identificarea autorului unor fragmente de cod sursă. Modelul $SoftId$

Acest studiu introduce un model de clasificare cu o singură clasă, denumit $SoftId$, pentru identificarea autorului unui fragment de cod sursă. Modelul propus este compus dintr-un *autoencoder*, care este antrenat să codifice și să recunoască stilul de programare al unui anumit set de dezvoltatori de software. Modelul $SoftId$ astfel antrenat va putea detecta dacă un anumit cod sursă este creat de un dezvoltator din setul original sau de un dezvoltator de software „necunoscut" (altul decât cei din setul original). La nivel decizional, este calculată probabilitatea ca fragmentul de cod sursă luat ca instanță de test să fi fost scris de un autor din setul original pe baza similarității dintre instanța de testare și informația codificată de *autoencoder*. În propunerea noastră, fragmentele de cod sursă sunt reprezentate folosind tehnici de reprezentare a textului din domeniul NLP [TCMM09]. Noutatea clasificatorului propus constă în rezolvarea task-ului de identificare a autorului unui program software într-o configurație *open-set* folosind un autoencoder profund și bazat pe reprezentări textuale ale codurilor sursă. Examinarea literaturii existente indică faptul că modelul $SoftId$ este nou în literatura de specialitate în ceea ce privește identificarea autorului unui fragment de cod sursă.

Scopul cercetării prezentate în acest studiu este de a răspunde la următoarele întrebări:

RQ1 Cum trebuie proiectat un clasificator cu o singură clasă bazat pe un *autoencoder* pentru a rezolva problema identificării autorului unui fragment de cod sursă ca o problemă de recunoaștere a autorului în configurație *open-set*?

RQ2 Care este relevanța reprezentării textuale a fragmentelor de cod sursă în diferențierea între dezvoltatorii de software originali (cunoscuți) și alții (necunoscuți)?

RQ3 Care dintre cele două reprezentări bazate pe corpus, *Term Frequency - Inverse Document Frequency (TF-IDF)* sau *Latent Semantic Indexing (LSI)*, este mai potrivită pentru abordarea noastră?

Concluzii

Domeniul prelucrării limbajului natural devine din ce în ce mai important pe măsură ce cantitatea de date textuale disponibile electronic crește, atât online, sub formă texte postate pe rețele de socializare și arhive de documente pentru servicii de extragere a informațiilor, cât și în baze de date private precum cele formate din dosare medicale sau documente legale salvate sub formă digitală. În fiecare task din acest domeniu, principala provocare se referă la găsirea unui model eficient de limbaj pe care calculatorul îl poate manipula pentru a rezolva task-urile îndeplinite de obicei de oameni. Numeroase abordări, de la definiții bazate pe reguli ale limbajului la modele probabilistice ale limbajului, au fost propuse și evaluate într-o varietate de task-uri.

În această teză ne-am concentrat pe două task-uri particulare de procesare a limbajului natural, și anume *detectarea emoțiilor în text* și *identificarea autorului unui text* dat. În cazul *detectării emoțiilor exprimate în text*, s-a urmărit dezvoltarea unor resurse care să sprijine elaborarea unei soluții pentru acest task pentru limba română și evaluarea calității și utilității acestor resurse în studiile privind asocierile dintre conținutul semantic și emoțional al cuvintelor și textelor. În ceea ce privește *identificarea autorului unui text dat*, s-au propus soluții generale și flexibile bazate pe învățarea profundă pentru a determina autorii unor texte literare și a unor fragmente de cod sursă.

Studiile originale propuse au fost împărțite în două categorii corespunzătoare acestor două task-uri: în primul rând, au fost prezentate resursele de detectare a emoțiilor dezvoltate pentru limba română, iar apoi au fost detaliate analizele nesupervizate ale conținutului semantico-emoțional în care au fost utilizate aceste resurse; în partea a doua a tezei, au fost prezentate trei modele de identificare a autorului unui text, unul evaluat pe texte literare și două pe seturi de date care conțin cod sursă.

În ceea ce privește detecția emoțiilor în text, în primul rând, a fost prezentat stadiul actual al celor trei resurse dezvoltate pentru recunoașterea și analiza emoțiilor în texte în limba română: RoEmoLex, RoWikiLit și RoEmoData.

RoEmoLex este un lexicon care conține 9177 de termeni adnotați cu cele opt emoții primare identificate de Robert Plutchik (*Anticipare, Bucurie, Dezgust, Frică, Furie, Încredere, Surpriză, Tristețe*) și două etichete de polaritate (*Pozitivitate, Negativitate*). Acest lexicon a trecut prin trei etape de dezvoltare, pornind de la o traducere a unui lexicon în limba engleză, iar contribuțiile noastre au constat în corectarea unor traduceri automate, alinierea cu alte resurse pentru limba română, precum RoWordNet, revizuirea și re-adnotarea etichetelor emoționale și de valență, și îmbogățirea lexiculonului cu termeni noi. Versiunea finală a RoEmoLex, RoEmoLex v.3., furnizează informații despre conținutul emoțional și de polaritate al termenilor, despre partea de vorbire aferentă termenului, precum și id-ul synset-ului RoWordNet corespondent pentru 8286 de intrări. Utilitatea lexiconului propus a fost evidențiată atât în studiile noastre asupra conținutului emoțional prezent în opere literare, cât și în munca altor cercetători, care au folosit RoEmoLex în studii din domenii precum științe politice și psihologie.

RoWikiLit este un corpus care conține texte literare ale unor autori români reprezentativi. Acest corpus include include diverse tipuri de text literar în limba română, cum ar fi nuvele, romane, poezie, eseuri sau basme. Acest set de date a fost alcătuit inițial pentru a ajuta la studiul conținutului

emoțional în texte în limba română, deoarece datele erau disponibile online, în mod gratuit, și avea avantajul de a avea asociate analize literare care ar putea ajuta la interpretarea rezultatelor obținute în ceea ce privește conținutul semantic și emoțional. Cu toate acestea, setul de date s-a dovedit a fi folositor și în alte task-uri de procesare a limbajului natural, cum ar fi identificarea autorului unui text.

În ceea ce privește resursele, ultima dintre acestea este **RoEmoData**, un corpus de propoziții și fraze în limba română adnotat din punct de vedere emoțional. Intrările din acest set de date au fost adnotate cu privire la 16 categorii de emoții și valențe prin *crowdsourcing*. Analiza fiabilității adnotărilor a arătat că în urma unor modificări, cum ar fi eliminarea anumitor etichete de emoții, setul de date este unul potrivit pentru a fi utilizat în task-urile de detectare a emoțiilor din text. Printr-o investigație asupra structurii setului de date, au apărut observații interesante cu privire la prezența și intensitatea emoțională, precum și la relațiile dintre etichetele emoționale care se referă la emoții primare și cele care se referă la emoții secundare. Această analiză va informa viitoarele experimente de detectare a emoțiilor, dezvoltarea cărora este obiectivul nostru actual.

În al doilea rând, în același domeniu al analizei emoțiilor, am detaliat abordările bazate pe învățarea nesupervizată utilizate în studiile noastre.

Primul dintre aceste studii s-a referit la analiza termenilor din RoEmoLex folosind FCA, care a condus la descoperirea unor grupuri semnificative de cuvinte în ceea ce privește conținutul emoțional (`EmoSynsets`) și la propunerea unei formule pentru calcularea unor scoruri relative pentru termenii RoEmoLex, scoruri care denotă intensitatea emoției exprimate.

Al doilea studiu a reprezentat o analiză computațională a unui corpus reprezentativ de poezie românească, opera poetică a lui Mihai Eminescu. Au fost examinate o serie de caracteristici care abordează aspecte cu privire la vocabularul utilizat, complexitatea limbajului și conținutul emoțional în fiecare fază a creației artistice. Rezultatele au arătat că, pentru o serie de măsuri, observațiile teoretice din domeniul criticii literare găsesc corespondență în analiza cantitativă, ceea ce indică faptul că această abordare și vizualizarea rezultatelor, în special, ar putea fi folosite ca instrument de sprijin pentru interpretarea operelor literare.

Al treilea studiu s-a concentrat pe o clusterizare ierarhică bazată pe conținutul emoțional al unor poezii în limba română, care a explorat asociațiile semantico-emoționale într-un subset de 131 de opere ale poetului Mihai Eminescu. Rezultatele au arătat că există o suprapunere parțială între conținutul afectiv și tematic, în concordanță cu evaluările literare ale acelorași opere. Mai mult, abordările computaționale au avantajul de a fi obiective și pot fi reproduse, cu tehnici nesupervizate precum HAC reprezentând un instrument valoros în explorarea operelor literare. Cu toate acestea, niciun tipar emoțional specific nu a putut fi pe deplin asociat cu anumite teme literare.

În ceea ce privește identificarea autorului din text, au fost propuse trei modele bazate pe învățarea profundă.

În primul rând, a fost propusă o tehnică auto-supervizată pentru identificarea autorului unor texte lirice. Pentru a răspunde acestui task, a fost propus ***AutoAt***, un model de clasificare profund bazat pe *autoencoder* care exploatează capacitatea *autoencoderilor* de a codifica tiparele semnificative din datele de intrare. Au fost efectuate experimente pe un set de date de 1571 de poezii scrise de 8 poeți români folosind o reprezentare distribuită a documentelor. Abordarea propusă a obținut rezultate comparabile sau mai bune în raport cu alți clasificatori de învățare automată. În plus, formularea modelului $AutoAt$ a permis calcularea probabilității ca o instanță de test să aparțină unei anumite clase de autor, ceea ce poate fi o proprietate utilă într-o varietate de aplicații de identificare a autorului. Acest aspect și faptul că $AutoAt$ a obținut rezultate bune în task-ul cu dificultate ridicată de identificare a autorului unor poezii fără ca informații specifice domeniului să fie introduse au arătat că modelul propus este unul general, cu potențial de a fi utilizat cu succes și în alte domenii.

În al doilea rând, am examinat performanța unei abordări similare pe un set de date de programe

software pentru a rezolva sarcina de identificare a autorului unor fragmente de cod. A fost propus modelul de clasificare $\boldsymbol{AutoSoft}$. Reprezentarea programelor software a fost inspirată din domeniul de procesare a limbajului natural, iar folosind această reprezentare, *autoencoderii* profunzi aplicați pe reprezentările distribuite ale documentelor (`doc2vec`) s-a dovedit că aceștia reușesc să codifice caracteristici latente relevante ale fragmentelor de cod sursă, care au dus la diferențierea cu succes între autori/programatori. Mai mult, o extensie a clasificatorului $AutoSoft$, $\boldsymbol{AutoSoft^{ext}}$ a fost propusă pentru a identifica instanțe aparținând unui „autor necunoscut". Pentru această extensie s-au obținut rezultate bune în comparație cu abordările existente de clasificare cu o singură clasă.

În cele din urmă, a fost propus modelul $\boldsymbol{SoftId}$ cu scopul de a rezolva problema de identificare a autorului software într-o configurație *open-set*, având o clasă originală, cunoscută, și posibilitatea de clasificare a unei instanțe de test ca aparținând unui autor necunoscut. $SoftId$ reprezintă un clasificator cu o singură clasă bazat pe un *autoencoder*, care a fost antrenat să învețe și să codifice stilul de programare al unui set dat ("original") de programatori. La momentul testării, $SoftId$ va distinge între codul scris de programatori din setul original și cel scris de alți programatori, pe baza asemănării dintre instanța testată și informațiile codificate de *autoencoder*. Experimentele efectuate pe un subset de programe Python [Goo] au demonstrat că $SoftId$ rezolvă cu succes task-ul de identificare a autorului în acest context, depășind clasificatorul *One-class SVM* într-o majoritate covârșitoare a configurațiilor de testare. Experimentele efectuate au evidențiat, de asemenea, relevanța reprezentărilor textuale a codurilor sursă utilizate de $SoftId$ pentru a distinge autorul unui fragment de cod ("cunoscut'' sau "necunoscut").

Direcții viitoare de cercetare

Continuarea cercetării pentru cele două direcții prezentate în această teză poate fi abordată în mai multe moduri.

În ceea ce privește dezvoltarea resurselor pentru detectarea emoțiilor din texte în limba română, obiectivul nostru principal este să finalizăm adnotarea datelor din categoriile rămase din setul de date **RoEmoData**. Pentru seturile de date adnotate ale RoEmoData, sunt propuse următoarele etape pentru a obține o resursă utilă și fiabilă: (1) completarea setului de date cu propoziții și fraze în care să nu fie exprimate emoții (aparținând unei clase *Neutru/Nicio emoție*), fie prin determinarea unui prag adecvat pentru scorurile emoționale pe seturile date adnotate și astfel luând în considerare unele instanțe RoEmoData drept *neutre*, sau prin adăugarea unor astfel de instanțe dintr-o sursă externă; (2) eliminarea etichetelor emoționale ambigue (de exemplu, *Anticipare*) care au o corelație a adnotării foarte scăzută între evaluatori și (3) definirea unei metode adecvate de agregare a scorurilor adnotatorilor.

Odată finalizați acești pași, ne propunem să proiectăm modele de învățare automată și învățare profundă folosind datele adnotate existente pentru a rezolva task-ul de recunoaștere a emoțiilor în texte în limba română. În timp ce seturile de date etichetate sunt baza abordărilor supervizate pe care le avem în vedere, intenționăm să folosim și RoEmoLex pentru a crea *reprezentări distribuite îmbogățite cu conținut emoțional* pentru texte. În ceea ce privește analiza nesupervizată a conținutului semantico-emoțional, o considerăm un pas esențial în înțelegerea resurselor și seturilor de date cu care lucrăm. Ca direcții viitoare de cercetare, propunem explorarea de sine stătătoare a emoției în diferite tipuri de texte literare din corpusul RoWikiLit și definirea unor măsuri mai fine care surprind nuanțele expresiei în astfel de texte, cum ar fi evoluția în timp a conținutului emoțional în fragmente de text.

În ceea ce privește identificarea autorului unui text, pentru identificarea autorilor în textele literare, propunem o combinație a trăsăturilor implicite (ascunse) extrase de `doc2vec` și a trăsăturilor stilistice specifice textelor literare (de exemplu, trăsături care surprind fenomene fonetice în poezii: eufonie, asonanță, aliterație, rimă și proprietăți ale cuvintelor: distribuția frecvenței, statistici privind

vocabularul folosit [PLTA15]), pentru reprezentarea textelor poetice.

În ceea ce privește $AutoSoft$, generalitatea clasificatorului propus, care se bazează pe *autoencoder*, a fost testată într-un task de atribuire a autorului codului sursă, pentru care sunt vizate următoarele extensii: rezolvarea problemei de identificare a co-autorilor prin adaptarea deciziei de clasificare a modelului $AutoSoft^{ext}$ și utilizarea reprezentării programelor software bazată pe o altă tehnică NLP, LSI [MM00], care este independentă de limbaj, în $AutoSoft$.

În ceea ce privește $SoftId$, este prevăzută realizarea de experimente pe seturi de date colectate de la echipe de dezvoltare de software din medii "reale" (vs. caracterul artificial al setului de date Google Code Jam) pentru evaluarea ulterioară a clasificatorului $SoftId$. Ca modalități suplimentare de a îmbunătăți $SoftId$, vor fi luate în considerare funcții alternative pentru calcularea distanței D dintre reprezentările vectoriale ale codului sursă, cum ar fi distanța euclidiană.

Bibliografie

[ARA+19] Mohammed Abuhamad, Ji-su Rhim, Tamer AbuHmed, Sana Ullah, Sanggil Kang, and DaeHun Nyang. Code authorship identification using convolutional neural networks. *Future Generation Computer Systems*, 95:104–115, 2019.

[BCL21] Anamaria Briciu, Gabriela Czibula, and Mihaiela Lupea. $AutoAt$: A deep autoencoder-based classification model for supervised authorship attribution. *Procedia Computer Science*, 192:119–128, 2021.

[Bir20] Daniel Biro. Emotions in the political discourse in Romania. A corpus-driven analysis of multiword expressions. *Bulletin of the Transilvania University of Braşov, Series IV: Philology & Cultural Studies*, 13(1):17–38, 2020.

[BKR+21] Egor Bogomolov, Vladimir Kovalenko, Yurii Rebryk, Alberto Bacchelli, and Timofey Bryksin. Authorship attribution of source code: A language-agnostic approach and applicability in software engineering. In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pages 932–944, 2021.

[BL17] Anamaria Briciu and Mihaiela Lupea. RoEmoLex - a Romanian Emotion Lexicon. *Studia Universitatis Babeș-Bolyai Informatica*, 62(2):45–56, 2017.

[BL18] Anamaria Briciu and Mihaiela Lupea. Studying the language of mental illness in romanian social media. In *2018 IEEE 14th International Conference on Intelligent Computer Communication and Processing (ICCP)*, pages 21–28. IEEE, 2018.

[Bri19] Anamaria Briciu. Quantitative analysis of style in Mihai Eminescu's poetry. *Studia Universitatis Babeș-Bolyai Informatica*, 64(2):80–95, 2019.

[BT07] Steven Burrows and Seyed M. M. Tahaghoghi. Source code authorship attribution using n-grams. In *Proceedings of the twelth Australasian document computing symposium, Melbourne, Australia, RMIT University*, pages 32–39, 2007.

[Bur09] Jill Burstein. Opportunities for natural language processing research in education. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, pages 6–27, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.

[CD21] Alexandra Ciobotaru and Liviu P. Dinu. RED: A novel dataset for Romanian emotion detection from tweets. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 291–300, Held Online, September 2021. INCOMA Ltd.

[Cho03] Gobinda G Chowdhury. Natural language processing. *Annual Review of Information Science and Technology (ARIST)*, 37:51–89, 2003.

[CJ20] Kakia Chatsiou and Slava Jankin. Deep learning for political science. *The SAGE Handbook of Research Methods in Political Science and International Relations*, 2020.

[CLB22] Gabriela Czibula, Mihaiela Lupea, and Anamaria Briciu. Enhancing the performance of software authorship attribution using an ensemble of deep autoencoders. *Mathematics, Special issue on Recent Advances in Artificial Intelligence and Machine Learning*, submitted, 2022.

[DFIC19] Steven HH Ding, Benjamin Fung, Farkhund Iqbal, and William K Cheung. Learning Stylometric Representations for Authorship Analysis. *IEEE Transactions on Cybernetics*, 49(1):107 – 121, 2019.

[FH+99] Carol Friedman, George Hripcsak, et al. Natural language processing and its future in medicine. *Acad Med*, 74(8):890–5, 1999.

[FRC13] Carol Friedman, Thomas C Rindflesch, and Milton Corn. Natural language processing: State of the art and prospects for significant progress. *Journal of biomedical informatics*, 46(5):765–773, 2013.

[GDKS20] Aaryan Gupta, Vinya Dengre, Hamza Abubakar Kheruwala, and Manan Shah. Comprehensive review of text-mining applications in finance. *Financial Innovation*, 6(1):1–25, 2020.

[Goo] Google. Google Code Jam Competition. https://codingcompetitions.withgoogle.com/codejam. Online; accessed 15 September 2021.

[GW97] Bernhard Ganter and Rudolf Wille. *Formal Concept Analysis: Mathematical Foundations*. Springer-Verlag, Berlin, Heidelberg, 1st edition, 1997.

[HLLA14] N.D. Hansen, C. Lioma, B. Larsen, and S. Alstrup. Temporal Context for Authorship Attribution. A Study of Danish Secondary Schools. *Multidisciplinary Information Retrieval. IRFC 2014. Lecture Notes in Computer Science*, 8849:22 – 40, 2014.

[HO20] Andreea Horoiţą and Adrian Opre. False memories: Romanian Deese-Roediger-Mcdermott lists of words. *Cognition, Brain, Behavior*, 24(2):163–186, 2020.

[Joc13] Matthew L Jockers. *Macroanalysis: Digital methods and literary history*. University of Illinois Press, 2013.

[Juo06] Patrick Juola. Authorship attribution. *Information Retrieval*, 1(3):233–334, 2006.

[KKG+19] Vaibhavi Kalgutkar, Ratinder Kaur, Hugo Gonzalez, Natalia Stakhanova, and Alina Matyukhina. Code Authorship Attribution: Methods and challenges. *ACM Computing Surveys (CSUR)*, 52(1):1 – 36, 2019.

[Kle] Dan Klein. CS 294-5: Statistical Natural Language Processing. https://people.eecs.berkeley.edu/~klein/cs294-5/index.html. Online; accessed 29 April 2022.

[LB17] Mihaiela Lupea and Anamaria Briciu. Formal Concept Analysis of a Romanian Emotion Lexicon. In *13th IEEE International Conference on Intelligent Computer Communication and Processing (ICCP 2017)*, pages 111–118. IEEE, 2017.

[LB19] Mihaiela Lupea and Anamaria Briciu. Studying emotions in Romanian words using Formal Concept Analysis. *Computer Speech & Language*, 57:128–145, 2019.

[LBB21] Mihaiela Lupea, Anamaria Briciu, and Elena Bostenaru. Emotion-based Hierarchical Clustering of Romanian Poetry. *Studies in Informatics and Control*, 30(1):109–118, 2021.

[LBCC22] Mihaiela Lupea, Anamaria Briciu, Istvan Gergely Czibula, and Gabriela Czibula. $SoftId$: An autoencoder-based one-class classification model for software authorship identification. In *Proceedings of the 26th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2022)*, page submitted, 2022.

[LM14] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning, Beijing, China, 2014*, volume 32, pages 1188–1196, 2014.

[MGAPD+17] Ilia Markov, Helena Gómez-Adorno, Juan-Pablo Posadas-Durán, Grigori Sidorov, and Alexander Gelbukh. Author profiling with doc2vec neural network-based document embeddings. In Obdulia Pichardo-Lagunas and Sabino Miranda-Jiménez, editors, *Advances in Soft Computing*, pages 117–131, Cham, 2017. Springer International Publishing.

[MM00] Jonathan I. Maletic and Andrian Marcus. Using Latent Semantic Analysis to identify similarities in source code to support program understanding. In *Proceedings 12th IEEE internationals conference on tools with artificial intelligence. ICTAI 2000*, pages 46–53. IEEE, 2000.

[MT10] Saif Mohammad and Peter Turney. Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, CAAGET '10, pages 26–34, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[Öhm20] Emily Öhman. Emotion annotation: Rethinking emotion categorization. *CEUR Workshop Proceedings*, 2865:134–144, 2020.

[PLTA15] Ioan-Iovitz Popescu, Mihaiela Lupea, Doina Tătar, and Gabriel Altmann. *Quantitative Analysis of Poetic Texts*. De Gruyter Mouton, 2015.

[Plu80] Robert Plutchik. A general psychoevolutionary theory of emotion. *Emotion: Theory, Research, and Experience*, 1:3–33, 1980.

[PT18] Vasile Păiș and Dan Tufiș. Computing distributed representations of words using the CoRoLa corpus. *Proceedings of the Romanian Academy*, 19(2):403–409, 2018.

[SAM96] Philip Sallis, Asbjorn Aakjaer, and Stephen MacDonell. Software forensics: Old methods for a new science. In *Proceedings 1996 International Conference Software Engineering: Education and Practice*, pages 481–485. IEEE, 1996.

[SMS17] S. Swain, G. Mishra, and C. Sindhu. Recent approaches on Authorship Attribution techniques — An overview. In *2017 International conference of Electronics, Communication and Aerospace Technology (ICECA)*, pages 557–566, 2017.

[STASH19] Sicong Shao, Cihan Tunc, Amany Al-Shawi, and Salim Hariri. One-class Classification with Deep Autoencoder Neural Networks for Author Verification in Internet Relay Chat. In *Proceedings of 16th IEEE/ACS International Conference on Computer Systems and Applications*, pages 1–8, 2019.

[STZ18] Armin Seyeditabari, Narges Tabari, and Wlodek Zadrozny. Emotion detection in text: a review. *CoRR*, abs/1806.00674, 2018.

[TBM14] Dan Tufiș and Verginica Barbu Mititelu. *The Lexical Ontology for Romanian*, volume 48 of *Text, Speech and Language Technology*, pages 491–504. Springer, 2014.

[TCMM09] Doina Tătar, Gabriela Serban Czibula, Andreea Diana Mihis, and Rada Mihalcea. Textual entailment as a directional relation. *Journal of Research and Practice in Information Technology*, 41(1):53–64, 2009.

[VHBT+05] Hans Van Halteren, Harald Baayen, Fiona Tweedie, Marco Haverkort, and Anneke Neijt. New machine learning methods demonstrate the existence of a human stylome. *Journal of Quantitative Linguistics*, 12(1):65–77, 2005.

[Wil05] Rudolf Wille. Formal Concept Analysis as Mathematical Theory of Concepts and Concept Hierarchies. In Bernhard Ganter, Gerd Stumme, and Rudolf Wille, editors, *Formal Concept Analysis*, Lecture Notes in Computer Science, vol 3626, pages 1–33. Springer-Verlag, Berlin, Heidelberg, 2005.