UNIVERSITATEA POLITEHNICA DIN BUCUREȘTI

Facultatea de Automatică și Calculatoare

TEZĂ DE DOCTORAT

Evaluarea scrierii și a performanțelor studenților folosind tehnici de prelucrare a limbajului natural și un cadru dialogic

Autor: Ing. Maria-Dorinela Dascălu (Sîrbu)

Conducător de doctorat: Prof.dr.ing. Ștefan Trăușan-Matu

COMITETUL DE DOCTORAT

Președinte	Prof.dr.ing. Florin Pop	de la	Universitatea Politehnica din București, România
Conducător de doctorat	Prof.dr.ing. Ștefan Trăușan-Matu	de la	Universitatea Politehnica din București, România
Membru	Conf.dr.ing. Traian Rebedea	de la	Universitatea Politehnica din București, România
Membru	Prof.dr. Philippe Dessus	de la	Universitatea Grenoble Alpes, Franța
Membru	Prof.dr. Nicolae Nistor	de la	Universitatea Ludwig Maximilian din Munchen, Germania

Rezumat

Prelucrarea Limbajului Natural (NLP) este o zonă de interes în cercetare în domeniul informaticii, inteligenței artificiale si lingvistică care explorează modul în care calculatoarele pot fi utilizate pentru a înțelege, manipula și interpreta textul sau vorbirea în limbaj natural. NLP include tehnici pentru modelarea și interpretarea limbajului uman și variază de la metode statistice, învățare automată, învățare profundă, până la abordări bazate pe reguli. Această teză se concentrează pe aplicații de prelucrare a limbajului natural, mai precis pe evaluarea scrierii, învățarea colaborativă și analiza computațională a discursului. Tehnici avansate de prelucrare a limbajului natural și învățare automată sunt aplicate pentru a analiza stilul de scriere, pentru a evalua performanțele online ale studenților și pentru a explora dialogismul ca model de discurs folosind modele de limbă.

Întrebările de cercetare ale acestei teze acoperă: (1) În ce măsură indicii de complexitate textuală sunt predictivi din prisma diferențelor în stilul de scriere atunci când se efectuează analize multilingve pe diferite corpuri și tipuri de discurs (spre exemplu, documente oficiale NATO și mass-media din România)? (2) În ce măsură feature-urile generate de platforma ReaderBench, inclusiv feature-uri derivate din graful CNA aplicat pe discuțiile de tip forum, indicii de complexitate textuală, analiza longitudinală și feature-urile derivate din datele click-stream, sunt capabile să prezică performanța studenților? (3) Care sunt diferențele dintre comportamentele și interacțiunile studenților înainte și în timpul pandemiei COVID-19 în două instanțe anuale consecutive ale unui curs universitar? (4) În ce măsură lanțurile semantice pot fi operaționalizate folosind modele de limbă predictive atât din prisma colaborării (spre exemplu, evaluarea implicării participanților în cadrul unei conversații), cât și în studiul individual (spre exemplu, evaluarea automată a eseurilor)? Pentru a răspunde la toate aceste întrebări de cercetare, această teză introduce o serie de studii experimentale.

În primul rând, această teză analizează stilul de scriere în ceea ce privește caracteristicile lingvistice, luând în considerare discursurile mass-media din România despre criza economică dintre anii 2008 și 2018, precum și discursurile militare ale documentelor oficiale NATO. Au fost efectuate analize statistice pentru a examina diferențele în stilurile de scriere, o analiză multivariată a variației urmată de o analiză discriminantă în trepte.

În al doilea rând, această teză evaluează activitatea online a studenților prin analiza comportamentelor acestora, modelarea interacțiunii și prezicerea notelor pe baza participării lor online. Caracteristicile globale (spre exemplu, indici de participare, indici de inițiere), analiza seriilor de timp și datele fluxului de clicuri sunt folosite pentru a prezice notele și succesul studenților. Au fost generate diverse sociograme pentru a modela tendințele participării studenților și pentru a vedea tiparele de interacțiune între participanți, împreună cu sociograme săptămânale pentru a urmări modul în care comunitatea evoluează de la o săptămână la alta. Modelele liniare au indicat faptul că succesul este legat de numărul de zilele petrecute pe forum și postarea în mod regulat. Comportamentele și interacțiunile studenților sunt comparate înainte și în timpul pandemiei COVID-19 utilizând două instanțe anuale consecutive ale unui curs universitar. Pandemia COVID-19 a generat o dezechilibrare, o creștere drastică a participării, urmată de o scădere spre sfârșitul semestrului, comparativ cu anul universitar anterior, când s-au observat fluctuații mai mici de participare.

În al treilea rând, această teză introduce o metodă bazată pe dialogism care identifică legăturile semantice folosind doar scorurile de atenție calculate folosind un model de limbă BERT. Deși diferențele de performanță nu sunt substanțiale, modelul are unele avantaje importante în comparație cu alte metode: lanțurile calculate de BERT sunt dependente de context, modelul se generalizează în relații multiple și detectează alte relații care au scoruri de atenție similare, chiar dacă modelele de predicție au fost antrenate pe reguli simple. Modelul a fost aplicat pe două sarcini: notarea automată a eseurilor și evaluarea implicării studenților în conversațiile de tip chat. Rezultatele au arătat că raportul lanțurilor, care indică acoperirea lanțurilor semantice utilizate de un participant în raport cu întreaga conversație, este pe departe cea mai predictivă caracteristică pentru participare. Mai mult,

continuările, care reflectă legăturile dintre doi participanți diferiți, împreună cu numărul și raporturile lanțurilor acoperite de participant, indică calitatea colaborării.

Cuprins

Listă figuri

1 Introducere

1.1 Prezentare generală

Prelucrarea Limbajului Natural (eng., Natural Language Porcessing; NLP) (Chowdhury, 2003; Jurafsky & Martin, 2008) și învățarea automată (eng., Machine Learning; ML) (Jordan & Mitchell, 2015; Mohri, Rostamizadeh, & Talwalkar, 2018) câștigă o atenție sporită datorită volumului crescut de date care este generat și care trebuie procesat. NLP explorează modul în care calculatoarele pot fi utilizate pentru a înțelege, manipula și interpreta limbajul natural. Sarcinile și aplicațiile NLP variază foarte mult, spre exemplu: răspuns automat la întrebări, generarea de text și întrebări, analiza sentimentelor, detectarea știrilor false, înțelegerea dialogului, rezumarea textului, înțelegerea citirii, extragerea cuvintelor cheie sau modelarea topicelor. Această teză prezintă o gamă largă de experimente bazate pe tehnici NLP ca domeniu de analiză, cu accent pe modelarea stilului de scriere, evaluarea performanței studenților în mediile de învățare online și explorarea dialogismului folosind modele de limbă.

Scrisul este o abilitate centrală de învățare care este strâns legată de înțelegerea textului. Abilitățile bune de scriere se dobândesc prin practică și se caracterizează printr-un limbaj clar și organizat, o utilizare corectă a gramaticii, o coeziune puternică a textului și o formulare sofisticată.

Mediile online de învățare colaborativă deschid noi oportunități de cercetare, spre exemplu, analiza rezultatelor învățării, identificarea tiparelor de învățare, predicția comportamentului elevilor, precum și modelarea și vizualizarea relațiilor sociale corelat cu tendințelor în rândul elevilor. Diverse platforme de învățare online sunt disponibile gratuit pentru a ajuta studenții în procesul de învățare. Cele mai cunoscute sunt Moodle (Moodle, 2021) și Massive Open Online Course (MOOC; McAuley, Stewart, Siemens, & Cormier, 2010). Moodle este o platformă educațională online concepută pentru a spori atât activitatea elevilor, cât și a profesorilor, și poate fi folosită eficient pentru a sprijini învățarea colaborativă. Moodle este adesea folosit pentru a pune întrebări cu privire la temele studenților, examene, pentru a solicita clarificări și pentru a face anunțuri. MOOC-urile au devenit o platformă importantă pentru predare și învățare datorită capacității lor de a oferi accesibilitate educațională oricând de oriunde. Evaluarea succesului și implicării elevilor în mediile de învățare online este o sarcină dificilă și care necesită mult timp pentru profesori. Un instrument de date care s-a dovedit eficient în explorarea succesului studenților la cursurile online a fost Cohesion Network Analysis (CNA; M. Dascalu, Trausan-Matu, McNamara, & Dessus, 2015), care oferă posibilitatea de analiză a structurii discursului în medii de învățare colaborativă și facilitează identificarea tiparelor de interacțiune a cursanților. Aceste tipare pot fi utilizate pentru a prezice comportamentele elevilor, cum ar fi ratele de abandon și performanța.

Dialogismul este o teorie filozofică centrată pe ideea că viața implică un dialog între mai multe voci care sunt într-o interacțiune continuă. Având în vedere limbajul uman, diferite idei sau puncte de vedere iau forma unor voci, care se răspândesc în orice discurs și îl influențează. Din punct de vedere al calculului, vocile pot fi operaționalizate ca lanțuri semantice care conțin cuvinte înrudite.

1.2 Obiective și Interese

Pe baza contextului anterior și a multiplelor domenii de cercetare pe care le oferă, această teză se concentrează pe trei obiective principale, dintre care se desprind patru întrebări de cercetare. Fiecare obiectiv este detaliat într-un capitol dedicat din partea de studii empirice a acestei teze:

1. Modelarea și investigarea stilului de scriere în diverse texte folosind indici de complexitate textuală;
2. Evaluarea activității online a studenților în medii de învățare online pe baza postărilor lor (spre exemplu, cursuri Moodle, MOOC) prin analiza comportamentelor elevilor, modelarea interacțiunii lor și prezicerea notelor elevilor pe baza participării lor;
3. Introducerea unei metode noi de identificare a lanțurilor semantice folosind modele de limbă de ultimă generație pentru a evalua implicarea elevilor, pe baza scrierii lor în conversațiile de chat online.

Pornind de la cele trei obiective principale anterioare, această teză răspunde la următoarele întrebări de cercetare:

- **RQ1**: În ce măsură indicii de complexitate textuală sunt predictivi din prisma diferențelor în stilul de scriere atunci când se efectuează analize multilingve pe diferite corpuri și tipuri de discurs (spre exemplu, documente oficiale NATO și mass-media din România)?
- **RQ2**: În ce măsură feature-urile generate de platforma ReaderBench, inclusiv feature-uri derivate din graful CNA aplicat pe discuțiile de tip forum, indicii de complexitate textuală, analiza longitudinală și feature-urile derivate din datele click-stream, sunt capabile să prezică performanța studenților?
- **RQ3**: Care sunt diferențele dintre comportamentele și interacțiunile studenților înainte și în timpul pandemiei COVID-19 în două instanțe anuale consecutive ale unui curs universitar?
- **RQ4**: În ce măsură lanțurile semantice pot fi operaționalizate folosind modele de limbă predictive atât din prisma colaborării (spre exemplu, evaluarea implicării participanților în cadrul unei conversații), cât și în studiul individual (spre exemplu, evaluarea automată a eseurilor)?

Prima întrebare de cercetare face referire la primul obiectiv, a doua și a treia întrebare de cercetare sunt legate de cel de-al doilea obiectiv, în timp ce ultima întrebare de cercetare este derivată din al treilea obiectiv.

1.3 Structura tezei

Teza este structurată în trei părți principale, aspecte teoretice, studii empirice și discuții și concluzii. Capitolele din partea Aspecte teoretice susțin capitolele din partea Studii empirice (vezi Figura 1). Fiecare subcapitol din Studiile empirice are un capitol corespunzător cu aspecte teoretice, care include modele, metode și sisteme de ultimă generație.

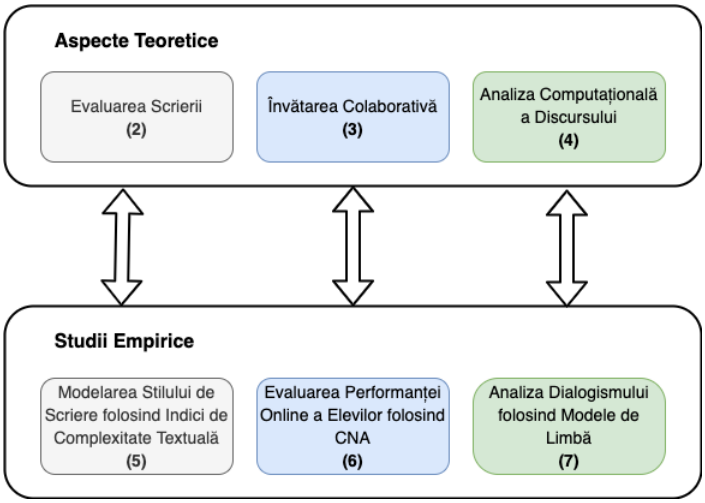


Figura 1. Structura tezei.

Teza începe cu aspecte teoretice și include modele, metode și sisteme de ultimă generație, relevante pentru studiile empirice. Această parte începe cu concepte de evaluare a scrierii (capitolul 2), definirea scrierii ca sarcină critică și descrierea diferitelor sisteme de evaluare automată a scrierii. Capitolul 2 continuă cu definiția dialogismului și a modelului polifonic și se încheie cu aspecte legate de stilometrie, inclusiv diferite sisteme de ultimă generație pentru calculul indicilor de complexitate textuală. Capitolul 3 introduce învățarea colaborativă și se concentrează pe mediile de învățare online care stimulează colaborarea (spre exemplu, sisteme de management a învățării, sisteme de gestionare a cursurilor, MOOC). Acest capitol introduce dialogismul drept cadru pentru mediile de învățare colaborativă computerizată (eng., Computer Supportive Collaborative Learning; CSCL), se încheie cu analiza rețelelor sociale (eng., Social Network Analysis; SNA) și cele mai comune metrici pentru a calcula participarea, colaborarea și importanța în comunitate. Capitolul 4 prezintă aspecte teoretice legate de analiza discursului computațional și include aspecte generale despre Prelucrarea Limbajului Natural, cu accent pe spaCy, modele semantice și modele de limbă de ultimă generație și analiza rețelei de coeziune (eng., Cohesion Network Analysis; CNA).

Teza continuă cu studii empirice și definește diverse studii de caz care se bazează pe conceptele teoretice descrise în prima parte, grupate în funcție de specificul lor. Fiecare capitol din această secțiune include descrierea experimentului, corpusul utilizat, metoda abordată, rezultatele obținute și discuții. Astfel, Capitolul 5 descrie două studii de caz care se concentrează pe modelarea stilului de scriere folosind indici de complexitate textuală. Capitolul 6 introduce două studii de caz axate pe evaluarea performanțelor studenților in mediile online prin analiza rețelei de coeziune, în timp ce capitolul 7 prezintă două studii de caz care analizează dialogismul folosind modele de limbă.

În continuare, capitolul *Discuții* descrie avantajele abordărilor noastre, problemele confruntate și soluțiile oferite, alături de implicațiile educaționale. Capitolul *Concluzii* prezintă contribuții personale și direcții pentru cercetări viitoare.

Partea 1 – Aspecte Teoretice

2 Evaluarea Scrierii

Scrierea este o activitate centrală de învățare care necesită practică și experiență (Light, 2004; McNamara, Crossley, & McCarthy, 2010; Witte & Faigley, 1981). Calitatea scrierii este un indicator al performanței. Un text bine scris necesită focus, consistență, coerență, și corectitudine (College, 2021). Mai mult, o scriere bună are o idee principală clară și fiecare paragraf are o singură propoziție principală sau o singură perspectivă principală. Accentul ar trebui să se concentreze pe o idee principală, care se răspândește în întreaga scriere. O scriere bună este consecventă, fiecare paragraf este legat de ideea principală, dar are propria sa perspectivă și nu o împărtășește cu un alt paragraf. Coerența este o calitate foarte importantă, deoarece ideile sau evenimentele ar trebui organizate logic. O scriere bună asigură un flux coeziv de informații și are sens pentru cititor. Calitatea scrierii este indicată și de corectitudine, deoarece o scriere corectă este scrisă corect, fără greșeli gramaticale, cuvintele sunt folosite corect, propozițiile sunt complete și au sens. Ideea principală este susținută și extinsă de fiecare paragraf. Ideile sunt descrise în detaliu, explicate și susținute de exemple.

Un text bine scris necesită și creativitate, în timp ce personalitatea autorului se regăsește în modul în care exprimă ideile și în combinația de cuvinte. Fiecare persoană are propriul său mod de a se exprima și un vocabular mai mult sau mai puțin complex. Un scriitor bun are claritate și disciplină în expresiile sale.

Un text bine scris necesită, de asemenea, ca textul să fie ușor de înțeles de publicul vizat. Lizibilitatea (Dale & Chall, 1949) decide cât de dificil sau ușor este un text de înțeles. Există diverse elemente care contribuie la lizibilitatea unui text: cuvintele, lungimea propozițiilor, structura propoziției, silabele medii pe cuvânt (DuBay, 2004). Aceste elemente combinate pot face un text greu de înțeles. O lizibilitate ridicată înseamnă că cititorul poate înțelege clar ideile textului, în timp ce ține cont de complexitatea conceptelor, poate procesa cu ușurință toate informațiile furnizate în text și reduce neînțelegerile.

2.1 Scrierea ca Sarcină Critică

Abilitățile bune de scriere se dobândesc prin practică și se caracterizează printr-un limbaj clar și organizat, o utilizare corectă a gramaticii, o coeziune puternică a textului și o formulare sofisticată. Furnizarea de feedback constructiv îi poate ajuta pe cursanți să-și îmbunătățească scrierea; cu toate acestea, furnizarea de feedback este un proces complex și care necesită mult timp.

Mulți studenți sunt interesați să își îmbunătățească abilitățile de scriere, deoarece calitatea scrierii este un aspect important în definirea capacităților intelectuale la aproape toate nivelurile academice. Furnizarea de feedback personalizat elevilor cu privire la capacitatea lor de a scrie este o componentă fundamentală în procesul de învățare. Cu toate acestea, mulți profesori nu au timp să ofere feedback într-un mod iterativ care susține cel mai bine învățarea elevilor. Ca rezultat, multe sisteme bazate pe calculator au fost dezvoltate pentru a sprijini studenții în procesul de scriere. Aceste sisteme sunt în general denumite sisteme de evaluare automată a scrierii (eng., Automated Writing Evaluation; AWE). Au fost dezvoltate diverse sisteme capabile să ofere instruire în strategia de scriere pentru a ajuta studenții în procesul de scriere (McNamara, Crossley, & Roscoe, 2013).

Potrivit lui Roscoe, Varner, Crossley, and McNamara (2013), există două categorii de sisteme de evaluare textuală care pot oferi feedback elevilor: sisteme automate de evaluare a eseurilor (eng., Automated Essay Scoring; AES) și sisteme automate de evaluare a scrierii (eng., Automated Writing Evaluation; AWE). Sistemele AES (Attali & Burstein, 2004) oferă un scor global, sumativ la un eseu, iar performanța lor poate fi evaluată în funcție de corelația dintre scorurile umane și cele automate. Sistemele AWE (Warschauer & Ware, 2006) sunt construite deasupra sistemelor AES, cu scopul de a oferi feedback constructiv care depășește scorul holistic (adică, formativ pe lângă scorul sumativ). Sistemele AWE sunt mai utile pentru dezvoltarea elevilor, deoarece feedbackul primit îi poate ajuta în ceea ce privește conștientizarea potențialelor probleme, precum și metodele de îmbunătățire. Ambele sisteme, AES și AWE, se pot baza pe indici de complexitate textuală care oferă informații cu privire la calitatea unui text sau la trăsăturile stilului de scriere, și pot măsura diferite aspecte, de la lizibilitate, metrici numerice la nivel de suprafață (cum ar fi numărul de paragrafe, numărul de silabe, entropia cuvintelor etc.), la coerența și coeziunea textului (Crossley, Kyle, & McNamara, 2015; McNamara, Crossley, Roscoe, Allen, & Dai, 2015).

Multe sisteme AWE au fost dezvoltate pentru a fi folosite in clasă cu studenții. O parte dintre aceste sisteme sunt disponibile gratuit (în majoritatea cazurilor, cele dezvoltate de cercetători), în timp ce sistemele dezvoltate de companii necesită de o taxă mică spre medie.

2.2 Dialogismul și Modelul Polifonic

Dialogismul este o teorie filozofică introdusă de Mihail Bakhtin (1981, 1984), centrată pe ideea că totul, chiar și viața, este dialogică, un schimb continuu și interacțiune între voci: „Viața prin natura sa este dialogică ... când dialogul se termină, totul se termină" (Bakhtin, 1984, pp. 293,252). Trausan-Matu, Stahl, and Sarmiento (2007) au extins conceptul de voce pentru analiza discursului, în general, și a învățării colaborative, în special. Autorii consideră că vocile sunt reprezentări generalizate ale diferitelor puncte de vedere sau idei, care se răspândesc în discurs și îl influențează. Există mai multe voci într-un dialog care susțin scopurile și interacțiunile comunicării, inter-animându-se prin divergențe și convergențe de puncte de vedere, oferind în același timp armonie discursului, într-un mod similar cu muzica polifonică (Trausan-Matu, 2020). Mergând mai larg, dialogismul poate fi perceput ca o teorie a unei lumi semnificative (Linell, 2009), care constă din idei, gânduri, acțiuni și procese semnificative.

Enunțurile, fie în vorbire, fie în text, sunt multivocale, tot ceea ce spunem sau scriem este umplut cu cuvintele și perspectivele altora și invers (Bakhtin, 1986). În acest sens, conceptul de dialog capătă alte conotații și poate fi extins pentru a include o gamă mai largă de activități lingvistice. Astfel, dialogismul și inter-animația vocilor sunt întâlnite în toate contextele, de la conversații în care mai mulți participanți discută despre diverse subiecte, la texte în general în care apar diferite poziții care aparțin anumitor autori - proza lui Dostoievski percepută ca fiind „o pluralitate de voci independente și neunite și conștiințe " (Bakhtin, 1984) - și chiar dialoguri interne între voci interioare care dezbat ideile și se contrazic reciproc (Marková, Linell, Grossen, & Salazar Orvig, 2007). Multivocalitatea duce la polifonie, un concept central în dialog și un punct central al modelului polifonic introdus de Trausan et al. (Trausan-Matu, 2020; Trausan-Matu & Rebedea, 2009; Trausan-Matu, Stahl, & Sarmiento, 2007).

Vocile au fost operaționalizate anterior de M. Dascalu, Trausan-Matu, and Dessus (2013) ca lanțuri semantice care au fost obținute prin combinarea lanțurilor lexicale, adică secvențe de cuvinte repetate sau înrudite, inclusiv sinonime sau hipernime (Mukherjee, Leroy, & Kauchak, 2018). Modul în care ideile se propagă în text, tipul de cuvinte folosite, modul de exprimare, sunt toate caracteristici ale stilului de scriere și definesc tipurile de texte. Pentru a sublinia caracteristicile stilului de scriere, secțiunea următoare descrie stilometria și sistemele care ajută la calcularea diferiților indici de complexitate textuală.

2.3 Stilometria

Stilometria este studiul stilului literar care definește modul în care autorul folosește cuvintele și modul în care organizează propoziții pentru a-și exprima ideile (Holmes, 1998). Oamenii se exprimă într-un anumit mod, variind de la cuvinte și fraze simple la complexe, de la propoziții scurte la paragrafe lungi cu propoziții lungi. Chiar dacă vocabularul bogat poate sugera o calitate literală superioară, nu este întotdeauna adevărat. Ernest Hemingway a câștigat Premiul Nobel în 1954 pentru literatură, chiar dacă a folosit un număr surprinzător de mic de cuvinte (Rice, 2018).

Cuvintele funcționale sunt cuvinte care nu conțin informații textuale sau au o semnificație lexicală redusă și sunt folosite pentru a exprima relații gramaticale între alte cuvinte. Acest tip de cuvânt specifică atitudinea sau starea de vorbire. Exemple de cuvinte funcționale sunt conjuncțiile, prepozițiile, articolele gramaticale, verbele auxiliare. Aceste cuvinte se mai numesc cuvinte stop, deoarece nu conțin informații textuale. Cuvintele care nu sunt cuvinte de oprire se numesc cuvinte de conținut, acestea sunt cuvinte care conțin informații textuale și includ substantive, adjective, verbe și adverbe. Cuvintele funcționale s-au dovedit a fi destul de puternice în modul în care sunt folosite de autori. Stamatatos (2009) a realizat un studiu asupra abordărilor de identificare automată a autorului. În studiul său, Stamatatos (2009) subliniază că cuvintele funcționale „sunt utilizate într-o manieră în mare măsură inconștientă de către autori și sunt independente de subiect" (Stamatatos, 2009). Astfel, utilizarea inconștientă a cuvintelor funcționale este un avantaj în analiza stilometriei deoarece variază mai puțin decât vocabularul unui autor.

Există diverse sisteme care calculează indici pentru complexitatea textuală, la diferite niveluri de procesare: suprafață, lexical, sintaxă, semantică, structura discursului și cuvânt. Coh-Metrix (McNamara, Graesser, & McCarthy, 2014) este unul dintre primele și cele mai cunoscute sisteme care calculează diverși indici la toate nivelurile. TAALES (Kyle & Crossley, 2015; Kyle, Crossley, & Berger, 2018) măsoară peste 400 de indici pentru sofisticare lexicală. TAALED (Kyle, Crossley, & Jarvis, 2020) calculează o mare varietate de indici ai diversității lexicale. TAASSC (Kyle, 2016) măsoară o serie de indici legați de sofisticarea și complexitatea sintactică. TAACO (Crossley, Kyle, & Dascalu, 2019; Crossley, Kyle, & McNamara, 2016) calculează 150 de indici ai coeziunii locale și globale. TAALES, TAALED, TAASSC și TAACO oferă o gamă largă de indici grupați pe categorii. ReaderBench (M. Dascalu, 2014; M. Dascalu, Crossley, McNamara, Dessus, & Trausan-Matu, 2018; M. Dascalu, Dessus, Bianco, Trausan-Matu, & Nardy, 2014) este un cadrul de procesare al limbajului natural avansat open-source bazat pe analiza rețelei de coeziune (CNA) și calculează diverși indici la toate nivelurile. ReaderBench este dezvoltat în grupul nostru de cercetare, încearcă să acopere cele mai comune caracteristici care sunt apoi integrate în secvențele de procesare ulterioare.

3 Învățarea Colaborativă

Învățarea colaborativă este un proces social de construire a cunoașterii prin colaborare (Stahl, 2006). Potrivit lui Stahl (2006), mai multe faze sunt într-un ciclu continuu și contribuie la construirea atât a cunoștințelor personale, cât și sociale.

Învățarea colaborativă susținută de calculator (eng., Computer Supported Collaborative Learning; CSCL) este o paradigmă emergentă a tehnologiei educaționale (Koschmann, 1996). Dezvoltarea tehnologiei ne-a modelat concepțiile despre cunoașterea și învățarea umană (Bolter, 1984). În CSCL, tehnologia întâlnește științele educației, psihologia și filosofia. Potrivit lui Lipponen (2002), primul atelier CSCL a avut loc în 1990, iar prima conferință internațională CSCL a avut loc în 1995 la Bloomington, Indiana. CSCL presupune că învățarea colaborativă este susținută de tehnologie pentru a îmbunătăți interacțiunea dintre participanți. Tehnologia facilitează schimbul de idei și opinii și ajută la distribuirea rapidă a cunoștințelor.

CSCL a devenit din ce în ce mai utilizat în contexte educaționale datorită efectelor sale sinergice în rândul colegilor. Cursanții își împărtășesc ideile și opiniile, învață unii de la alții, în timp ce au acces la o gamă largă de materiale. Chat-urile și forumurile, cele mai frecvent utilizate medii CSCL, oferă elevilor posibilitatea de a lucra împreună pentru a rezolva probleme și de a cere ajutor atunci când întâmpină probleme - mai general, elevii construiesc în comun cunoștințe și le împărtășesc tuturor participanților (Stahl, 2006). Procesele de învățare colectivă și individuală se împletesc între ele pentru a crea cunoștințe colaborative, care sunt răspândite între toți participanții (Cress, 2013).

În mediile CSCL, tehnologia permite comunicarea și colaborarea pe tot parcursul procesului de învățare. Cu toate acestea, din perspectiva tutorelui, analiza conversațiilor rezultate este o sarcină care consumă mult timp, datorită volumului crescut al acestora. Prin urmare, instrumentele automate care analizează conversațiile și evaluează colaborarea dintre participanți au devenit o necesitate. Mai mult, colaborarea și creativitatea între colegi pot fi stimulate folosind procese automate (Stamati, Dascalu, & Trausan-Matu, 2015). În plus, sarcinile repetabile, care nu necesită un răspuns uman, pot fi efectuate automat de către aceste sisteme în timp ce se utilizează agenți de conversație (Gabriel, Hahne, Zimmermann, & Lenk, 2021).

3.1 Dialogismul, Cadru pentru CSCL

Dialogismul este considerat cadrul teoretic sau paradigma CSCL (Koschmann, 1999; Trausan-Matu, Stahl, & Sarmiento, 2007), o componentă centrală pentru înțelegerea modului în care informațiile sunt propagate prin conversații și fluxul de informații în discuții (M. Dascalu, Trausan-Matu,

McNamara, & Dessus, 2015). Vocile (punctele de vedere) iau forma unor concepte sau evenimente care sunt propagate de-a lungul conversației de către participanții care împărtășesc perspective convergente sau divergente (Trausan-Matu, 2010). Multivocalitatea este centrată pe dialogul dintre voci și semnificații multiple, în timp ce polifonia se concentrează pe inter-animația dintre voci și relațiile care apar prin suprapunerea acestora. În modelul polifonic, vocile nu sunt asociate cu un singur individ, pot deveni subiecte sau teme emergente din discurs. Astfel, o multitudine de voci se animă între ele, fiecare cu propria identitate și muzicalitate, formând împreună o entitate inteligibilă (Trausan-Matu & Rebedea, 2009).

Inter-animarea vocală și polifonia reflectă esența unei bune sesiuni de conversație CSCL, în care interacționează multiple perspective și puncte de vedere, străduindu-se spre unitate sau diversitate. Referințele explicite și implicite din cadrul conversațiilor introduc dimensiuni longitudinale ale dialogului (de-a lungul timpului) urmărind cronologic evenimentele din cadrul discuției. Interacțiunile dintre membri se reflectă în vocile lor; astfel, polifonia poate fi percepută ca un indicator al colaborării (M. Dascalu, Trausan-Matu, McNamara, & Dessus, 2015). Mai mult, modul în care vocile evoluează în timp, influența lor asupra altor participanți, inter-animația lor, toate oferă o analiză aprofundată a vorbirii. Schimbul de idei, puncte de vedere și subiecte de interes, divergențe și convergențe reflectă o dimensiune transversală (în timp), care surprinde mișcarea atunci când membrii unei conversații se schimbă pentru a discuta alte subiecte (Trausan-Matu, Stahl, & Sarmiento, 2007). În orice moment, un subiect nou poate apărea într-o conversație și mai multe voci sunt prezente simultan. Prin urmare, o discuție reprezintă inter-animația mai multor voci, fie ele armonice sau disonante. Evoluția vocilor de-a lungul unei conversații și influența lor asupra altor participanți oferă informații valoroase despre colaborare.

3.2 Medii de Învățare Online pentru Susținerea Colaborării

Mediile de învățare online (eng., Online Learning Environments; OLE) sunt din ce în ce mai utilizate de oameni din întreaga lume, deoarece facilitează accesul rapid la resurse și informații, permit utilizatorilor să împărtășească opinii și idei și chiar să se angajeze în dezbateri deschise. Cursanții își împărtășesc experiențele și opiniile și caută răspunsuri în mediile online, în timp ce tutorii și instructorii își împărtășesc cunoștințele și expertiza. Mai mult, pe lângă facilitățile aduse cursanților și instructorilor, OLE au deschis noi domenii de cercetare, cum ar fi modelarea participării membrilor, analiza interacțiunilor și identificarea tiparelor de interacțiune în cadrul comunității (Moore, Dickson-Deane, & Galyen, 2011; Tu, 2002; Weidlich & Bastiaens, 2019).

Învățarea online este în continuă evoluție și s-a schimbat considerabil de la prima sa apariție. În 1989, Universitatea din Phoenix, unul dintre pionierii educației online, a oferit primul program online care

a devenit obiectivul principal al școlii (Britannica, 2021). Zece ani mai târziu, prima universitate complet bazată pe web - Universitatea Jones - a devenit acreditată și, doar un an mai târziu, Universitatea din Texas a oferit o serie de cursuri online pe un site web care a inclus teste, sondaje, note și calendare. Astfel, a apărut o nouă industrie - tehnologia educației online. Dave Cormier de la Universitatea din Insula Prințului Edward a inventat termenul Massive Open Online Course (MOOC) în 2008. Au apărut diverse medii MOOC în anii următori, inclusiv Coursera, edX și Udacity (Achieve Virtual, n.d.). Învățarea online a îmbunătățit învățarea la distanță și a început să acționeze ca un substitut pentru cursurile față în față. Deoarece tehnologia continuă să evolueze într-un ritm rapid, instituțiile care se străduiesc să ofere educație de înaltă calitate sunt necesare pentru a utiliza cele mai bune alternative (Hiltz & Turoff, 2005).

Au fost dezvoltate diferite tipuri de sisteme pentru a sprijini învăţarea online. Sistemele de management al învățării (eng., Learning Managemenr Systems; LMS) (Watson & Watson, 2007) oferă un mediu de învățare complex, care este capabil să gestioneze toate elementele procesului de învățare. Sistemele de management al cursurilor (eng., Course Management Systems; CMS) (Simonson, 2007) au o funcționalitate mai limitată decât LMS-urile și oferă un mediu de învățare care încurajează colaborarea între studenți. În timp ce LMS-urile se concentrează pe cursanți și organizații, CMS-urile se concentrează pe tutori și studenți. Cursurile online deschise masive (eng., Massive Open Online Courses; MOOC) sunt cursuri online disponibile oricui. MOOC-urile sunt gratuite și oferă oportunități de învățare la distanță pentru oamenii din întreaga lume.

3.3 Analiza Rețelelor Sociale pentru Modelarea Colaborării

Analiza rețelelor sociale (eng., Social Network Analysis; SNA) este o abordare frecvent utilizată în explorarea și evaluarea structurile sociale (SNA; Scott, 1988). Analiza rețelelor sociale, utilizează teoria grafurilor și rețelelor pentru a reprezenta și a înțelege structurile sociale. SNA ajută la înțelegerea comportamentului unei rețele, cum sunt conectate nodurile, care sunt relațiile dintre noduri și care sunt cele mai importante noduri. O rețea socială este un set de entități și relații (Knoke & Yang, 2019) și „oferă un model puternic pentru structura socială” (Scott, 1988).

Diverse măsuri și metrici pot fi utilizate în analiza rețelelor sociale, la nivel de entitate (nod) sau rețea. Măsurile de centralitate oferă informații despre importanța nodurilor și a muchiilor din cadrul unei rețele sociale: centralitate de grad, centralitate între, centralitate de apropiere, centralitate PageRank, centralitate proprie. Un scor ridicat de centralitate indică influență, putere în rețeaua socială. La nivel de rețea, măsurile includ densitatea, dimensiunea rețelei.

Centralitatea gradului unui nod indică numărul de noduri care sunt legate direct de nodul curent (Das, Samanta, & Pal, 2018). Centralitatea gradului este cea mai simplă măsură a conectivității unui nod și

identifică nodurile cu cele mai multe conexiuni la alte noduri din rețea. Astfel, măsura centralității gradului identifică care este cel mai popular nod dintr-o rețea și care este cel mai puțin popular. Nodurile cu un grad înalt de centralitate au cele mai bune conexiuni și pot fi puternice în rețea. Centralitatea gradului definește numărul de muchii pentru un anumit nod. Dacă rețeaua este direcționată, atunci există două măsuri de centralitate a gradului: *indegree* și *outdegree*. Indegree este numărul de margini care ies din nodul curent către alte noduri din rețea, în timp ce outdegree este numărul de margini care intră în nodul curent.

Centralitatea între a unui nod reprezintă numărul celor mai scurte căi între un nod sursă și un nod destinație, care trec prin nodul curent. A fost introdus mai întâi de Shaw (1954) pentru rețelele umane și apoi de Freeman (1977) pentru rețelele sociale. Cel mai adesea, nodurile cu un scor de centralitate mare între ele acționează ca „punte" între celelalte noduri și creează căi scurte în cadrul rețelei. Astfel de noduri joacă un rol important în rețea, deoarece controlează fluxul cel mai puternic de informații între noduri, iar eliminarea acestuia provoacă o întrerupere destul de mare (Arif, 2015).

Măsura de centralitate a apropierii reprezintă distanța medie între un nod și fiecare nod din rețea (Hansen, Shneiderman, Smith, & Himelboim, 2011). Această măsură verifică cât de puternic este conectat un nod cu fiecare nod din rețea și găsește toate nodurile care sunt cele mai apropiate de alte noduri. Pentru a face acest lucru, măsura centralității de apropiere găsește toate căile cele mai scurte din rețea. După aceea, un scor pentru fiecare nod este calculat folosind cele mai scurte căi ale sale.

Măsura centralității proprii găsește nodurile cu cea mai mare influență în cadrul unei rețele. Centralitatea Eigen este o extensie a centralității gradului, începe prin măsurarea centralității gradului pentru nodul curent, după care măsoară centralitatea gradului tuturor nodurilor la care nodul curent se conectează și așa mai departe, până la sfârșitul rețelei. Algoritmul utilizează metoda de iterație a puterii (Cambridge Intelligence, 2021).

Algoritmul *Page Rank* (Brin & Page, 1998; Page, Brin, Motwani, & Winograd, 1999) calculează importanța sau influența fiecărui nod, pe baza relațiilor de intrare și a importanței nodurilor sursă echivalente. Folosind algoritmul Page Rank, puteți afla cine are o influență largă la nivel de rețea sau cine este important în cadrul rețelei la nivel macro. Chiar dacă Page Rank este o variantă a Eigen Centrality care a fost concepută pentru a clasifica conținutul web, folosind hyperlinkurile dintre pagini, poate fi utilizată pentru orice tip de rețea. Principala diferență între centralitatea proprie și Page Rank este că Page Rank are în vedere direcția marginii..

Dimensiunea rețelei reprezintă numărul de noduri din rețea, nu include numărul de margini. Densitatea rețelei se calculează ca raport între numărul de muchii existente și numărul total de muchii posibile.

4 Analiza Computațională a Discursului

Prelucrarea limbajului natural (eng., Natural Language Processing; NLP) (Jurafsky & Martin, 2008) este o zonă foarte populară de cercetare care explorează modul în care calculatoarele pot fi utilizate pentru a înțelege, manipula și interpreta textul sau vorbirea în limbaj natural (Chowdhury, 2003). Identificarea modului în care ființele umane înțeleg și utilizează limbajul natural scris sau vorbit subliniază dezvoltarea unor instrumente și tehnici speciale pentru sistemele informatice de înțelegere a limbajului natural. NLP se bazează pe mai multe discipline, inclusiv informatică și lingvistică de computațională. Prin combinarea lingvisticii computaționale cu învățarea automată, metodele statistice și modelele de învățare profundă, calculatoarele sunt capabile să proceseze limbajul uman, să îl înțeleagă, pentru a extrage diverse caracteristici utilizate în diferite sarcini NLP.

Limbajul natural este foarte complex și divers. Fiecare limbă are propriile reguli, atât gramaticale, cât și sintactice. Oamenii se exprimă în multe feluri, foarte diferit, atât în scris, cât și verbal. În limbajul scris, de cele mai multe ori oamenii prescurtează cuvinte, fac greșeli gramaticale sau sintactice, omit semnele de punctuație. În limba vorbită, împrumutăm uneori termeni din alte limbi, folosim regionalisme sau accente diferite. Astfel, pe lângă modelarea limbajului uman, înțelegerea sintactică și semantică sunt foarte importante pentru sarcinile NLP.

Sarcinile NLP includ recunoașterea vorbirii, analiza sentimentelor, etichetarea parțială a vorbirii, dezambiguizarea sensului cuvântului, recunoașterea entității denumite, rezoluția de corespondență, generarea limbajului natural, înțelegerea limbajului natural. Aplicațiile NLP acoperă o varietate de domenii: traducere automată, chatbots, asistenți virtuali, detectare spam, rezumare, analiză sentiment, întrebări și răspunsuri.

4.1 Secvența de Prelucrare a Limbajului Natural

Secvența de prelucrare a limbajului natural este un lanț de pași care se alimentează reciproc pentru a rezolva o anumită problemă. Secvența comună constă din trei pași: procesarea textului, extragerea caracteristicilor și modelarea. Fiecare pas returnează un rezultat care este utilizat în pasul următor. Pasul de procesare a textului se referă la curățarea textului, tokenizarea, lematizarea, etichetarea parțială a vorbirii. Extragerea caracteristicilor obține reprezentări de caracteristici adecvate pentru sarcina și modelul NLP dorit. Modelarea presupune adaptarea parametrilor la datele utilizate în instruire și utilizarea acestuia pentru predicții.

Au fost dezvoltate diverse biblioteci pentru sprijinirea sarcinilor NLP: Spark NLP (John Snow Labs Inc, 2021), spaCy (Explosion, 2016-2021), Allen NLP (The Allen Institute for Artificial Intelligence,

2021), Stanford Core NLP (Stanford NLP Group, 2021), Gensim (LGPLv2.1, 2009-2021), Rasa (Rasa Technologies Inc, 2021).

SpaCy (Explosion, 2016-2021) este o bibliotecă open-source Python pentru procesare avansată a limbajului natural. SpaCy poate fi utilizat pentru procesarea sau pre-procesarea textelor pentru a construi sisteme de extragere a informațiilor, rezumare, înțelegere a limbajului natural, întrebări și răspunsuri. SpaCy oferă suport în mai multe limbi, iar secvențele sale de procesare sunt antrenate pentru fiecare limbă. SpaCy are diverse caracteristici referitoare la noțiuni lingvistice sau funcționalități pentru învățarea automată generală. Caracteristicile SpaCy includ: Tokenizarea, Etichetarea în parte a vorbirii (eng., Part-of-speech; POS), Analiza dependenței, Lematizarea, Detectarea limitelor frazelor, Recunoașterea entității denumite, Legătura entităților, Similaritatea, Clasificarea textelor, Potrivirea bazată pe reguli, Antrenarea și Serializarea.

4.2 Modele Semantice și de Limbă

Reprezentări de Cuvinte folosind Modele Semantice

Reprezentările de cuvinte au un rol important în NLP și se referă la reprezentarea unui cuvânt folosind un vector (Liu, Lin, & Sun, 2020). Cuvintele reprezintă cea mai mică unitate semnificativă folosită de oameni în vorbire sau scriere. Propozițiile și paragrafele conțin grupuri de cuvinte aranjate astfel încât textul rezultat să aibă sens și să fie ușor de înțeles. Pentru a înțelege un astfel de text, oamenii trebuie să înțeleagă corect fiecare cuvânt pe care îl conține și semnificația acestuia. Cu toate acestea, pentru sarcinile NLP, fiecare cuvânt trebuie să aibă o reprezentare clară pentru a ajuta modelele să înțeleagă cât mai bine semnificațiile cuvintelor.

Reprezentările vectoriale ale cuvintelor codifică cuvinte similare în codificări similară folosind un spațiu n-dimensional. O reprezentare vectorială este un vector dens, ale cărui valori sunt puncte flotante și reprezintă greutăți învățate de un model. Incorporările de cuvinte au fost utilizate în diverse sarcini NLP, cum ar fi: analiza sentimentelor, clasificarea documentelor, sistemele de recomandare, rezumarea textului.

Analiza Semantică Latentă (eng., Latent Semantic Analysis; LSA) (LSA; Dumais, 2004; Landauer & Dumais, 2008; Landauer & Dumais, 1997) este o metodă statistică și matematică pentru extragerea semnificației cuvintelor din texte cu scopul de a deduce conceptele. LSA folosește modelul Bag of Words (BoW) care reprezintă un text folosind aparițiile fiecărui cuvânt pe care îl conține.

Alocarea Latentă Dirichlet (eng., Latent Dirichlet Allocation; LDA) (LDA; Blei, Ng, & Jordan, 2003) este un model probabilistic pentru modelarea subiectelor. În cadrul unui document sau al unei colecții de documente, pot exista mai multe subiecte. Modelarea subiectelor ajută la înțelegerea și

organizarea automată a colecțiilor de documente. În același timp, modelarea subiectelor este foarte utilă pentru a rezuma conținutul documentelor sau pentru a căuta informații în colecțiile de documente.

Word2vec (Mikolov, Chen, Corrado, & Dean, 2013) este o metodă bine cunoscută pentru reprezentarea vectorială a cuvintelor. Word2vec ia ca intrare un corpus de texte și returnează un set de vectori care înseamnă reprezentarea vectorială a cuvintelor din corpus. Vectorii sunt aleși folosind funcția de similaritatea cosinus. Word2vec se bazează pe o arhitectură de tip rețea neuronală și constă din două modele: modelul continuu Bag-of-Words (CBOW) și modelul continuu skip-gram (SKIP – GRAM).

Reprezentări Contextualizate utilizând Modele de limbă bazate pe Transformatoare

Transformer (Vaswani et al., 2017) este un model de arhitectură de rețea neuronală, care se bazează pe un mecanism de atenție, o auto-atenție mai specifică. Mecanismul de atenție este utilizat pentru a extrage dependențele dintre intrare și ieșire uzitându-se la întreaga intrare, nu numai la ultimul segment. Spre exemplu, dacă avem următoarea intrare „The cat saw a mouse. It was in the mood to play, so it tried to catch it, but it got into the hole in time. ", Mecanismul de atenție poate detecta la cine se referă fiecare „it" folosit. Potrivit lui Vaswani et al. (2017) „Arhitectura de tip Transformer este primul model de transducție care se bazează în întregime pe auto-atenție pentru a calcula reprezentări ale intrării și ieșirii sale, fără a utiliza rețele neuronale recurente (RNN) aliniate la secvență sau convoluție".

Bidirectional Encoder Representations from Transformers (BERT; Devlin, Chang, Lee, & Toutanova, 2018) este un model de reprezentare a limbajului, care a adus cele mai recente rezultate de ultimă generație în diferite sarcini NLP, cum ar fi clasificarea textului, întrebare și răspuns, rezumarea textului, clasificarea sentimentului, predicția propozițiilor, inferența limbajului natural, dezambiguizarea sensului cuvântului. BERT a fost conceput pentru a pregăti reprezentări bidirecționale profunde din texte neetichetate, utilizând modele de limbă mascat. BERT se bazează exclusiv pe mecanismul de atenție. Cuvintele își schimbă sensul pe măsură ce propoziția sau paragraful se dezvoltă. Cuvintele pe care modelul e centrat devin mai ambigue dacă propoziția sau paragraful conține mai multe cuvinte, iar semnificația lor crește. BERT explică efectul pe care îl au alte cuvinte asupra cuvântului focus. Fiind un model bidirecțional, în timpul fazei de antrenament BERT învață informații atât din partea stângă, cât și din partea dreaptă a contextului unui anumit simbol.

4.3 Analiza Rețelei de Coeziune

Analiza rețelei de coeziune (eng., Cohesion Network Analysis; CNA) (CNA; M. Dascalu, Trausan-Matu, McNamara, & Dessus, 2015) combină abordările NLP avansate cu analiza rețelelor sociale (SNA; Scott, 2017) pentru a analiza și a oferi o viziune aprofundată a structurii discursului centrată pe coeziunea textului. SNA reprezintă și examinează structurile sociale folosind teoria grafurilor; CNA este strâns corelat cu SNA deoarece oferă indici echivalenți pentru a evalua participarea utilizând grafuri de rețea care utilizează estimări ale coeziunii discursului pentru a simula informații schimbate între participanți. CNA îmbunătățește SNA deoarece consideră coeziunea semantică bazată pe NLP atunci când modelează interacțiunile dintre participanți. În consecință, CNA consideră atât interacțiunile elevilor, cât și conținutul discursului pentru a efectua o analiză aprofundată a interacțiunilor acestora.

În cadrul CNA, coeziunea este calculată utilizând diferite măsuri de similaritate din diferite modele semantice, și anume: Analiza Semantică Latentă (LSA; Landauer & Dumais, 1997), Alocarea Latentă Dirichlet (LDA; Blei, Ng, & Jordan, 2003) sau word2vec (Mikolov, Chen, Corrado, & Dean, 2013). Graful de coeziune reprezintă o structură cu mai multe straturi, alcătuită din noduri diferite și legăturile între ele, și poate fi folosit ca un proxy pentru conținutul semantic al discursului (M. Dascalu, 2014).. Graful de coeziune constă dintr-un nod central, care reprezintă firul conversației. Apoi, nodul central este împărțit în contribuții, care sunt împărțite în continuare în propoziții și cuvinte. Legăturile sunt construite pentru a calcula un scor de coeziune care denotă relevanța unei contribuții într-o conversație sau impactul unui cuvânt într-o propoziție sau contribuție. Graful include, de asemenea, linkuri explicite adăugate de participanții la conversații, cum ar fi „răspuns la". Pe lângă prezicerea colaborării (M. Dascalu, McNamara, Trausan-Matu, & Allen, 2018) și a cursurilor (M. Dascalu, McNamara, Trausan-Matu, & Allen, 2018), CNA a fost de asemenea angajat cu succes pentru a prezice răspunsul bloggerilor comunității la cererile pentru noilor veniți prin evaluarea automată a dialogului (Nistor, Dascalu, Serafin, & Trausan-Matu, 2018; Nistor, Dascalu, Tarnai, & Trausan-Matu, 2020).

Înțelegerea este un proces dificil și provocator, pentru care elevii trebuie să înțeleagă cuvinte și propoziții, să conecteze ideile și să le lege de cunoștințele anterioare, creând în același timp o reprezentare mentală coerentă a textului citit. Un factor important în procesul de înțelegere se referă la coeziunea textului (McNamara, 2004), care ia în considerare gradul în care există legături semantice între idei într-un text. Coeziunea este mai mare atunci când există mai multe idei și cuvinte care se suprapun și când conexiunile dintre idei sunt explicite. Textul cu coeziune redusă este mai dificil de înțeles, în special pentru cititorii slabi și pentru cititorii mai puțin calificați (O'Reilly & McNamara, 2007). Procesul de depășire a lacunelor de coeziune este și mai dificil atunci când cursanții se

confruntă cu documente multiple care necesită stabilirea conexiunilor atât în interiorul, cât și între fragmente de text disparate. Realizarea conexiunilor între mai multe texte este considerabil mai dificilă decât analiza unui singur text. Unele fragmente de text pot fi legate semantic, în timp ce altele pot fi izolate, distale și, astfel, mai dificil de recunoscut sau dedus.

Partea 2 – Studii Empirice

5 Modelarea Stilului de Scriere folosind Indici de Complexitate Textuală

Acest capitol se concentrează pe două studii de caz axate pe analiza stilului de scriere în discursurile mass-media românești despre criza economică și discursul militar.

5.1 Studii de Caz

Primul studiu de caz (Terian, Cotet, Sirbu, Dascalu, & Trausan-Matu, 2019) analizează discursurile din mass-media din România despre criza economică din anii 2008 - 2018. Nu mulți oameni sunt specialiști în economie, dar toată lumea tinde să aibă impresia de a ști ceva despre crizele economice - și, mai presus de toate, este dornică să vorbească despre asta. Criza economică pare să fi fost subiectul dominant al ultimului deceniu cel puțin în România, generând o pluralitate de discursuri paralele sau chiar contradictorii. Prin urmare, utilizarea instrumentelor automate pentru analiza acestor texte este binevenită.

Al doilea studiu de caz (Dragomir, Dascalu, Dascalu, Terian, & Trausan-Matu, 2020) explorează discursul militar dintr-o perspectivă diacronică axată pe o analiză lingvistică aprofundată a diferențelor dintre stilurile de scriere ale documentelor oficiale NATO. Abordarea constă într-o investigație cantitativă a evoluției acestui limbaj specific de-a lungul a două perioade diferite - Războiul Rece (1949 - 1990) și perioada post-Război Rece (1991 - 2018) -, cu accent explicit pe examinarea discursului corespunzător la dinamica puterii. Scopul principal al studiului este de a descrie și explica modul în care două relații de putere predominante - integrative și contradictorii - au fost reificate în discursul Alianței pe o perioadă de șaptezeci de ani, strâns corelată cu contextul istoric și politic.

5.2 Întrebări de cercetare

Cele două studii de caz menționate în secțiunea 5.1 adresează următoarea întrebare de cercetare:

- **RQ1:** În ce măsură indicii de complexitate textuală sunt predictivi din prisma diferențelor în stilul de scriere atunci când se efectuează analize multilingve pe diferite corpuri și tipuri de discurs (spre exemplu, documente oficiale NATO și mass-media din România)?

5.3 Metodă

Criza economică reflectată în mass-media din România - Corpus

Corpusul acestui studiu de caz a inclus 200 de texte, grupate in 4 categorii (publicații academice specializate, publicații non-academice specializate, publicații non-academice nespecializate, dezbateri informale), care acoperă în mod egal categoriile - în termeni de sumă (50), nu neapărat în funcție de lungime. Toate textele selectate au fost publicate în perioada 1 ianuarie 2008 - 31 decembrie 2018. Ne-am propus să acoperim toți cei unsprezece ani calendaristici incluși în acest interval de timp, dar nu am intenționat să obținem o distribuție egală a textelor pe an, întrucât o astfel de omogenizare ar fi fii artificial. În schimb, în ceea ce privește ultimele două categorii, am căutat să includem sursele din care am putea identifica texte care acoperă o gamă largă de perspective ideologice (cel puțin din punctul de vedere al stângii versus opoziția de dreapta).

Discursuri NATO - Corpus

Corpusul pentru acest studiu de caz a fost selectat din arhivele NATO, disponibile online ca documente publice. Selecția corpusului a fost gestionată luând în considerare câteva criterii importante: tipul de putere investigat (integrativ versus contradictoriu), intervalul de timp (1949 - 2018), relevanța textelor pertinente alese cu atenție dintr-o colecție cuprinzătoare de documente (peste 1.000), dimensiunea și accesibilitatea arhivelor (dintre care unele sunt învechite sau conțin linkuri lipsă) și varietatea actuală a textelor (peste 25 de categorii tematice și peste 150 de subiecte). Materialul empiric utilizat ca bază pentru analiză este compus în principal din documente oficiale NATO rezultate din 114 reuniuni ministeriale (63 la nivelul miniștrilor apărării și 51 la nivelul miniștrilor de externe) și 30 de summit-uri, care au avut loc între 1949 și 2018.

Indici de complexitate textuală multilingv

Indicii de complexitate textuală pentru limba română (studiu de caz mass-media din România) și limba engleză (studiu de caz NATO) au fost generați utilizând ReaderBench (M. Dascalu, 2014; M. Dascalu, Dessus, Bianco, Trausan-Matu, & Nardy, 2014). Luând în considerare dimensiunea corpusului nostru pentru limba română care depășește 800 de milioane de cuvinte extrase din surse publice online, doar modelul word2vec a fost utilizat în experimentele actuale, pe lângă distanțele semantice WordNet (Miller, 1995). Aproximativ 200 de indici de complexitate personalizați pentru limba română au fost generați utilizând ReaderBench și au fost utilizați să analizeze diferențele dintre

cele patru stiluri de scriere, toate abordând criza economică. Indicii de complexitate textuală care acoperă trăsături lexicale, semantice și coezive au fost utilizați în analizele statistice pentru a evidenția diferențele în stilul de scriere între documentele selectate. În ceea ce privește studiul de caz NATO, mai mult de 800 de indici de complexitate, inclusiv toți indicii listei de cuvinte, au fost calculați utilizând ReaderBench. Modelele semantice de bază au fost antrenate folosind corpusul Corpus of Contemporary American English (COCA) (Davies, 2010).

Analize Statistice

ReaderBench facilitează o analiză statistică aprofundată a diferitelor proprietăți, inclusiv lexic, sintaxă, semantică, coeziune a textului, precum și structura discursului. Indicii generați de ReaderBench (pentru ambele studii de caz) au fost filtrați în continuare pe criterii diferite, așa cum este descris mai jos.

Primul criteriu a presupus acoperirea lingvistică, și anume faptul că un index este relevant și poate fi calculat pentru cel puțin 20% din documente; nu s-au luat în considerare indicii cu acoperire lingvistică redusă. Majoritatea acestor indici erau legați de numărul de liste de cuvinte specifice. În al doilea rând, a fost verificată normalitatea distribuției fiecărui indice; mai precis, au fost eliminați indicii cu valoarea absolută a asimetriei sau kurtozei mai mari de 2. În al treilea rând, s-au efectuat teste de multicoliniaritate bazate pe comparații pe perechi ($r > .90$) și s-au păstrat doar cei mai predictivi indici. Al patrulea criteriu a constat în testul lui Levene privind egalitatea variațiilor de eroare și ignorarea indicilor ale căror valori p rezultate sunt semnificative ($p < .05$).

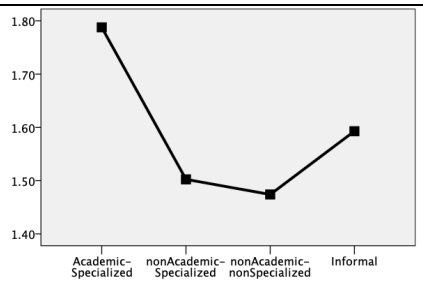
Pentru studiul de caz al presei române s-au efectuat analize statistice pentru a examina diferențele în stilurile de scriere din cele 4 tipuri de documente (academice specializate, non-academice specializate, nespecializate neacademice, informale). Pentru studiul de caz NATO, au fost efectuate analize statistice pentru a investiga diferențele în stilurile de scriere ale discursurilor NATO pe baza tipului și perioadei de timp (2 intervale între 1949 și 2018) în care au fost produse. Analiza noastră s-a axat pe proprietățile lexicale, semantice și coezive ale documentelor analizate. ANOVA-uri unidirecționale pentru fiecare indice de complexitate textuală au fost realizate pentru a identifica diferențe semnificative în stilurile de scriere. Ulterior, ambele studii de caz au efectuat o analiză funcțională discriminantă (eng., Discriminant Function Analysis; DFA) în trepte (Klecka, 1980).

5.4 Rezultate

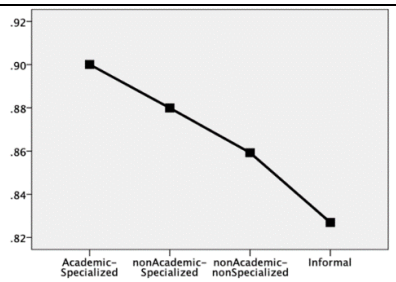
Diferențe de discurs în mass-media din România

Analiza a pornit de la aproximativ 200 de indici de complexitate textuală pentru limba română, care au fost filtrați în continuare pe diferite criterii, așa cum este descris în secțiunea Analize Statistice.

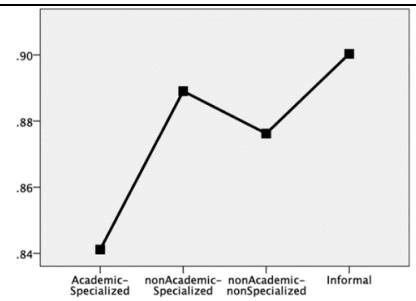
După ce cele patru filtre au fost aplicate secvențial, au trecut 17 indici care au fost supuși unei ANOVA unidirecționale pentru a identifica diferențe semnificative în stilurile de scriere din cele 4 tipuri de documente. Cei mai reprezentativi 8 indici din punct de vedere al fluctuațiilor sunt afișați în Figura 2.



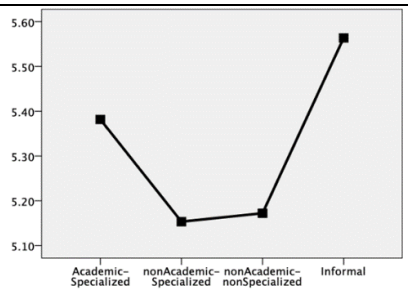
(a) Diferența de lungime a cuvântului între tulpină și forma originală (M)



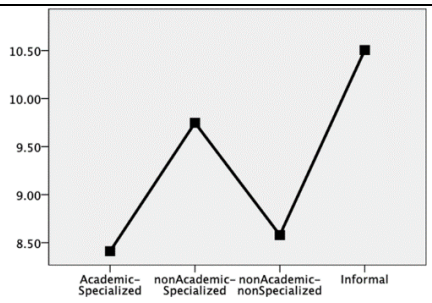
(b) *Coeziune locală* – Distanța Wu-Palmer între propoziții și paragrafe (M)



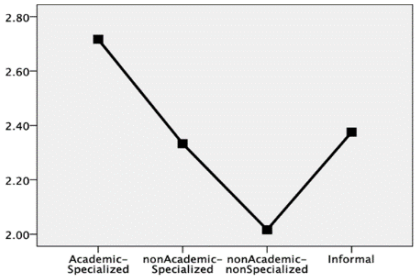
(c) *Coeziune globală* – similaritate word2vec între paragrafe și document (M)



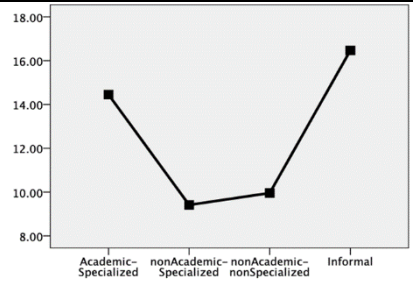
(d) Entropie de cuvinte



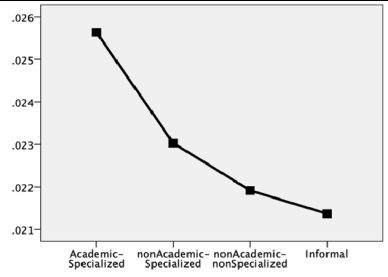
(e) Cuvinte de conținut unic pe propoziție (M)



(f) Conjuncții coordonatoare pe propoziție (M)



(g) Scorul CNA per paragraf (SD)



(h) Procentul de cuvinte incluse în lanțurile lexicale

Figura 2. a-h: Comparație între stilurile de scriere în termeni de indici de complexitate textuală.

A fost efectuat un stepwise DFA pentru a prezice tipul de document pe baza proprietăților stilului de scriere care stau la baza acestuia. DFA a raportat o acuratețe de 71,05% pentru clasificarea corectă a 143 de documente (48 + 27 + 29 + 39) din totalul de 200. Rezultate similare au fost obținute pentru LOOCV (eng., Leave - One - Out Cross - Validation) care a raportat o precizie de 64,50%. Toate scorurile F1 depășesc dublul valorii inițiale aleatorii, cu toate acestea diferențele în scorurile F1 pentru cele 4 categorii sunt considerabile. Se poate remarca o distincție clară între documentele academice de specialitate și toate celelalte tipuri (scor F1 de 96%), urmată de texte informale (scor F1 de 67%), în timp ce clasificatorul are cele mai mari probleme în identificarea diferențelor dintre lucrările specializate și non-specializate non-academice, care sunt similare în stilul de scriere.

Discursuri NATO Contradictorii versus Integrative

În primul rând, 91 dintre indicii de complexitate textuală calculați de ReaderBench au fost eliminați din cauza unei acoperiri lingvistice reduse, deoarece elementele textului de bază nu au fost întâlnite frecvent în discursul NATO (adică nu au fost prezente în cel puțin 20% din toate documentele) . Pe baza specificităților discursului, corpusul nostru nu conține un număr semnificativ de apariții ale anumitor fraze de repere sau dependențe lexicale.

În al doilea rând, normalitatea a fost verificată în termeni de Kurtosis și Skewness ale căror valori absolute trebuie să fie mai mici sau egale cu 2; toate variabilele care prezintă valori mai mari au fost ignorate în analizele de urmărire.

În al treilea rând, toate variabilele au fost testate folosind testul Levene al egalității variațiilor de eroare și acei indici pentru care valorile p rezultate sunt semnificative ($p < .05$) au fost ignorate deoarece au prezentat o diferență între varianțele din populație. Astfel, 341 de indici au fost păstrați și au fost introduși în ANOVA bidirecțional pentru a examina dacă proprietățile documentelor diferă în funcție de tipul documentului (integrativ și contradictoriu) și perioada (1949 - 1990 și 1991 - 2018). Pentru a înțelege mai bine diferențele, prezentăm în Figura 3 diagramele de profil calculate folosind mijloacele marginale estimate pentru cei mai reprezentativi indici de complexitate textuală. Aceste valori, coroborate cu reprezentarea lor vizuală din Figura 3, stau la baza discuțiilor detaliate din secțiunea următoare. Cei mai predictivi indici au explicat o variație considerabil mai mare în termeni de perioadă (parțială $\eta^2 = .473$, $p < .001$) spre deosebire de tip (parțială $\eta^2 = .208$, $p < .001$), denotând că au existat diferențe mai mari în timp decât între tipuri.

Ulterior, a fost efectuat un DFA treptat pentru a prezice tipul și perioada unui text dat pe baza proprietăților stilului de scriere care stau la baza. Toate variabilele rămase au fost eliminate și au fost considerate predictori nesemnificativi. Rezultatele demonstrează că DFA care utilizează acești cinci indici a diferențiat semnificativ textele, Wilks ' $\lambda = .849$, $\chi^2 (df = 3) = 29.301$, $p < .001$. DFA a

alocat corect 104 (21 + 21 + 23 + 39) din cele 184 de documente din setul total, rezultând o precizie de 56,50% (nivelul de șansă pentru această analiză este de 25%). Pentru LOOCV, analiza discriminantă a alocat 101 (20 + 20 + 23 + 38) din cele 184 de texte pentru o precizie de 54,90%.

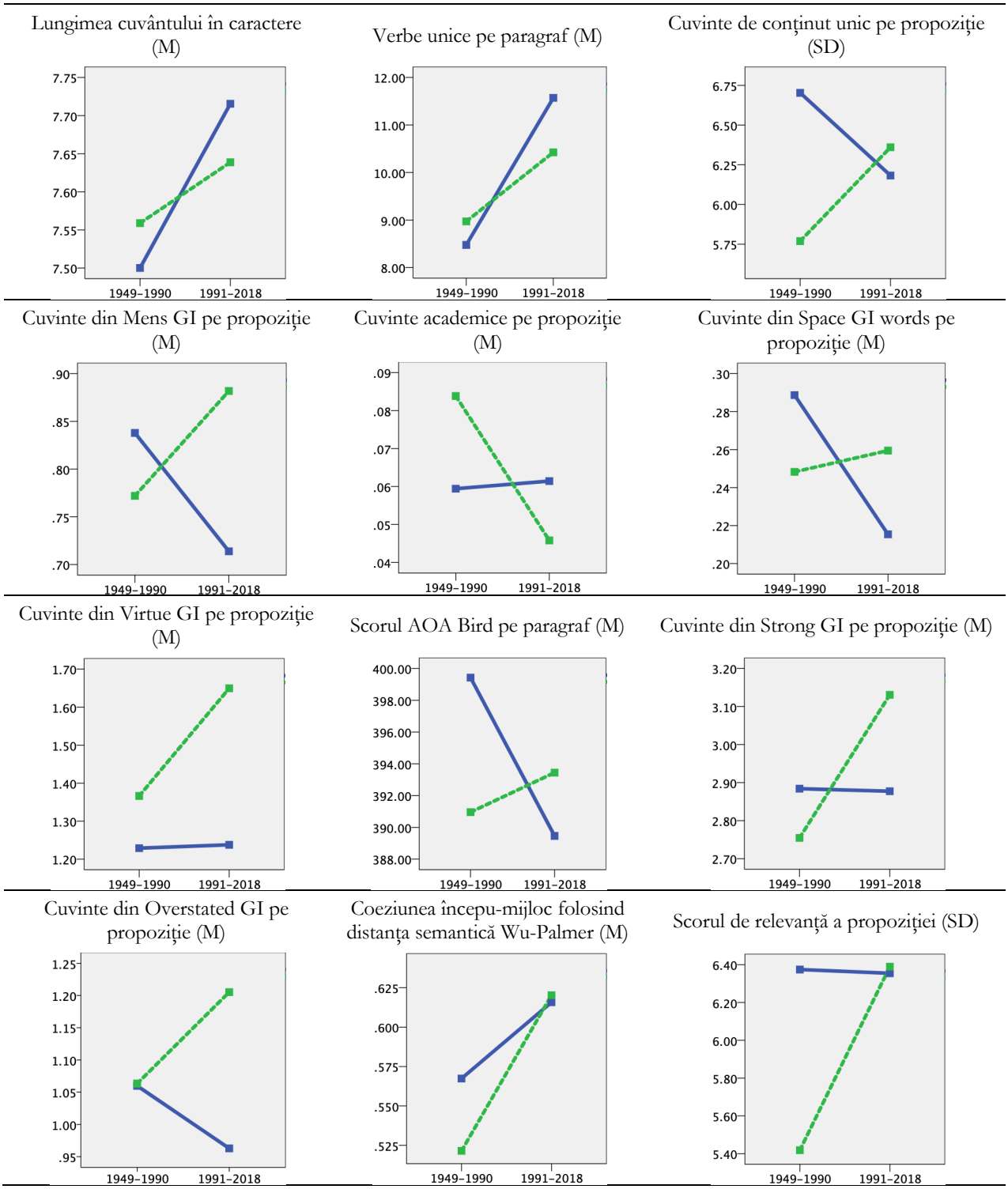


Figura 3. Graficele profilurilor bazate pe mijloacele marginale estimate ale fiecărui indice de complexitate textuală (Tipul contradictoriu este albastru, în timp ce Integrativ este punctat verde).

6 Evaluarea folosind CNA a Performanțelor Online a Elevilor

Acest capitol prezintă două studii de caz axate pe evaluarea performanței online a studenților folosind graful CNA.

6.1 Studii de Caz

Primul studiu examinează succesul matematic în cadrul unui curs mixt de licență, care a constat în prelegeri față în față și sprijin cu instrumente online, inclusiv un forum standard întrebare-răspuns, numit Piazza (Crossley et al., 2018). Metoda noastră se bazează pe variabile CNA și date click-stream pentru a prezice succesul matematic. Mai mult, diferite sociograme (de ex., grafuri de interacțiune între participanți) au fost generate prin intermediul CNA pentru a analiza implicarea și interacțiunile elevilor și pentru a identifica tendințele temporale dintre studenți (Sirbu et al., 2018a). În plus, a fost efectuată o analiză longitudinală (Sirbu, Dascalu, Crossley, McNamara, & Trausan-Matu, 2019), și s-au generat diferite puncte de vedere pentru a modela tendințele participării studenților, precum și a hărților conceptuale bazate pe relația semantică dintre cuvintele cheie găsite în forumurile de discuții.

Al doilea studiu evaluează activitatea online a studenților la un curs Moodle din România, prin analiza comportamentelor elevilor, modelarea interacțiunii acestora și prezicerea notelor studenților pe baza participării lor online (M.-D. Dascalu et al., 2021). O rețea neuronală recurentă cu celule LSTM care combină caracteristici globale, inclusiv indici de participare și inițiere, cu o analiză a seriei temporale pe intervale de timp este utilizată pentru a prezice notele elevilor, în timp ce sunt generate mai multe sociograme pentru a observa tiparele de interacțiune. Comportamentele și interacțiunile studenților sunt comparate înainte și în timpul pandemiei COVID-19 folosind două instanțe anuale consecutive ale unui curs universitar de Proiectarea Algoritmilor, desfășurat în limba română folosind Moodle.

6.2 Întrebări de cercetare

Cele două studii de caz menționate în secțiunea 6.1, adresează următoarele întrebări de cercetare:

- **RQ2:** În ce măsură feature-urile generate de platforma ReaderBench, inclusiv feature-uri derivate din graful CNA aplicat pe discuțiile de tip forum, indicii de complexitate textuală, analiza longitudinală și feature-urile derivate din datele click-stream, sunt capabile să prezică performanța studenților?
- **RQ3:** Care sunt diferențele dintre comportamentele și interacțiunile studenților înainte și în timpul pandemiei COVID-19 în două instanțe anuale consecutive ale unui curs universitar?

6.3 Metodă

Acest capitol prezintă metoda generală utilizată pentru ambele studii: modelarea succesului la un curs de matematică și performanța elevilor la cursurile Moodle.

Modelarea Succesului la un Curs Online Matematică- Corpus

Pentru acest studiu de caz, am folosit datele de la un curs de matematică pentru studenții de licență dintr-un departament de informatică (Crossley, Barnes, Lynch, & McNamara, 2017). Cursul a fost combinat și a inclus prelegeri standard (față în față), ore de birou și asistență folosind instrumente online. Instrumentele includeau un forum standard de răspuns la întrebări pentru studenți (Piazza), asistenți didactici și instructori.

Performanța Studenților la Cursul Moodle - Corpus

Pentru acest studiu de caz am colectat datele de la două instanțe Moodle diferite din anii universitari 2018-2019 (condiții normale) și 2019-2020 (condiții pandemice COVID-19) pentru un curs Moodle din semestrul II, desfășurat în limba română și centrat pe proiectarea algoritmilor. Datele au inclus postări pe forum ale studenților, lectorilor și asistenților didactici, precum și activităților lor online extrase din datele jurnalului fluxului de clicuri. Informațiile colectate de la postările de pe forum au constat din nume de utilizator, marcaje temporale ale contribuțiilor, link-uri de răspuns și textele reale din postări.

Pașii de procesare

Abordarea noastră se bazează pe CNA combinat cu tehnici de învățare automată și este utilizată pentru a evalua și modela participarea și interacțiunile elevilor. Spre deosebire de studiile anterioare efectuate de M.-D. Dascalu et al. (2020), M. Dascalu, McNamara, Trausan-Matu, and Allen (2018) and Crossley, Paquette, Dascalu, McNamara, and Baker (2016), introducem o secvență integrată care înglobează toate tipurile de indici (CNA, timp serie și complexitate textuală), acțiuni derivate din jurnalele fluxului de clicuri și o rețea neuronală pentru prezicerea notelor cursului. Considerăm comportamentele studenților ca fiind dezvăluite prin acțiuni din cadrul OLE (spre exemplu, vizualizarea sarcinilor, finalizarea sarcinilor, postarea comentariilor), precum și interacțiunile sociale și conținutul semantic al contribuțiilor lor online.

Figura 4 prezintă pașii de procesare automată care includ două etape importante. Prima etapă constă într-un proces ETL (eng., Extract Transform Load) care începe cu colectarea datelor click-stream și discuțiilor de pe forum de pe platforma Moodle care au fost exportate din baza de date relațională (spre exemplu, MariaDB). A doua etapă este centrată pe secvență de procesare automată din cadrul ReaderBench, care are ca date de intrare cele două seturi de date generate în prima etapă.

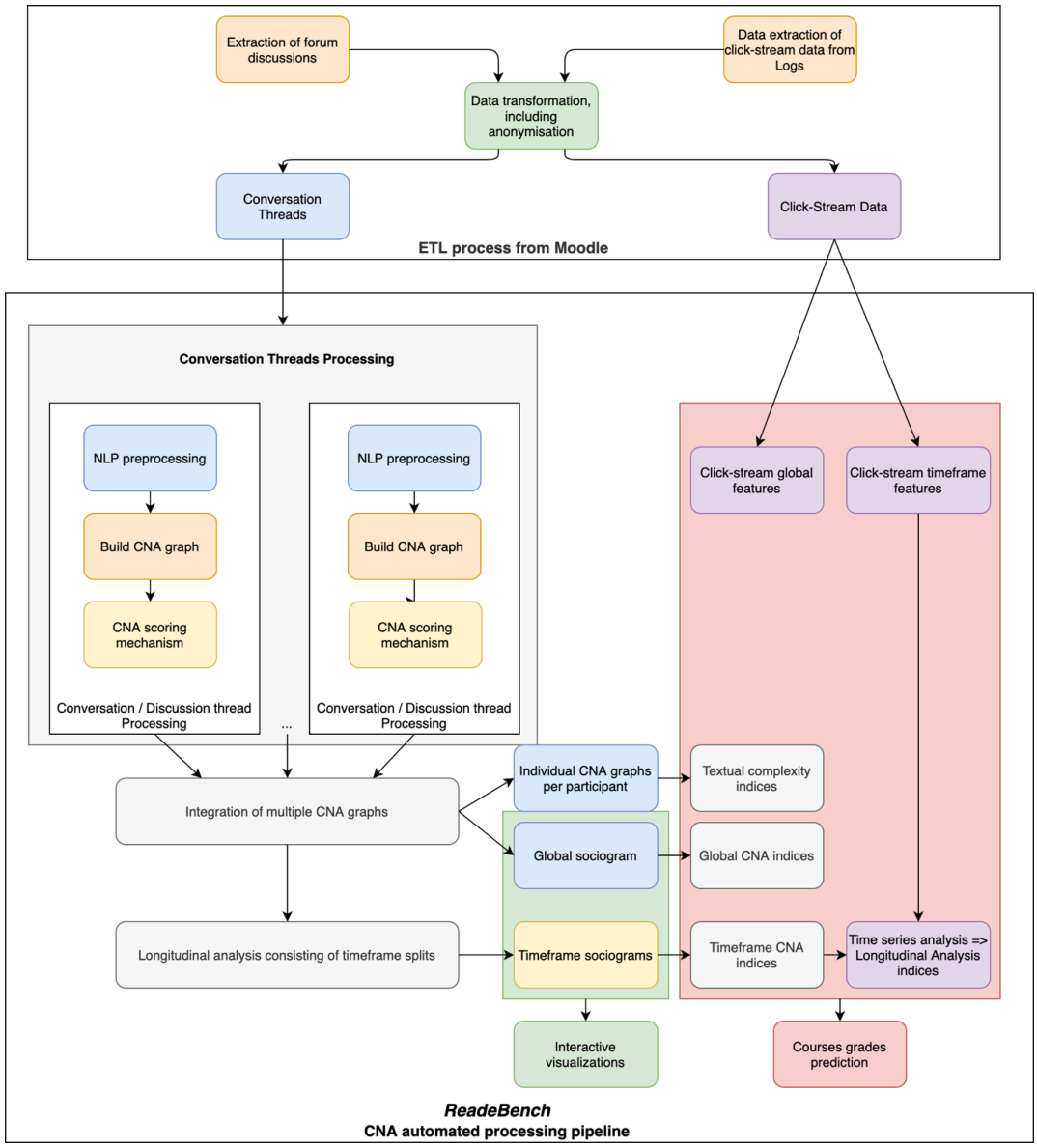


Figura 4. *ReaderBench* – Pașii de procesare pentru prezicerea notelor și generarea de vizualizări interactive.

Prezicerea Notelor Cursului

Toți indicii, inclusiv CNA, complexitatea textuală, analiza longitudinală, precum și seriile de timp aplicate pe perioade și date de flux de clicuri sunt folosite pentru a prezice notele cursurilor studenților folosind diverși algoritmi de învățare automată. Toți indicii anteriori oferă informații valoroase în ceea ce privește participarea și colaborarea (indicii CNA), stilometria (indicii de complexitate textuală), activitatea online (date click-stream), precum și regularitatea și evoluția în timp (analiza longitudinală).

Modelul constă dintr-o rețea neuronală recurentă (RNN) cu celule LSTM (Hochreiter & Schmidhuber, 1997) care combină caracteristici globale (CNA, complexitate textuală și indici LA) cu o analiză a seriilor temporale pe intervale de timp pentru a prezice notele elevilor. În cazul nostru,

intrările sunt activități săptămânale ale studenților, derivate atât din CNA indicativ al participării online, cât și din datele click-stream care reflectă interacțiunile generale cu Moodle.

Generarea de Vizualizări Interactive

Vizualizările interactive sunt generate folosind sociogramele globale și de timp, pentru a evidenția interacțiunea dintre participanți și pentru a descrie evoluția comunității de la o săptămână la alta. Astfel, sunt redate mai multe tipuri de vizualizări pentru a descrie evoluția, comportamentele și tiparele de interacțiune ale elevilor utilizând biblioteca d3.js (Bostock, 2021). Aplicația web a fost construită folosind Angular 6, în timp ce diverse biblioteci JavaScript au fost integrate pentru a crea sociogramele interactive.

6.4 Rezultate

Prezicerea Succesului la un Curs Online de Matematică

Perspectiva ierarhică de grupare a muchiilor din Figura 5 arată interacțiunea într-o manieră radială în care dependențele sunt grupate în legături de tip „spline" și participanții sunt grupați în grupul / stratul lor corespunzător. Participanții centrali sunt colorați în albastru, membrii activi sunt afișați în verde, iar membrii periferici în portocaliu. La trecerea cu mouse-ul, utilizatorul poate vedea muchiile de intrare și ieșire. Muchiile de intrare și nodurile corespunzătoare (dependente) sunt afișate în albastru închis, în timp ce legăturile de ieșire și nodurile de ieșire (dependențe) sunt colorate în roșu (consultați Figura 5 pentru participantul cu ID-ul 303190984).

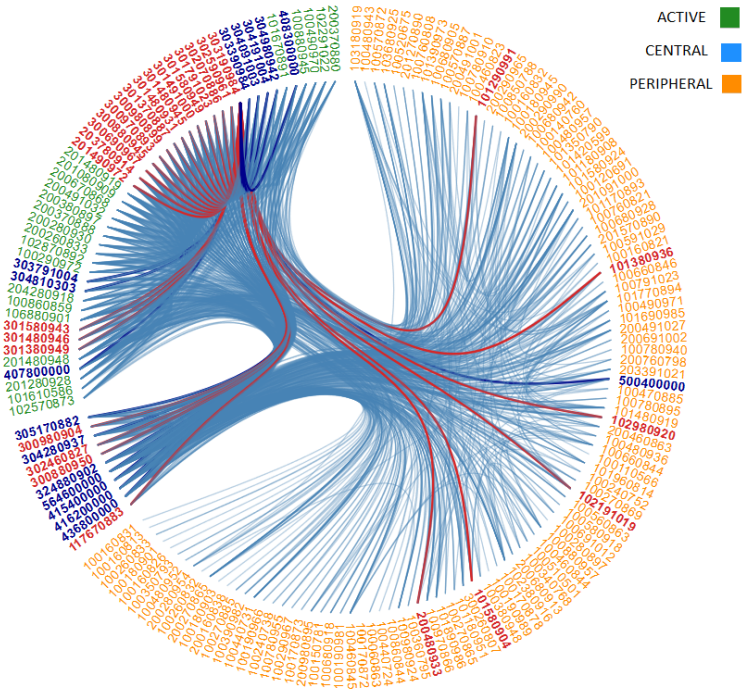


Figura 5. Sociogramă globală pentru întreaga perioadă a cursului (23 august - 24 decembrie 2013).

Figura 6 prezintă un grafic direcționat care ia în considerare puterea comunicării dintre noduri (adică elevi). Lățimea marginilor este proporțională cu calitatea textului din CNA (adică scorurile de contribuție cumulative ale mesajelor de schimb), în timp ce lungimea fiecărei margini este redată automat de biblioteca de vizualizare. În plus, dimensiunea fiecărui nod este proporțională cu scorurile medii la nivel inferior și inferior ale fiecărui participant. Vizualizările săptămânale sunt generate pentru a examina modul în care participarea în comunitate evoluează de la o săptămână la alta.

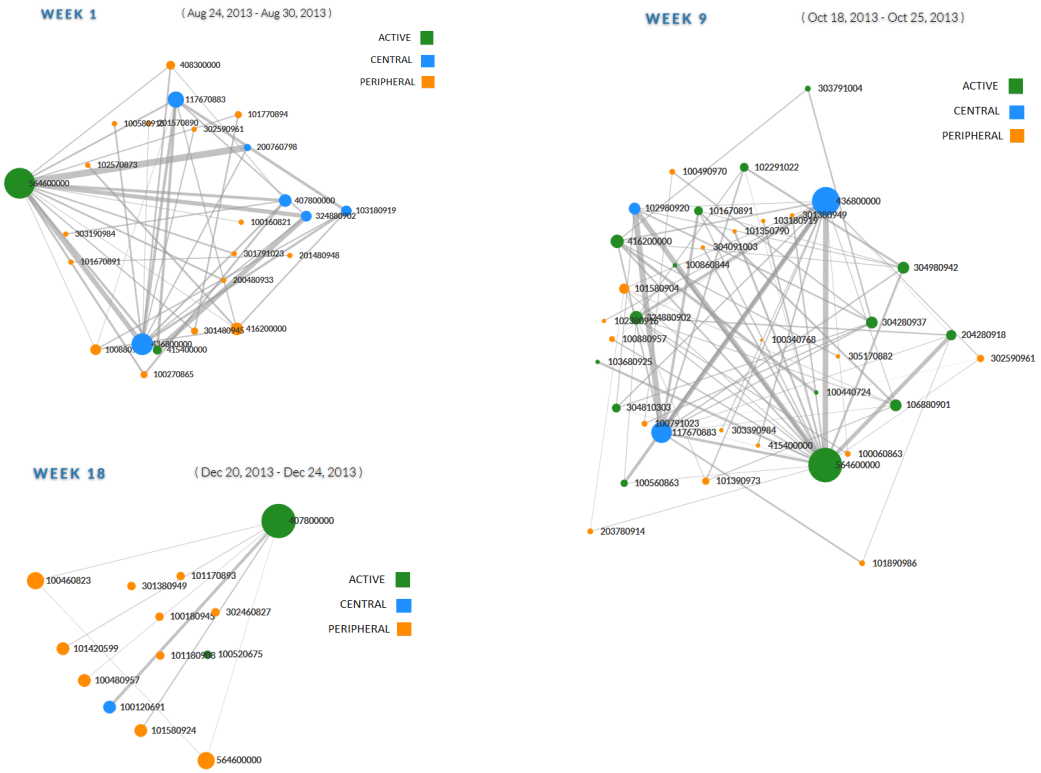


Figura 6. Vizualizări săptămânale.

Trăsăturile specifice modelelor de interacțiune comportamentală pot fi observate din cele două tipuri de vizualizări, și anume: a) o dominanță a membrilor periferici care au cel mai mic număr de interacțiuni; b) un grad crescut de colaborare de la nivelul periferic la membrii activi și, mai important, la participanții centrali (inclusiv instructorii cursului și asistenții didactici); c) un început destul de lent al comunității (săptămâna 1), urmat de o creștere a participării (săptămâna 9) și o scădere drastică în ultima săptămână; și d) deși sunt direcționate către participanți mai centrali, pot fi observate discuții libere, iar cursul nu este dominat de o singură persoană, situație frecventă în multe MOOC-uri și comunități online.

Variabilele click-stream și CNA au explicat în jur de 20% din varianță și au indicat că scorurile mai mari la matematică au fost cel mai bine prezise de către studenții care au avut o recurență mai mică (adică distanța exprimată ca număr de săptămâni între două perioade consecutive cu valori diferite

de zero) de postări care au generat efecte de colaborare (atât Social KB, cât și CNA, fără grad) cu alți participanți. Dacă participanții contribuie în mod regulat la fiecare interval de timp consecutiv din analiza longitudinală, scorul lor de recurență corespunzător este zero; în schimb, dacă sunt omise săptămâni suplimentare succesive, scorurile lor de recurență cresc. Spre deosebire de scorurile CNA care denotă implicarea activă, recurența cuantifică dezechilibrul și participarea inconsistentă în timp.

Prezicerea Notelor unui Curs pe Moodle

Au fost mulți studenți care au fost izolați pe forum în primul an universitar, fără a interacționa cu colegii sau a răspunde la întrebările altor studenți. De asemenea, doar doi asistenți didactice au fost mai activi în toate conversațiile, în timp ce lectorii au fost mai puțin angajați. În schimb, studenții au fost mai implicați în discuțiile din anul universitar 2019-2020, au colaborat mai mult și au existat mai puțini studenți care au introdus doar postări izolate ca noi fire de conversație.

În plus față de analiza comportamentului și interacțiunilor elevilor, am examinat ce cuvinte cheie au fost utilizate cel mai frecvent în discuțiile de pe forum. Utilizarea unor concepte depindea de săptămână (cu sau fără termenele limită pentru teme), în timp ce altele erau frecvent utilizate pe tot parcursul anului universitar. În săptămânile cu termenele limită pentru teme (spre exemplu, săptămânile 8 și 13), cuvinte cheie precum „lucru", „problemă" sau „merge" au fost întâlnit (vezi Figura 7). Comparativ cu anul universitar anterior, cuvinte cheie precum „problemă" și „muncă" au fost utilizate intens pe parcursul întregului curs, nu numai în săptămânile cu termenele limită pentru teme.

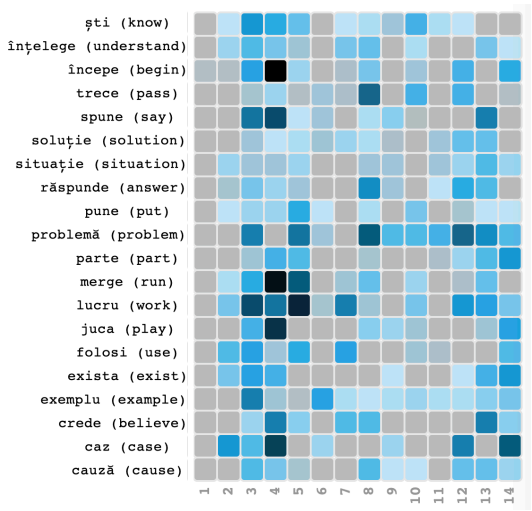


Figura 7. Harta de concepte 2020 - Cele mai discutate subiecte.

Am explorat diferențele în comportamentele și interacțiunile studenților înainte și în timpul COVID-19 folosind indicii CNA, complexitatea textuală și analiza longitudinală care au intrat în analiză după verificările de normalitate și multi-colinearitate. Așa cum era de așteptat, s-a observat o creștere

semnificativă a activității online (spre exemplu, scoruri mai mari ale contribuției CNA) în cursul anului universitar 2019-2020 afectat de COVID-19. În plus, au fost inițiate considerabil mai multe fire și rețeaua generală este mai conectată (spre exemplu, valori mai mari ale vectorului propriu CNA).

Indicii de complexitate textuală au denotat un discurs mai elaborat și mai sofisticat în al doilea an universitar. Indicii analizei longitudinale au relevat, de asemenea, rezultate interesante. O activitate sporită spre sfârșitul semestrului a fost observată în anul universitar 2018-19 (spre exemplu, panta generală este pozitivă). În schimb, focarul COVID-19 a generat un dezechilibru, o creștere drastică a participării, care a scăzut ulterior spre sfârșitul semestrului (spre exemplu, panta a fost aproape de zero sau negativă, cu abateri standard mai mari).

Modelele au fost antrenate individual pentru fiecare an universitar. Modelele pentru ambii ani academici au fost antrenate pentru maximum 2000 de epoci, cu oprire timpurie pe baza RMSE privind pierderea antrenamentului, cu o răbdare de 200 de epoci. Un strat de tip dropout de 0,2 a fost introdus înainte de ultimul strat ascuns pentru a reduce overfitting-ul și pentru a îmbunătăți generalizarea modelului. Acești parametri au fost aleși pe baza performanței obținute la validarea încrucișată.

Pentru anul universitar 2018-2019, modelul care include toate componentele, cu excepția feature-urilor săptămânale CNA și care utilizează o dimensiune a celulei LSTM de 24 și o dimensiune a stratului ascuns de 16, a explicat cel mai mare procent de varianță (R^2 = .27). Pentru anul următor, în timpul COVID-19, cea mai bună performanță a fost obținută cu aceleași caracteristici, dar cu o dimensiune a stratului ascuns de 24 în loc de 16. Acest model a obținut un R^2 mai mare de .34. Indicii CNA globali pot fi percepuți ca sume de indici CNA săptămânali din analiza longitudinală; astfel, nu este cu adevărat surprinzător faptul că acestea nu sunt întotdeauna utile în arhitectura RNN. O varianță explicată mai mare în al doilea an universitar este justificabilă, având în vedere cantitatea mai mare de date colectate (spre exemplu, mai multe postări și o rețea mai densă) care ajută la crearea unui model mai predictiv. În ambii ani academici, eliminarea indicilor de complexitate textuală duce la cea mai mare scădere a variației explicate.

7 Analiza Dialogismului folosind Modele de Limbă

Acest capitol se concentrează pe două studii de caz axate pe analiza dialogismului folosind modele de limbă.

7.1 Studii de Caz

Primul studiu de caz introduce o metodă bazată pe dialogism pentru evaluarea implicării elevilor în conversațiile de chat bazate pe lanțuri semantice calculate folosind modele lingvistice. Aceste lanțuri semantice reflectă voci emergente din dialogism care se întind și interacționează pe tot parcursul conversației.

Al doilea studiu de caz aplică modelul nostru pe o sarcină automată de evaluare a eseurilor. Scopul nostru a fost să evaluăm acuratețea predicției caracteristicilor derivate din lanțurile semantice în sarcina de evaluare automată a textului.

7.2 Întrebări de Cercetare

Cele două studii de caz menționate în secțiunea 7.1, adresează următoarea întrebare de cercetare:

- **RQ4:** În ce măsură lanțurile semantice pot fi operaționalizate folosind modele de limbă predictive atât din prisma colaborării (spre exemplu, evaluarea implicării participanților în cadrul unei conversații), cât și în studiul individual (spre exemplu, evaluarea automată a eseurilor)?

7.3 Metodă

Măsurarea Implicării în Conversaţiile CSCL - Corpus

Analiza noastră este efectuată pe aceleași conversații de chat procesate în detaliu de Dascalu și colab. (M. Dascalu, McNamara, Trausan-Matu, & Allen, 2018; M. Dascalu, Trausan-Matu, McNamara, & Dessus, 2015). Acest corpus este format din 10 discuții selectate dintr-un corpus de peste 100 de conversații care au fost evaluate de 4 evaluatori. Conversațiile au avut loc între patru și cinci studenți care studiază Computer Human Interaction, și care au dezbătut despre avantajele și dezavantajele tehnologiilor specifice CSCL (Trausan-Matu, Dascalu, Rebedea, & Gartner, 2010).

Evaluarea Automată a Eseurilor - Corpus

Am folosit setul de date Automated Student Assessment Prize (ASAP) (Kaggle Inc, 2021) pentru a evalua lanțurile semantice. Corpul eseului este format din 8 seturi diferite de eseuri, fiecare cu propria

scară de notare, toate clasele fiind normalizate la 0-1 în funcție de tipul lor. Fiecare eseu din setul de date ASAP are o lungime de aproximativ 150-550 de cuvinte.

Corpus pentru Construirea Lanțurilor Semantice

Pentru a găsi capetele capabile să detecteze legături semantice între cuvintele care aparțin aceluiași lanț, a trebuit să fie compilat un set de date specific care conține exemple de legături. Un set de euristici simple au fost folosite pentru a extrage legături din textele eșantionului, pentru toate perechile de cuvinte etichetate ca substantiv, verb sau pronume care verifică una dintre următoarele condiții: repetări de cuvinte având aceeași lemă; sinonime, hipernime din taxonomia WordNet; coreferențe identificate folosind spaCy. Corpusul TASA (Zeno, Ivens, Millard, & Duvvuri, 1995) a fost selectat drept referință și perechile anterioare de cuvinte au fost marcate pe întregul set de date.

Predicțiile Cuvintelor Asociate Semantic

Niciun cap de atenție nu este suficient de precis pentru a prezice aceste tipuri de relații semantice între cuvinte. Prin urmare, un model de predicție care învață să combine valorile de atenție din toate capetele de atenție între două cuvinte a fost antrenat pe setul de date construit pe baza TASA (Zeno, Ivens, Millard, & Duvvuri, 1995). Luând în considerare ambele direcții ale capetelor de atenție, au fost utilizate în total 288 scoruri, similar cu abordarea utilizată de Clark, Khandelwal, Levy și Manning (2019). Au fost antrenate și evaluate diferite arhitecturi: un model liniar care calculează doar o pondere pentru fiecare cap de atenție și perceptron multi-strat (eng., multi-layer perceptron; MLP) cu unul sau două straturi ascunse. Toate modelele returnează un număr trecut printr-o activare sigmoidă.

Cele mai performante modele din acest set de date nu sunt neapărat cele mai bune pentru a fi utilizate în analize ulterioare. Scopul nostru a fost să detectăm perechi de cuvinte care aparțin aceluiași lanț semantic. Ținta utilizată pentru a construi acest set de date este una foarte simplă și nu dorim ca modelul de atenție să-l prezică perfect, ci mai degrabă să fie capabil să generalizeze către alte tipuri similare de linkuri. În plus, un model liniar este mai ușor de interpretat, ceea ce ar putea oferi un avantaj suplimentar. Prin urmare, această validare inițială a fost utilizată în primul rând pentru a găsi cei mai buni parametri pentru diferite tipuri de modele, care ulterior sunt evaluați utilizând o sarcină diferită.

Modelul de predicție descris anterior poate fi utilizat pentru a înscrie toate perechile de cuvinte care se află la o anumită distanță în text. Următorul pas constă în gruparea acestor perechi de cuvinte în seturi de cuvinte legate semantic, adică lanțuri semantice. Pentru a filtra legăturile pe baza greutății prezise, a fost utilizat un prag fix. Lanțurile semantice sunt selectate sub formă de componente conectate din graful rezultat.

Construirea Lanțurilor Semantice

Metoda noastră folosește informații contextuale capturate de BERT (Devlin, Chang, Lee, & Toutanova, 2018) pentru a identifica verigile dintr-un lanț semantic. Modelul antrenat a fost apoi utilizat pentru a identifica toate legăturile semantice potențiale într-o conversație, în timp ce contabiliza toate perechile de cuvinte. Pornind de la aceste legături, lanțurile semantice sunt generate sub forma unor componente conectate în graful obținut prin interconectarea tuturor legăturilor care depășesc un prag de similaritate impus.

7.4 Rezultate

Vizualizarea Lanțurilor Semantice

Au fost introduse vizualizări interactive pentru a evidenția atât o propagare longitudinală a vocilor (vezi Figura 8), cât și o suprapunere transversală a lanțurilor semantice între participanți (vezi Figura 9). Vizualizările au fost dezvoltate folosind Angular 6 (Google Inc, 2010-2021), în timp ce legăturile dintre cuvinte au fost desenate folosind SVG. Același conversații de chat din Figura 8 au fost analizate de M. Dascalu, Trausan-Matu, McNamara, and Dessus (2015). Cuvintele sunt colorate în funcție de lanțul semantic doi pe care le aparțin, precum și de verigile corespunzătoare. Fiecare rând reprezintă o enunț, îmbogățit cu următoarele detalii: identificator, marcaj de timp și identificator de participant, care a fost generat incremental pentru anonimizare și urmat de tehnologia susținută de fiecare participant.

Id	Time	Participant ID (technology)	Utterance
222	02:43	1 (blog)	wiki for **documentation** and faqs.
223	02:43	2 (forum)	and a **forum** for technical **support** .
224	02:43	3 (wave)	**forum** for technical **support** and maybe **chat** for live support **wave** for collaboration/brainstorming/ **document sharing** .
226	02:43	2 (forum)	**chat** for live **support** inside the company .
227	02:43	4 (wiki)	yes, live **support** is a good ideea .
228	02:44	5 (chat)	**we** could also use **chat** for **meetings** ....for peope who can't come to the **meeting** .
229	02:44	3 (wave)	**wave** is better for that **you** can **share documents** etc, view what the other person is typing etc.
231	02:45	3 (wave)	**i** think that about **wraps it** up.
232	02:45	5 (chat)	does **it support** live **video** feed?
233	02:45	4 (wiki)	**wave** have also **support** for audio or **video** conversation?

Figura 8. Vizualizare longitudinală a lanțurilor semantice în cadrul unei conversații.

În contrast cu constatările inițiale ale lui M. Dascalu, Trausan-Matu, McNamara, and Dessus (2015), metoda noastră a identificat mai multe lanțuri semantice cu mai multe cuvinte înrudite - spre

exemplu, noi lanțuri semantice - concepte legate de documentare și acțiuni; mai multe cuvinte înrudite - „întâlniri” (eng., meetings) și „videoclip” (eng., video) sunt legate de „val” (eng., wave); chat-urile și forumul sunt acum agregate împreună, pe măsură ce modelele lingvistice percep similaritățile dintre concepte. Figura 9 prezintă o vedere transversală a aparițiilor legăturilor semantice ale căror concepte de bază sunt rostite de diferiți participanți. Cuvintele și legăturile sunt colorate în funcție de lanțul lor semantic corespunzător; culorile diferă între vizualizări, deoarece sunt selectate aleatoriu.

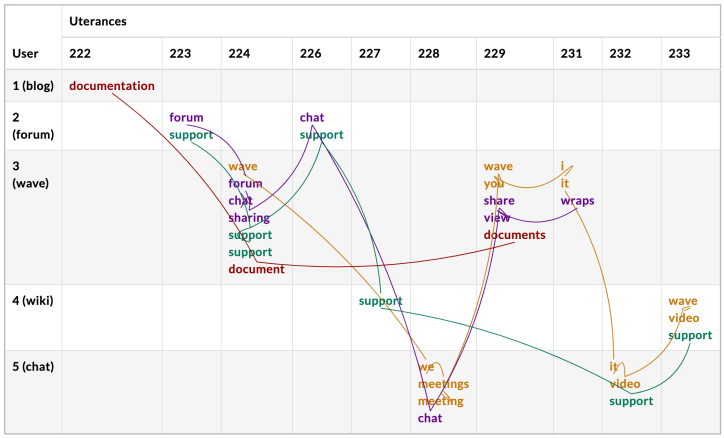
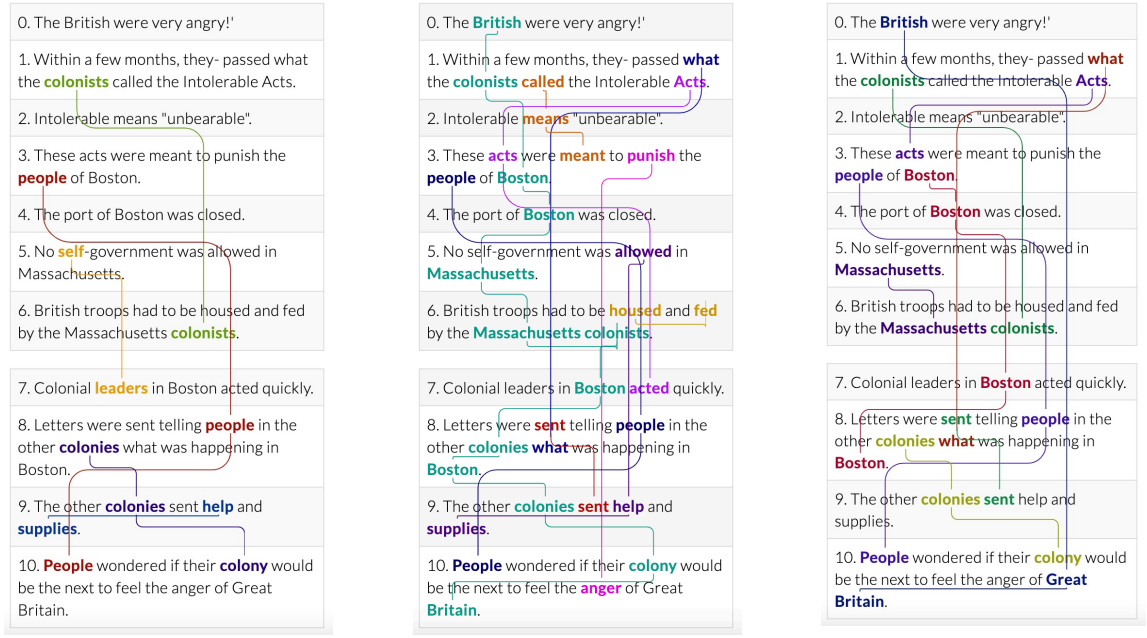


Figura 9. Vizualizare transversală a lanțurilor semantice în cadrul unei conversații.

Mai mult, a fost introdusă o vizualizare interactivă pentru a vizualiza lanțurile semantice dintr-un document care este împărțit în paragrafe și ulterior în propoziții (vezi Figura 10). Fiecare propoziție reprezintă un rând, în timp ce rândurile sunt grupate în paragraful lor corespunzător. Cuvintele care aparțin lanțurilor semantice diferite sunt colorate corespunzător lanțului lor, în timp ce cuvintele dintr-un singur lanț semantic sunt legate folosind o cale cu aceeași culoare ca lanțul.

În primul rând, putem observa densitatea mare a lanțurilor extrase cu metoda noastră în comparație cu lanțurile lexicale clasice. În plus, relațiile surprinzătoare care nu erau prezente în setul de date construit pot fi văzute în lanțurile generate. Modelul liniar a găsit conexiuni între „colonișți” (eng., „colonists”) și „Boston”, sau între „ajutor” (eng., „help”) și „provizii” (eng., „supplies”), în timp ce modelul MLP a identificat conexiunile dintre „britanici” și „Marea Britanie” ca un cuvânt compus.



(a) lanțuri lexicale

(b) lanțuri semantice folosind modelul liniar

(c) lanțuri semantice folosind modelul MLP

Figura 10. Vizualizări ale lanțurilor lexicale și semantice.

Predicția Implicării în CSCL

Mai mulți algoritmi de învățare automată au fost testați folosind caracteristicile introduse anterior pentru a evalua performanțele lanțurilor semantice în evaluarea implicării elevilor în conversații. Deoarece setul de date conținea 10 conversații distincte, a fost efectuată o validare încrucișată de 10 ori, lăsând un chat în afara pentru testare în fiecare dosar. Două clase diferite, și anume participarea (spre exemplu, reflectând implicarea activă) și colaborarea (spre exemplu, interacțiunile cu colegii) notate manual între 1 și 10 au fost prezise cu două modele separate.

Performanța umană obținută pe acest set de date arată un MAE mai mic de 1 din 10, indicând faptul că evaluatorii erau apropiați unul de altul, dar nu erau perfect de acord între ei (în mod inerent, o parte era mai relaxată, în timp ce cealaltă era mai fastidioasă); ca atare, s-au adunat 4 evaluări pentru fiecare participant, iar sistemul prezice ratingul mediu al evaluatorilor atât pentru participare, cât și pentru colaborare. Algoritmul Random Forest a fost cel mai predictiv atingând un MAE de 0,55; scăderea numărului de arbori generați are un impact benefic asupra performanței, având în vedere numărul limitat de caracteristici și exemple. În general, modelele de învățare automată par să capteze mai bine scorurile medii ale evaluatorilor, având avantajul generalizării tuturor participanților.

O analiză ulterioară a fost efectuată pentru a înțelege importanța fiecărei caracteristici pentru cele două sarcini diferite.

Prezicerea Scorurilor Eseurilor

Lanțurile semantice nu pot fi evaluate direct în absența unui set de date adnotat, așa că am decis să le evaluăm indirect pentru o altă sarcină. Presupunerea noastră a fost că lanțurile semantice pot reflecta mai bine coeziunea textului și calitatea eseurilor scrise. Prin urmare, am folosit setul de date Automated Student Assessment Prize (ASAP) (Kaggle Inc, 2021) pentru a evalua lanțurile semantice. Având în vedere că partițiile de validare și testare ale setului de date nu mai sunt disponibile, modelele au fost evaluate cu validări încrucișate de 5 ori pe partiția de antrenament. Pentru a reduce varianța rezultatelor și influența greutăților inițiale, scorurile au fost mediate pe 10 probe diferite.

Au fost definite mai multe caracteristici pe baza lanțurilor semantice extrase, unele dintre ele inspirate din setul LEX-1 (Somasundaran, Burstein, & Chodorow, 2014). Scopul a fost de a evalua precizia de predicție a trăsăturilor derivate din lanțurile semantice și descriptive ale coeziunii textului în sarcina de evaluare automată a textului. Caracteristicile noastre sunt descriptive atât pentru coeziunea textului local, cât și global, pe măsură ce lanțurile semantice se întind pe text, dar și pentru fluxul semantic (O'Rourke & Calvo, 2009), luând în considerare recurența și suprapunerea vocilor.

Diferite modele de predicție a legăturilor și pragurile de filtrare au fost comparate cu un model de bază bazat pe algoritmul de construcție a lanțurilor lexicale Galley și McKeown (Galley & McKeown, 2003). Aceleași caracteristici și algoritm de clasificare au fost utilizate pentru toate metodele de descoperire a lanțului. Prin urmare, diferențele de performanță ale modelelor ar trebui să depindă doar de calitatea lanțurilor semantice extrase.

O rețea neuronală simplă cu un singur strat ascuns de dimensiunea 16 a fost antrenată folosind caracteristicile descrise anterior pentru a prezice nota acordată eseurilor. Au fost inspectate vizual mai multe valori prag pentru filtrarea perechilor de cuvinte. Utilizarea pragurilor mici duce la componente mari conectate și nu separă bine vocile din text, astfel încât sunt prezentate doar rezultatele obținute cu praguri mari. Cea mai bună eroare medie absolută (MAE) a fost obținută cu modelul MLP, în timp ce modelul liniar a obținut cel mai bun scor R^2; ambele modele au folosit un prag de 0.9. Cu toate acestea, modelele reprezintă doar aproximativ 20% din varianță; acest lucru este de înțeles, deoarece fluxul este doar o parte din ceea ce este prezentat într-un eseu.

8 Discuții

8.1 Avantajele Abordării Noastre

Studiile de caz prezentate în această teză și care încorporează cele trei obiective prezentate în capitolul 1.2, se bazează pe tehnici NLP avansate și se concentrează pe următoarele sarcini: identificarea și modelarea stilului de scriere în texte, evaluarea performanței elevilor în mediile online de învățare, identificarea tiparelor de interacțiune între elevi, prezicerea notele elevilor pe baza activității lor online, identificarea lanțurilor semantice în cadrul textelor, evaluarea colaborării pe baza lanțurilor semantice în conversațiile chat.

Pentru a sublinia avantajele abordărilor noastre, prezentăm pentru fiecare întrebare de cercetare din capitolul 1.2, metodele propuse și rezultatele obținute.

- **RQ1:** În ce măsură indicii de complexitate textuală sunt predictivi din prisma diferențelor în stilul de scriere atunci când se efectuează analize multilingve pe diferite corpuri și tipuri de discurs (spre exemplu, documente oficiale NATO și mass-media din România)?

Complexitatea textuală generată de ReaderBench oferă atât informații despre proprietățile numerice ale diferitelor caracteristici textuale, cât și feature-uri desprinse din textul, cum ar fi lizibilitatea, coeziunea locală sau globală sau complexitatea cuvintelor. Capitolul 5 prezintă două studii de caz care investighează stilul de scriere în discursurile mass-media din România despre criza economică și documentele oficiale NATO, bazate pe indici de complexitate textuală generați de ReaderBench. Indicii de complexitate textuală generați pentru fiecare studiu de caz au fost folosiți pentru explorarea diferențelor dintre cele patru stiluri de scriere. Mai mult, indicii de complexitate textuală care acoperă trăsături lexicale, semantice și coezive au fost utilizați în analizele statistice pentru a evidenția diferențele în stilul de scriere. Pe lângă distanțele semantice WordNet, folosim modelul semantic word2vec.

Analizele ale variației au fost efectuate pentru a identifica indicii de complexitate textuală care au prezentat cele mai mari diferențe semnificative în stilurile de scriere. Ulterior, s-a efectuat o analiză funcțională discriminantă pas cu pas pentru ambele studii de caz.

Pentru mass-media din România, s-au efectuat analize statistice pentru a examina diferențele în stilurile de scriere din patru tipuri de documente (academice specializate, non-academice specializate, nespecializate neacademice, informale). Pentru documentele NATO, s-au efectuat analize statistice pentru a investiga diferențele în stilurile de scriere ale discursurilor NATO în funcție de tipul și perioada de timp în care au fost produse (integrative și contradictorii). Rezultatele relevă diferențe

semnificative statistic și interesante în ceea ce privește gradul de elaborare a cuvintelor (lungimea și numărul polisemiei), numărul de verbe unice pe paragraf, scorul Age of Acquisition (AoA) per paragraf, precum și numărul de cuvinte din cadrul unor liste multiple.

- **RQ2:** În ce măsură feature-urile generate de platforma ReaderBench, inclusiv feature-uri derivate din graful CNA aplicat pe discuțiile de tip forum, indicii de complexitate textuală, analiza longitudinală și feature-urile derivate din datele click-stream, sunt capabile să prezică performanța studenților?

Capitolul 6 prezintă două studii de caz care se concentrează pe prezicerea succesului matematic și a notelor elevilor pe baza participării lor online în două medii online: MOOC și Moodle. Secvența noastră de procesare integrează indici de analiză a rețelei de coeziune, indici de analiză longitudinală, indici de complexitate textuală și date click-stream pentru a prezice notele elevilor utilizând diverși algoritmi de învățare automată. Toți indicii anteriori oferă informații valoroase în ceea ce privește participarea și colaborarea (indicii CNA), stilometria (indicii de complexitate textuală), activitatea online (date click-stream), precum și regularitatea și evoluția în timp (analiza longitudinală).

Modelele liniare au indicat faptul că succesul matematic era legat de zilele petrecute pe forum și de studenții care postau mai regulat pe forum.

- **RQ3:** Care sunt diferențele dintre comportamentele și interacțiunile studenților înainte și în timpul pandemiei COVID-19 în două instanțe anuale consecutive ale unui curs universitar?

Folosind indicii de analiză a rețelei de coeziune, au fost generate diverse vizualizări interactive pentru a evidenția interacțiunea dintre participanți și pentru a descrie evoluția comunității de la o săptămână la alta. Sunt redate mai multe tipuri de vizualizări (spre exemplu, force-directed graph, hierarchical edge bundling, radar chart, parallel coordinates chart) pentru a descrie evoluția, comportamentele și tiparele de interacțiune ale elevilor folosind biblioteca d3.js. Mai mult, am examinat ce cuvinte cheie sunt cele mai frecvente în discuțiile elevilor și reprezentăm conceptele folosind o hartă de concepte. Toate vizualizările sunt prezentate în Capitolul 6.

Datorită pandemiei COVID-19, tranziția de la fizic la complet online a fost drastică, adaptarea a fost necesară într-un timp scurt, iar mediile de învățare inteligente au facilitat această tranziție, sprijinind în același timp atât elevii, cât și profesorii. Întrucât totul s-a mutat online, discuțiile față în față la curs au trecut la forumuri, chat-uri și întâlniri video. Capitolul 6 prezintă o analiză a activității online a studenților într-un curs Moodle timp de doi ani, înainte și în timpul pandemiei COVID-19. Comportamentele și interacțiunile studenților sunt comparate înainte și în timpul COVID-19 folosind două instanțe anuale consecutive ale unui curs universitar de proiectare a algoritmilor,

desfășurat în limba română folosind Moodle. Comportamentele și interacțiunile studenților sunt comparate în termeni de diferențe derivate din gama noastră largă de indici, luând în considerare, de asemenea, vizualizări interactive comparative care ilustrează interacțiunile lor și oferă o analiză calitativă din perspectiva tutorelui. Rezultatele au arătat că pandemia COVID-19 a generat un dezechilibru, o creștere drastică a participării, urmată de o scădere spre sfârșitul semestrului, comparativ cu anul universitar 2018-2019 când s-au observat fluctuații mai mici de participare.

- **RQ4:** În ce măsură lanțurile semantice pot fi operaționalizate folosind modele de limbă predictive atât din prisma colaborării (spre exemplu, evaluarea implicării participanților în cadrul unei conversații), cât și în studiul individual (spre exemplu, evaluarea automată a eseurilor)?

Modelele bazate pe arhitectura de tip Transformer (spre exemplu, BERT) oferă informații despre importanța și relațiile dintre cuvinte, analizând valorile atenției. Astfel, am introdus și evaluat o metodă care utilizează informațiile contextuale capturate de BERT pentru a găsi legăturile semantice, luând în considerare doar greutățile de atenție de la diferite capete (Capitolul 7). Modelul rezultat se generalizează la relații multiple, inclusiv repetări, concepte legate semantic din WordNet, precum și rezoluții pronominale. Spre deosebire de rezultatele anterioare, noua noastră metodă identifică mai multe lanțuri semantice cu mai multe cuvinte înrudite și include rezoluția coreferenței (spre exemplu, rezoluția pronumelui).

Folosind lanțurile semantice identificate cu informații contextuale capturate de BERT, am testat mai mulți algoritmi de învățare automată pentru a evalua măsura în care lanțurile semantice sunt predictive pentru implicarea elevilor în conversațiile chat. Raportul lanțurilor, care indică acoperirea lanțurilor semantice utilizate de un participant în raport cu întreaga conversație, este de departe cea mai predictivă caracteristică. Continuările, care reflectă legăturile dintre doi participanți diferiți, împreună cu numărul și raporturile lanțurilor acoperite de participant sunt cei mai buni predictori pentru colaborare. Mai mult decât atât, modelul nostru a fost utilizat cu succes pentru evaluarea automată a eseurilor. Scopul nostru a fost de a evalua acuratețea de predicție a caracteristicilor derivate din lanțurile semantice și descriptive ale coeziunii textului în sarcina de notare automată a textului.

8.2 Probleme Întâmpinate și Soluții Furnizate

O limitare potențială a studiului cu privire la performanța elevilor din cadrul cursului Moodle (Capitolul 6.1.2) presupune generalizarea modelului. Un factor care trebuie luat în considerare se referă la studenții care au prezentat comportamente ascunse și nu au avut postări active. Acești elevi nu au putut fi incluși în analize, deoarece nu au existat urme textuale de analizat. Acest factor poate

influența acuratețea predicțiilor modelului atunci când se ia în considerare numărul total de studenți. Astfel, mecanismele suplimentare centrate pe analiza jurnalelor (spre exemplu, accesul la postări specifice, ratele de finalizare a temelor) trebuie luate în considerare pentru a construi modele de bază capabile să genereze predicții pentru această categorie de studenți. Mai mult, experimentele actuale trebuie să fie aplicate pe un interval de timp mai mare și pe cursuri suplimentare pentru a crea modele generalizabile în diferite subiecte și situații ale cursului. Perioade mai lungi de timp (adică mai multe rate) și cursuri suplimentare vor fi luate în considerare pentru a crea modele predictive cu un grad mai mare de generalizare, care este greu, dacă nu imposibil, de obținut cu modificările induse de pandemie. Cu toate acestea, trebuie luate în considerare specificitățile fiecărui curs, la nivel de subiecte, activități și modele de interacțiune, în timp ce se construiesc modele de predicție.

O limitare a abordării noastre în ceea ce privește explorarea dialogismului folosind modele de limbă (Capitolul 7) este raportul mic de legături de coreferență din setul de date generat, care este vizibil și în lanțurile semantice construite. O modalitate de a îmbunătăți acest lucru în viitor va echilibra regulile utilizate pentru a genera setul de date și va include mai multe.

8.3 Implicații Educaționale

Sociogramele generate prezentate în capitolul 6 pot fi utilizate de profesori pentru a urmări evoluția elevilor în termeni de interacțiuni, interactivitate și participare online, permițându-le astfel să intervină atunci când observă o scădere a participării sau a inactivității. Șansele elevilor de a trece sau de a obține o notă mai bună pot fi crescute dacă profesorii îi încurajează să fie mai implicați pe tot parcursul cursului. Pe măsură ce mai multe instituții de învățământ s-au mutat online, iar activitățile față în față au fost reduse drastic, un mecanism care ține evidența activităților elevilor și a probabilității lor de a obține note bune ar fi benefic atât studenților, cât și profesorilor.

Mai mult, vizualizările pot fi utilizate de instructori și cercetători pentru a înțelege mai bine participarea și colaborarea în medii de învățare pe scară largă. Vizualizările săptămânale pot fi folosite pentru a înțelege mai bine evoluția elevilor de-a lungul cursului, în timp ce activitățile lor sunt corelate cu evenimente specifice cursului (spre exemplu, termene limită pentru teme, teste, vacanțe, examen etc.).

9 Concluzii

9.1 Contribuții Personale

A Modelarea Stilului de Scriere folosind Indici de Complexitate Textuală

Au fost aduse următoarele contribuții la modelarea stilului de scriere prin indicii de complexitate textuală:

- Analiza detaliată a stilului de scriere în discursurile mass-media din România despre criza economică evidențiază diferențe semnificative în ceea ce privește caracteristicile lingvistice;
- Identificarea și compararea caracteristicilor lingvistice specifice discursurilor mass-media din România despre criza economică din anii 2008 - 2018;
- Au fost efectuate analize detaliate ale discursului militar dintr-o perspectivă diacronică axată pe o analiză lingvistică aprofundată a diferențelor dintre stilurile de scriere a documentelor oficiale NATO;
- A fost efectuată o investigație cantitativă a evoluției limbajului documentelor oficiale NATO de-a lungul a două perioade diferite - Războiul Rece (1949 - 1990) și perioada postbelică (1991 - 2018) -, cu accent pe examinarea discursului corespunzător puterii dinamice;
- A fost descris și explicat modul în care două relații de putere predominante - integrative și contradictorii - au fost reificate în discursul Alianței pe o perioadă de șaptezeci de ani, strâns corelată cu contextul istoric și politic.

B Evaluarea Performanței Online a Elevilor folosind CNA

Au fost aduce următoarele contribuții în ceea ce privește evaluarea performanței online a elevilor folosind CNA:

- Examinarea succesului studenților în cadrul unui curs mixt de licență care utilizează CNA, controlând în același timp diferențele individuale și variabilele generate din datele de tip click-stream;
- Evaluarea comportamentelor și tiparelor de interacțiune ale studenților într-un curs de licență de proiectarea algoritmului oferit la Universitatea Politehnica din București, atât pentru anii universitari 2018 - 2019, cât și 2019 - 2020, fiecare reflectând condiții și cerințe externe diferite, înainte și în timpul pandemiei COVID – 19;

- Performanța estimată a elevilor utilizând un model de rețea neuronală recurentă care ia în considerare analizele seriilor de timp pe perioade de timp, pe lângă indicii CNA, complexitatea textuală și analiza longitudinală;
- Generarea diverselor sociograme pentru a arăta interacțiunea dintre participanți, pe baza indicilor CNA;
- Generarea sociogramelor săptămânale pentru a examina modul în care interacțiunile dintre studenți și tutorii evoluează de la o săptămână la alta. Aceste vizualizări oferă informații despre comportamentele studenților în asociere cu evenimentele cursului, cum ar fi termenele limită, sarcinile, testele și examenele;
- Examinarea proprietăților lingvistice ale forumurilor de discuții online folosind CNA;
- Identificarea tendințelor temporale în rândul studenților;
- Introducerea unei analize longitudinale și vizualizări detaliate ale participării, utile pentru a obține o mai bună înțelegere a modului în care elevii interacționează în timpul unui curs de matematică online.

C Analiza Dialogismului folosind Modele de limbă

Au fost aduse următoarele contribuții privind analiza dialogismului folosind modele de limbă:

- A fost introdusă și evaluată o metodă nouă de identificare a lanțurilor semantice utilizând modele de limbă de ultimă generație, și anume BERT;
- Lanțurile semantice sunt generalizate la relații multiple, inclusiv repetiții, concepte semantic legate de WordNet (spre exemplu, sinonime, hipernimi, hiponime), precum și rezoluții pronominale;
- A fost introdusă o metodă bazată pe dialogism pentru evaluarea implicării elevilor în conversațiile de chat bazate pe lanțuri semantice calculate folosind modele de limbă;
- Evaluarea performanței de predicție a feature-urilor derivate din lanțurile semantice în sarcina de notare automată a textului;
- A fost introduse vizualizări interactive pentru a evidenția atât o propagare longitudinală a vocilor, cât și o suprapunere transversală a lanțurilor semantice între participanții la discuții;
- Au fost introduse vizualizării interactive pentru a vedea lanțurile semantice dintr-un document care este împărțit în paragrafe și ulterior în propoziție.

D Contribuții Suplimentare

În plus față de contribuțiile menționate mai sus cu privire la studiile empirice, numeroase contribuții au fost făcute în diferite studii:

- Crearea unui nou mediu de învățare inteligent pentru a genera feedback personalizat pentru scrierile studenților atât pentru limba engleză (Botarleanu, Dascalu, Sirbu, Crossley, & Trausan-Matu, 2018; Sirbu, Botarleanu, Dascalu, Crossley, & Trausan-Matu, 2018), cât și pentru limba română (Sirbu et al., 2018b);
- Analiza interacțiunii participanților pe două tipuri de comunități: OKBC și comunități de jocuri, pentru a observa trăsături importante și modele specifice de interacțiune (Sirbu et al., 2017);
- Prezicerea notele de curs pe baza participării studenților și evaluarea interacțiunilor dintre participanți și impactul termenelor și testelor pentru contribuțiile lor online prin intermediul sociogramelor interactive într-un curs Moodle centrat pe Utilizarea sistemelor de operare pentru studenții din anul I (M. Dascalu et al., 2018; M.-D. Dascalu et al., 2020; Gutu-Robu et al., 2018);
- Efectuarea unei analize a unei comunități de dezbateri de la Reddit, care se concentrează pe discutarea ideilor politice și a credințelor legate de socialism și comunism, pentru a modela interacțiunea dintre participanți, a analiza comportamentele utilizatorilor pe baza măsurilor de regularitate, a extrage subiectele cele mai discutate, și prezice rangurile utilizatorilor pe baza participării lor (Fetoiu et al., 2020);
- Introducerea tehnicilor de prelucrare a textului și de regăsire a informațiilor pentru implementarea unei platforme inteligente pentru căutarea avansată a medicamentelor din România și a interacțiunilor dintre acestea (M.-D. Dascalu et al., 2019; Ene, Sirbu, Dascalu, Trausan-Matu, & Nuta, 2019; Paraschiv et al., 2019);
- Construirea unei baze de cunoștințe pentru administrarea medicamentelor din România (Nicula, Dascalu, Sirbu, Trausan-Matu, & Nuta, 2019);
- Extinderea structurii grafului CNA (M.-D. Dascalu, Ruseti, Dascalu, McNamara, & Trausan-Matu, 2020) cu legături suprapuse lexicale de mai multe tipuri, împreună cu legături coreferențiale pentru evidențierea dependențelor dintre fragmentele de text de diferite granularități. Au fost introduse două vizualizări ale grafului CNA care susțin explorarea vizuală a legăturilor intra-textuale și inter-textuale.

9.2 Direcții pentru Cercetări Viitoare

În ceea ce privește studiile viitoare privind evaluarea performanței online a studenților, scopul nostru este de a introduce evaluări în clasă pentru evaluare impactului utilizării instrumentelor automate pe parcursul anului universitar care permit reacții în timp util din partea tutorilor, spre deosebire de evaluarea a posteriori. În plus, dorim să introducem configurații personalizate care iau în considerare modele semantice specifice adaptate pentru tema fiecărui curs. Un alt obiectiv general este de a oferi

feedback automat atât studenților, cât și profesorilor cu privire la implicarea studenților în cadrul curs. În acest scop, CNA va fi extins astfel încât tutorii să poată evalua periodic evoluția studenților lor, să identifice studenții cu risc și să ia măsuri în timp util. Sunt prevăzute alte modificări în care elevii pot fi stimulați să se implice activ și să colaboreze mai mult cu colegii lor. În cele din urmă, obiectivul este de a maximiza șansele de succes ale cursanților.

În plus, ne gândim să aplicăm modelul dialogic pe alte seturi de date și să efectuăm analize conexe. Spre exemplu, dialogismul poate fi utilizat pentru a monitoriza implicarea studenților în cursurile online și feature-urile introduse pot fi utilizate pentru a prezice abandonul sau notele cursului. Și mai mult, prin evaluarea lanțurilor semantice introduse, discuțiile și contribuțiile corespunzătoare pot fi catalogate ca fiind specifice cursului, administrative sau off-topic, pot fi introduse mecanisme de îndrumare automate, în timp ce vizează stimularea creativității. Mai mult, dorim să extindem în continuare acest model cu caracteristici de analiză a sentimentelor derivate din contexte locale capturate de BERT, îmbogățind astfel analiza cu identificarea punctelor de vedere convergente și divergente.

Lista de Publicații

Articole ISI

Conferințe de top (categoria A în portalul de clasare CORE)

1. Ruseti, S., Dascalu, M.-D., Corlatescu, D.-G., Dascalu, M., Trausan-Matu, S., & McNamara, D. S. (in press). *Exploring Dialogism using Language Models*. In 22nd Int. Conf. on Artificial Intelligence in Education (AIED 2021). Utrech, Netherlands (Online): Springer.
2. Dascalu, M.-D., Ruseti, S., Dascalu, M., McNamara, D. S., & Trausan-Matu, S. (2020). *Multi-Document Cohesion Network Analysis: Visualizing Intratextual and Intertextual Links*. Paper presented at the 21st Int. Conf. on Artificial Intelligence in Education (AIED 2020), Online: Springer.
3. Dascalu, M.-D., Ruseti, S., Carabas, M., Dascalu, M., Trausan-Matu, S., & McNamara, D. S. (2020). *Cohesion Network Analysis: Predicting Course Grades and Generating Sociograms for a Romanian Moodle Course*. Paper presented at the 16h Int. Conf. on Intelligent Tutoring Systems (ITS 2020), Online: Springer. (Best Full Paper Award)
4. Samoilescu, R.-F., Dascalu, M., Sirbu, M.-D., Trausan-Matu, S., & Crossley, S. A. (2019). *Modeling Collaboration in Online Conversations using Time Series Analysis and Dialogism*. In 20th Int. Conf. on Artificial Intelligence in Education (AIED 2019) (pp. 458–468). Chicago, IL: Springer.
5. Crossley, S. A., Sirbu, M.-D., Dascalu, M., Barnes, T., Lynch, C. F., & McNamara, D. S. (2018). *Modeling Math Success using Cohesion Network Analysis*. In C. P. Rosé, R. Martínez-Maldonado, U. Hoppe, R. Luckin, M. Mavrikis, K. Porayska-Pomsta, B. McLaren & B. d. Boulay (Eds.), 19th Int. Conf. on Artificial Intelligence in Education (AIED 2018), Part II (pp. 63–67). London, UK: Springer.
6. Sirbu, M.-D., Dascalu, M., Crossley, S. A., McNamara, D. S., Barnes, T., Lynch, C. F., & Trausan-Matu, S. (2018). *Exploring Online Course Sociograms using Cohesion Network Analysis*. In C. P. Rosé, R. Martínez-Maldonado, U. Hoppe, R. Luckin, M. Mavrikis, K. Porayska-Pomsta, B. McLaren & B. d. Boulay (Eds.), 19th Int. Conf. on Artificial Intelligence in Education (AIED 2018), Part II (pp. 337–342). London, UK: Springer.

Alte Conferințe ISI

1. Dascalu, M.-D., Paraschiv, I. C., Nicula, B., Dascalu, M., Trausan-Matu, S., & Nuta, A. C. (2019). *Intelligent Platform for the Analysis of Drug Leaflets Using NLP Techniques*. In 18th Int.

Conf. on Networking in Education and Research (RoEduNet) (pp. 1–6). Galati, Romania: IEEE.

2. Ruseti, S., Sirbu, M.-D., Calin, M. A., Dascalu, M., Trausan-Matu, S., & Militaru, G. (2019). *Comprehensive Exploration of Game Reviews Extraction and Opinion Mining using NLP Techniques*. In 5th Int. Congress on Information and Communication Technology (pp. 323–331). London, UK: Springer.
3. Ene, O.-G., Sirbu, M.-D., Dascalu, M., Trausan-Matu, S., & Nuta, A. C. (2019). *PIAM – Intelligent Platform for Retrieving Relevant Information on Drugs Marketed in Romania*. In 4th Int. Workshop on Design and Spontaneity in Computer-Supported Collaborative Learning (DS-CSCL-2019), in conjunction with the 22nd Int. Conf. on Control Systems and Computer Science (CSCS22) (pp. 420–425). Bucharest, Romania: IEEE.
4. Stamati, D., Sirbu, M.-D., Dascalu, M., & Trausan-Matu, S. (2018). *Exploring General Morphological Analysis and Providing Personalized Recommendations to Stimulate Creativity with ReaderBench*. In M. Chang, E. Popescu, Kinshuk, N.-S. Chen, M. Jemni, R. Huang & J. M. Spector (Eds.), Int. Conf. on Smart Learning Environments (ICSLE 2018) (pp. 41–50). Beijing, China: Springer.
5. Sirbu, M.-D., Botarleanu, R., Dascalu, M., Crossley, S. A., & Trausan-Matu, S. (2018). *ReadME – Enhancing Automated Writing Evaluation*. In 18th Int. Conf. on Artificial Intelligence: Methodology, Systems, and Applications (AIMSA 2018) (pp. 281–285). Varna, Bulgaria: Springer.
6. Dascalu, M., Sirbu, M.-D., Gutu-Robu, G., Ruseti, S., Crossley, S. A., & Trausan-Matu, S. (2018). *Cohesion-Centered Analysis of Sociograms for Online Communities and Courses using ReaderBench*. In 13th European Conference on Technology Enhanced Learning (EC-TEL 2018) (pp. 622–626). Leeds, UK: Springer.
7. Sirbu, M.-D., Panaite, M., Secui, A., Dascalu, M., Nistor, N., & Trausan-Matu, S. (2017). *ReaderBench: Building Comprehensive Sociograms of Online Communities*. In 19th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC 2017) (pp. 225–231). Timisoara, Romania: IEEE.
8. Sirbu, M.-D., Secui, A., Dascalu, M., Crossley, S. A., Ruseti, S., & Trausan-Matu, S. (2016). *Extracting Gamers' Opinions from Reviews*. In 18th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC 2016) (pp. 227–232). Timisoara, Romania: IEEE.
9. Secui, A., Sirbu, M.-D., Dascalu, M., Crossley, S. A., Ruseti, S., & Trausan-Matu, S. (2016). *Expressing Sentiments in Game Reviews*. In 17th Int. Conf. on Artificial Intelligence:

Methodology, Systems, and Applications (AIMSA 2016) (pp. 352–355). Varna, Bulgaria: Springer.

10. Sirbu, M.-D., Dascalu, M., Gifu, D., Cotet, T.-M., Tosca, A., & Trausan-Matu, S. (2018). *ReadME – Improving Writing Skills in Romanian Language*. In V. Pais, D. Gifu, D. Trandabat, D. Cristea & D. Tufis (Eds.), 13th Int. Conf. on Linguistic Resources and Tools for Processing Romanian Language (ConsILR 2018) (pp. 135–145). Iasi, Romania.

Articole BDI

Conferințe de top (categoria A în portalul de clasare CORE)

1. Sirbu, M.-D., Dascalu, M., Crossley, S. A., McNamara, D. S., & Trausan-Matu, S. (2019). *Longitudinal Analysis of Participation in Online Courses Powered by Cohesion Network Analysis*. In 13th Int. Conf. on Computer-Supported Collaborative Learning (CSCL 2019) (pp. 640–643). Lyon, France: ISLS.

Alte Conferințe BDI

1. Fetoiu, C.-E., Dascalu, M.-D., Calin, M. A., Dascalu, M., Trausan-Matu, S., & Militaru, G. (2020). *Cohesion Network Analysis for Predicting User Ranks in Reddit Communities*. In 5th Int. Conf. on Smart Learning Ecosystems and Regional Development (SLERD 2020) (pp. 173–185). Online: Springer.
2. Paraschiv, I. C., Sirbu, M.-D., Nicula, B., Dascalu, M., Trausan-Matu, S., & Nuta, A. C. (2019). *Designing an Intelligent Platform for Drugs Administration*. In International Conference on Human-Computer Interaction (RoCHI2019) (pp. 53–59). Bucharest, Romania: MatrixRom.
3. Nicula, B., Dascalu, M., Sirbu, M.-D., Trausan-Matu, S., & Nuta, A. C. (2019). *Building a Comprehensive Romanian Knowledge Base for Drug Administration*. In G. Angelova, R. Mitkov, I. Nikolova & I. Temnikova (Eds.), Int. Conf. on Recent Advances in Natural Language Processing (RANLP 2019) (pp. 829–836). Varna, Bulgaria.
4. Florea, A.-M., Dascalu, M., Sirbu, M.-D., & Trausan-Matu, S. (2019). *Improving Writing for Romanian Language*. In 4th Int. Conf. on Smart Learning Ecosystems and Regional Development (SLERD 2019) (pp. 131–141). Rome, Italy: Springer.
5. Botarleanu, R.-M., Dascalu, M., Sirbu, M.-D., Crossley, S. A., & Trausan-Matu, S. (2018). *ReadME – Generating Personalized Feedback for Essay Writing using the ReaderBench Framework*. In H. Knoche, E. Popescu & A. Cartelli (Eds.), 3rd Int. Conf. on Smart Learning Ecosystems and Regional Development (SLERD 2018) (pp. 133–145). Aalborg, Denmark: Springer.
6. Dascalu, M.-D., Ruseti, S., Dascalu, M., McNamara, D. S., & Trausan-Matu, S. (in press). *Dialogism Meets Language Models for Evaluating Involvement in CSCL Conversations*. In 6th Int.

Conf. on Smart Learning Ecosystems and Regional Development (SLERD 2021) Bucharest, Romania (Online).

Jurnale

Jurnale ISI

1. Dascalu, M.-D., Ruseti, S., Dascalu, M., McNamara, D. S., Carabas, M., Rebedea, T., & Trausan-Matu, S. (2021). *Before and during COVID-19: A Cohesion Network Analysis of Students' Online Participation in Moodle Courses*. Computers in Human Behavior, 121 (**Q1 Journal**).
2. Dragomir, I. A., Dascalu, M.-D., Dascalu, M., Terian, S.-M., & Trausan-Matu, S. (2020). *Exploring Differences in NATO Discourses using the ReaderBench Framework*. Scientific Bulletin, University Politehnica of Bucharest, Series C, 82(1), 3–18.
3. Terian, S.-M., Cotet, T.-M., Sirbu, M.-D., Dascalu, M., & Trausan-Matu, S. (2019). *Discourses of Economic Crisis in Romanian Media: An Automated Analysis using the ReaderBench Framework*. Transylvanian Review, XXVIII (Supplement No. 1), 245–262.

Alte Journale

1. Gutu-Robu, G., Sirbu, M.-D., Paraschiv, I. C., Dascalu, M., Dessus, P., & Trausan-Matu, S. (2018). *Liftoff - ReaderBench introduces new online functionalities*. Romanian Journal of Human - Computer Interaction, 11(1), 76–91.

Cereri de brevet

1. Trausan-Matu, S., Dascalu, M., Dascalu, M.-D., Paraschiv, I. C., Nicula, B., Nuta, A. C., & Cioroboiu, G.-C. (2019). Romania Patent No. Cererea A/00877 / 09.12.2019. OSIM.

Referințe

Achieve Virtual. (n.d.). Infographic: Evolution of Virtual Education. Retrieved September 1st, 2020, from https://achievevirtual.org/infographic-evolution-of-virtual-education/

Arif, T. (2015). The mathematics of social network analysis: metrics for academic social networks. *International Journal of Computer Applications Technology and Research, 4*(12), 889-893.

Attali, Y., & Burstein, J. (2004). Automated essay scoring with e-rater V.2.0. In *Annual Meeting of the International Association for Educational Assessment* (pp. 23). Philadelphia, PA: Association for Educational Assessment.

Bakhtin, M. M. (1981). *The dialogic imagination: Four essays* (C. Emerson & M. Holquist, Trans.). Austin, TX, USA and London, UK: The University of Texas Press.

Bakhtin, M. M. (1984). *Problems of Dostoevsky's poetics* (C. Emerson, Trans. C. Emerson Ed.). Minneapolis: University of Minnesota Press.

Bakhtin, M. M. (1986). *Speech genres and other late essays* (V. W. McGee, Trans.). Austin: University of Texas.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research, 3*(4-5), 993–1022.

Bolter, J. D. (1984). *Turing's man: Western culture in the computer age.* Chapel Hill, North Carolina, USA: UNC Press Books.

Bostock, M. (2021). Data-Driven Documents. Retrieved July, 22nd, 2021, from https://d3js.org/

Botarleanu, R.-M., Dascalu, M., Sirbu, M.-D., Crossley, S. A., & Trausan-Matu, S. (2018). ReadME – Generating Personalized Feedback for Essay Writing using the ReaderBench Framework. In H. Knoche, E. Popescu & A. Cartelli (Eds.), *3rd Int. Conf. on Smart Learning Ecosystems and Regional Development (SLERD 2018)* (pp. 133–145). Aalborg, Denmark: Springer.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems, 30*, 1–7.

Britannica, E. (2021). Britannica – University of Phoenix. Retrieved June, 9th, 2021, from https://www.britannica.com/topic/University-of-Phoenix

Cambridge Intelligence. (2021). Social Network Analysis. Retrieved May, 30th, 2021, from https://cambridge-intelligence.com/social-network-analysis/

Chowdhury, G. G. (2003). Natural language processing. *Annual Review of Information Science and Technology, 37*(1), 51-89.

Referințe

College, H. (2021). The Writing Process. Retrieved July, 14th, 2021, from http://www.hunter.cuny.edu/rwc/handouts/the-writing-process-1/invention/Five-Qualities-of-Good-Writing

Cress, U. (2013). Mass collaboration and learning. In R. Luckin, S. Puntambekar, P. Goodyear, B. Grabowski, J. Underwood & N. Winters (Eds.), *Handbook of design in educational technology* (pp. 416–424). New York: Routledge.

Crossley, S. A., Barnes, T., Lynch, C., & McNamara, D. S. (2017). Linking language to math success in a blended course. In *10th Int. Conf, on Educational Data Mining (EDM)* (pp. 180–185). Wuhan, China.

Crossley, S. A., Kyle, K., & Dascalu, M. (2019). The Tool for the Automatic Analysis of Cohesion 2.0: Integrating Semantic Similarity and Text Overlap. *Behavior research methods, 51*(1), 14–27.

Crossley, S. A., Kyle, K., & McNamara, D. S. (2015). To aggregate or not? Linguistic features in automatic essay scoring and feedback systems. *Journal of Writing Assessment, 8*(1), 19.

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior research methods, 48*(4), 1227–1237.

Crossley, S. A., Paquette, L., Dascalu, M., McNamara, D. S., & Baker, R. S. (2016). Combining Click-Stream Data with NLP Tools to Better Understand MOOC Completion. In *6th Int. Conf. on Learning Analytics & Knowledge (LAK '16)* (pp. 6–14). Edingurgh, UK: ACM.

Crossley, S. A., Sirbu, M.-D., Dascalu, M., Barnes, T., Lynch, C. F., & McNamara, D. S. (2018). Modeling Math Success using Cohesion Network Analysis. In C. P. Rosé, R. Martínez-Maldonado, U. Hoppe, R. Luckin, M. Mavrikis, K. Porayska-Pomsta, B. McLaren & B. d. Boulay (Eds.), *19th Int. Conf. on Artificial Intelligence in Education (AIED 2018), Part II* (pp. 63–67). London, UK: Springer.

Dale, E., & Chall, J. S. (1949). The concept of readability. *Elementary English, 26*(1), 19-26.

Das, K., Samanta, S., & Pal, M. (2018). Study on centrality measures in social networks: a survey. *Social network analysis and mining, 8*(1), 1-11.

Dascalu, M. (2014). *Analyzing discourse and text complexity for learning and collaborating, Studies in Computational Intelligence* (Vol. 534). Switzerland: Springer.

Dascalu, M., Crossley, S. A., McNamara, D. S., Dessus, P., & Trausan-Matu, S. (2018). Please ReaderBench this Text: A Multi-Dimensional Textual Complexity Assessment Framework. In S. Craig (Ed.), *Tutoring and Intelligent Tutoring Systems* (pp. 251–271). Hauppauge, NY, USA: Nova Science Publishers, Inc.

Dascalu, M., Dessus, P., Bianco, M., Trausan-Matu, S., & Nardy, A. (2014). Mining texts, learner productions and strategies with ReaderBench. In A. Peña-Ayala (Ed.), *Educational Data Mining: Applications and Trends* (pp. 345–377). Cham, Switzerland: Springer.

Referințe

Dascalu, M., McNamara, D. S., Trausan-Matu, S., & Allen, L. K. (2018). Cohesion Network Analysis of CSCL Participation. *Behavior research methods, 50*(2), 604–619. doi: 10.3758/s13428-017-0888-4

Dascalu, M., Sirbu, M. D., Gutu-Robu, G., Ruseti, S., Crossley, S. A., & Trausan-Matu, S. (2018). Cohesion-Centered Analysis of Sociograms for Online Communities and Courses using ReaderBench. In *13th European Conference on Technology Enhanced Learning (EC-TEL 2018)* (pp. 622–626). Leeds, UK: Springer.

Dascalu, M., Trausan-Matu, S., & Dessus, P. (2013). Voices' inter-animation detection with ReaderBench – Modelling and assessing polyphony in CSCL chats as voice synergy. In *2nd Int. Workshop on Semantic and Collaborative Technologies for the Web, in conjunction with the 2nd Int. Conf. on Systems and Computer Science (ICSCS)* (pp. 280–285). Villeneuve d'Ascq, France: IEEE.

Dascalu, M., Trausan-Matu, S., McNamara, D. S., & Dessus, P. (2015). ReaderBench – Automated Evaluation of Collaboration based on Cohesion and Dialogism. *International Journal of Computer-Supported Collaborative Learning, 10*(4), 395–423.

Dascalu, M.-D., Paraschiv, I. C., Nicula, B., Dascalu, M., Trausan-Matu, S., & Nuta, A. C. (2019). Intelligent Platform for the Analysis of Drug Leaflets Using NLP Techniques. In *18th Int. Conf. on Networking in Education and Research (RoEduNet)* (pp. 1–6). Galati, Romania: IEEE.

Dascalu, M.-D., Ruseti, S., Carabas, M., Dascalu, M., Trausan-Matu, S., & McNamara, D. S. (2020). Cohesion Network Analysis: Predicting Course Grades and Generating Sociograms for a Romanian Moodle Course. In *16h Int. Conf. on Intelligent Tutoring Systems (ITS 2020)*. Online: Springer.

Dascalu, M.-D., Ruseti, S., Dascalu, M., McNamara, D. S., Carabas, M., Rebedea, T., & Trausan-Matu, S. (2021). Before and during COVID-19: A Cohesion Network Analysis of Students' Online Participation in Moodle Courses. *Computers in Human Behavior, 121*.

Dascalu, M.-D., Ruseti, S., Dascalu, M., McNamara, D. S., & Trausan-Matu, S. (2020). Multi-Document Cohesion Network Analysis: Visualizing Intratextual and Intertextual Links. In *21st Int. Conf. on Artificial Intelligence in Education (AIED 2020)*. Online: Springer.

Davies, M. (2010). The Corpus of Contemporary American English as the First Reliable Monitor Corpus of English. *Literary and Linguistic Computing, 25*(4), 447–465.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NACL 2018)*. Minneapolis, Minnesota, USA: ACL.

Dragomir, I. A., Dascalu, M.-D., Dascalu, M., Terian, S.-M., & Trausan-Matu, S. (2020). Exploring Differences in NATO Discourses using the ReaderBench Framework. *Scientific Bulletin, University Politehnica of Bucharest, Series C, 82*(1), 3–18.

DuBay, W. H. (2004). *The Principles of Readability*. Costa Mesa, CA, USA: Impact Information.

Referințe

Dumais, S. T. (2004). Latent semantic analysis. *Annual Review of Information Science and Technology, 38*(1), 188–230.

Ene, O.-G., Sirbu, M.-D., Dascalu, M., Trausan-Matu, S., & Nuta, A. C. (2019). PIAM – Intelligent Platform for Retrieving Relevant Information on Drugs Marketed in Romania. In *4th Int. Workshop on Design and Spontaneity in Computer-Supported Collaborative Learning (DS-CSCL-2019), in conjunction with the 22nd Int. Conf. on Control Systems and Computer Science (CSCS22)* (pp. 420–425). Bucharest, Romania: IEEE.

Explosion. (2016-2021). spaCy. Retrieved July, 27th, 2021, from https://spacy.io

Fetoiu, C.-E., Dascalu, M.-D., Calin, M. A., Dascalu, M., Trausan-Matu, S., & Militaru, G. (2020). Cohesion Network Analysis for Predicting User Ranks in Reddit Communities. In *5th Int. Conf. on Smart Learning Ecosystems and Regional Development (SLERD 2020)* (pp. 173–185). Bucharest, Romania (Online): Springer.

Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, 35-41.

Gabriel, C., Hahne, C., Zimmermann, A., & Lenk, F. (2021). The Virtual Tutor: Tasks for conversational agents in Online Collaborative Learning Environments. In *Proceedings of the 54th Hawaii International Conference on System Sciences* (pp. 104–113). Hawaii, USA: Online http://hdl.handle.net/10125/70623.

Galley, M., & McKeown, K. (2003). Improving word sense disambiguation in lexical chaining. In G. Gottlob & T. Walsh (Eds.), *18th International Joint Conference on Artificial Intelligence (IJCAI'03)* (pp. 1486–1488). Acapulco, Mexico: Morgan Kaufmann Publishers, Inc.

Google Inc. (2010-2021). Angular website. Retrieved July, 22nd, 2021, from https://angular.io

Gutu-Robu, G., Sirbu, M.-D., Paraschiv, I. C., Dascalu, M., Dessus, P., & Trausan-Matu, S. (2018). Liftoff - ReaderBench introduces new online functionalities. *Romanian Journal of Human - Computer Interaction, 11*(1), 76–91.

Hansen, D. L., Shneiderman, B., Smith, M. A., & Himelboim, I. (2011). Social network analysis: measuring, mapping, and modeling collections of connections. In B. Shneiderman, D. Hansen & M. A. Smith (Eds.), *Analyzing social media networks with NodeXL* (pp. 31–51). Amsterdam, Netherlands: Elsevier.

Hiltz, S. R., & Turoff, M. (2005). Education goes digital: The evolution of online learning and the revolution in higher education. *Communications of the ACM, 48*(10), 59–64.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation, 9*(8), 1735–1780.

Holmes, D. I. (1998). The evolution of stylometry in humanities scholarship. *Literary and Linguistic Computing, 13*(3), 111-117.

John Snow Labs Inc. (2021). Spark NLP. Retrieved July, 5th, 2021, from https://nlp.johnsnowlabs.com

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science, 349*(6245), 255-260.

Referințe

Jurafsky, D., & Martin, J. H. (2008). *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics* (2nd ed.). London: Pearson Prentice Hall.

Kaggle Inc. (2021). Automated Student Assessment Prize. Retrieved July, 5th, 2021, from https://www.kaggle.com/c/asap-aes/overview/description

Klecka, W. R. (1980). Discriminant analysis *Quantitative Applications in the Social Sciences Series* (Vol. 19). Thousand Oaks, CA: Sage Publications.

Knoke, D., & Yang, S. (2019). *Social network analysis*: Sage Publications.

Koschmann, T. (1996). Paradigm shifts and instructional technology: an introduction. In T. Koschmann (Ed.), *CSCL, Theory and Practice of an Emerging Paradigm*. Mahwah, New Jersey, USA: Routledge.

Koschmann, T. (1999). Toward a dialogic theory of learning: Bakhtin's contribution to understanding learning in settings of collaboration. In C. M. Hoadley & J. Roschelle (Eds.), *Int. Conf. on Computer Support for Collaborative Learning (CSCL'99)* (pp. 308–313). Palo Alto: ISLS.

Kyle, K. (2016). Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication.

Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly, 49*, 757–786. doi: 10.1002/tesq.194

Kyle, K., Crossley, S. A., & Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): version 2.0. *Behavior research methods, 50*(3), 1030–1046. doi: 10.3758/s13428-017-0924-4

Kyle, K., Crossley, S. A., & Jarvis, S. (2020). Assessing the Validity of Lexical Diversity Indices Using Direct Judgements. *Language Assessment Quarterly*, 1-17.

Landauer, T. K., & Dumais, S. (2008). Latent semantic analysis. *Scholarpedia, 3*(11), 4356.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. *Psychological Review, 104*(2), 211–240.

LGPLv2.1, G. (2009-2021). Gensim. Retrieved July, 5th, 2021, from https://pypi.org/project/gensim/

Light, R. J. (2004). *Making the most of college: Students speak their minds*: Harvard University Press.

Linell, P. (2009). *Rethinking language, mind, and world dialogically: Interactional and contextual theories of human sense-making*. Information Age Publishing: Charlotte, NC.

Lipponen, L. (2002). Exploring foundations for computer-supported collaborative learning.

Liu, Z., Lin, Y., & Sun, M. (2020). Word Representation *Representation Learning for Natural Language Processing* (pp. 13-41): Springer.

Marková, I., Linell, P., Grossen, M., & Salazar Orvig, A. (2007). *Dialogue in focus groups: Exploring socially shared knowledge*. London, UK: Equinox.

McAuley, A., Stewart, B., Siemens, G., & Cormier, D. (2010). The MOOC model for digital practice.

Referințe

McNamara, D. S. (2004). SERT: Self-Explanation Reading Training. *Discourse Processes, 38*, 1–30.

McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). The linguistic features of quality writing. *Written Communication, 27*(1), 57–86.

McNamara, D. S., Crossley, S. A., & Roscoe, R. (2013). Natural language processing in an intelligent writing strategy tutoring system. *Behavior research methods, 45*(2), 499-515.

McNamara, D. S., Crossley, S. A., Roscoe, R., Allen, L. K., & Dai, J. (2015). A hierarchical classification approach to automated essay scoring. *Assessing Writing, 23*, 35–59.

McNamara, D. S., Graesser, A. C., & McCarthy, P. M. (2014). Coh-Metrix. Retrieved May, 26th, 2021, from http://cohmetrix.com

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representation in Vector Space. In *Workshop at ICLR*. Scottsdale, AZ.

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM, 38*(11), 39–41.

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). *Foundations of machine learning*: MIT press.

Moodle. (2021). Moodle website. Retrieved May, 28th, 2021, from https://moodle.org

Moore, J. L., Dickson-Deane, C., & Galyen, K. (2011). e-Learning, online learning, and distance learning environments: Are they the same? *The Internet and Higher Education, 14*(2), 129–135.

Moravcsik, J. E., & Kintsch, W. (1993). Writing quality, reading skills, and domain knowledge as factors in text comprehension. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale, 47*(2), 360.

Mukherjee, P., Leroy, G., & Kauchak, D. (2018). Using Lexical Chains to Identify Text Difficulty: A Corpus Statistics and Classification Study. *IEEE journal of biomedical and health informatics, 23*(5), 2164-2173.

Nicula, B., Dascalu, M., Sirbu, M.-D., Trausan-Matu, S., & Nuta, A. C. (2019). Building a Comprehensive Romanian Knowledge Base for Drug Administration. In G. Angelova, R. Mitkov, I. Nikolova & I. Temnikova (Eds.), *Int. Conf. on Recent Advances in Natural Language Processing (RANLP 2019)* (pp. 829–836). Varna, Bulgaria.

Nistor, N., Dascalu, M., Serafin, Y., & Trausan-Matu, S. (2018). Automated Dialog Analysis to Predict Blogger Community Response to Newcomer Inquiries. *Computers in Human Behavior, 89*, 349–354. doi: https://doi.org/10.1016/j.chb.2018.08.034

Nistor, N., Dascalu, M., Tarnai, C., & Trausan-Matu, S. (2020). Predicting newcomer integration in online learning communities: Automated dialog assessment in blogger communities. *Computers in Human Behavior, 105*.

O'Rourke, S. T., & Calvo, R. A. (2009). Analysing semantic flow in academic writing. In V. Dimitrova, R. Mizoguchi, B. du Boulay & A. C. Graesser (Eds.), *Artificial Intelligence in Education. Building learning*

systems that care: From knowledge representation to affective modelling (AIED2009) (pp. 173–180). Amsterdam, The Netherlands: IOS Press.

O'Reilly, T., & McNamara, D. S. (2007). Reversing the reverse cohesion effect: Good texts can be better for strategic, high-knowledge readers. *Discourse Processes, 43*(2), 121–152.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The PageRank citation ranking: Bringing order to the web: Stanford InfoLab.

Paraschiv, I. C., Sirbu, M.-D., Nicula, B., Dascalu, M., Trausan-Matu, S., & Nuta, A. C. (2019). Designing an Intelligent Platform for Drugs Administration. In *International Conference on Human-Computer Interaction (RoCHI2019)* (pp. 53–59). Bucharest, Romania: MatrixRom.

Rasa Technologies Inc. (2021). Rasa website. Retrieved July, 5th, 2021, from https://rasa.com

Rice, J. (2018). What Makes Hemingway Hemingway? *The LitCharts Blog, 3*.

Roscoe, R. D., Varner, L. K., Crossley, S. A., & McNamara, D. S. (2013). Developing pedagogically-guided algorithms for intelligent writing feedback. *International Journal of Learning Technology 25, 8*(4), 362–381.

Scott, J. (1988). Social network analysis. *Sociology, 22*(1), 109-127.

Scott, J. (2017). *Social Network Analysis*. London, UK: Sage.

Shaw, M. E. (1954). Group structure and the behavior of individuals in small groups. *The Journal of psychology, 38*(1), 139-149.

Simonson, M. (2007). Course management systems. *Quarterly review of distance education, 8*(1), 7-9.

Sirbu, M.-D., Botarleanu, R., Dascalu, M., Crossley, S. A., & Trausan-Matu, S. (2018). ReadME – Enhancing Automated Writing Evaluation. In *18th Int. Conf. on Artificial Intelligence: Methodology, Systems, and Applications (AIMSA 2018)* (pp. 281–285). Varna, Bulgaria: Springer.

Sirbu, M.-D., Dascalu, M., Crossley, S. A., McNamara, D. S., Barnes, T., Lynch, C. F., & Trausan-Matu, S. (2018a). Exploring Online Course Sociograms using Cohesion Network Analysis. In C. P. Rosé, R. Martínez-Maldonado, U. Hoppe, R. Luckin, M. Mavrikis, K. Porayska-Pomsta, B. McLaren & B. d. Boulay (Eds.), *19th Int. Conf. on Artificial Intelligence in Education (AIED 2018), Part II* (pp. 337–342). London, UK: Springer.

Sirbu, M.-D., Dascalu, M., Crossley, S. A., McNamara, D. S., & Trausan-Matu, S. (2019). Longitudinal Analysis of Participation in Online Courses Powered by Cohesion Network Analysis. In *13th Int. Conf. on Computer-Supported Collaborative Learning (CSCL 2019)* (pp. 640–643). Lyon, France: ISLS.

Sirbu, M.-D., Dascalu, M., Gifu, D., Cotet, T.-M., Tosca, A., & Trausan-Matu, S. (2018b). ReadME – Improving Writing Skills in Romanian Language. In V. Pais, D. Gifu, D. Trandabat, D. Cristea & D. Tufis (Eds.), *13th Int. Conf. on Linguistic Resources and Tools for Processing Romanian Language (ConsILR 2018)* (pp. 135–145). Iasi, Romania.

Referințe

Sirbu, M.-D., Panaite, M., Secui, A., Dascalu, M., Nistor, N., & Trausan-Matu, S. (2017). ReaderBench: Building Comprehensive Sociograms of Online Communities. In *19th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC 2017)* (pp. 225–231). Timisoara, Romania: IEEE.

Somasundaran, S., Burstein, J., & Chodorow, M. (2014). Lexical chaining for measuring discourse coherence quality in test-taker essays. In *Proceedings of COLING 2014, the 25th International conference on computational linguistics: Technical papers* (pp. 950-961).

Stahl, G. (2006). *Group cognition. Computer support for building collaborative knowledge*. Cambridge, MA: MIT Press.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology, 60*(3), 538-556.

Stamati, D., Dascalu, M., & Trausan-Matu, S. (2015). Creativity stimulation in chat conversations through morphological analysis. *University Politehnica of Bucharest Scientific Bulletin Series C-Electrical Engineering and Computer Science, 77*(4), 17–30.

Stanford NLP Group. (2021). Stanford CoreNLP. Retrieved July, 5th, 2021, from https://stanfordnlp.github.io/CoreNLP/

Terian, S.-M., Cotet, T.-M., Sirbu, M.-D., Dascalu, M., & Trausan-Matu, S. (2019). Discourses of Economic Crisis in Romanian Media: An Automated Analysis using the ReaderBench Framework. *Transylvanian Review, XXVIII*(Supplement No. 1), 245–262.

The Allen Institute for Artificial Intelligence. (2021). Allen NLP. Retrieved July, 5th, 2021, from https://allennlp.org

Trausan-Matu, S. (2010). Automatic Support for the Analysis of Online Collaborative Learning Chat Conversations. In P. M. Tsang, S. K. S. Cheung, V. S. K. Lee & R. Huang (Eds.), *3rd Int. Conf. on Hybrid Learning* (pp. 383–394). Beijing, China: Springer.

Trausan-Matu, S. (2020). The Polyphonic Model of Collaborative Learning. In N. Mercer, R. Wegerif & L. Major (Eds.), *The Routledge International Handbook of Research on Dialogic Education* (pp. 454–468). Abingdon, UK: Routledge.

Trausan-Matu, S., Dascalu, M., Rebedea, T., & Gartner, A. (2010). Corpus de conversatii multi-participant si editor pentru adnotarea lui. *Revista Romana de Interactiune Om-Calculator, 3*(1), 53–64.

Trausan-Matu, S., & Rebedea, T. (2009). Polyphonic inter-animation of voices in VMT. In G. Stahl (Ed.), *Studying Virtual Math Teams* (pp. 451–473). Boston, MA: Springer.

Trausan-Matu, S., Stahl, G., & Sarmiento, J. (2007). Supporting polyphonic collaborative learning. *E-service Journal, Indiana University Press, 6*(1), 58–74.

Tu, C.-H. (2002). The measurement of social presence in an online learning environment. *International Journal on E-learning, 1*(2), 34–45.

Referințe

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *31st Conf. on Neural Information Processing Systems (NIPS 2017)* (pp. 5998–6008). Long Beach, CA, USA.

Warschauer, M., & Ware, P. (2006). Automated writing evaluation: defining the classroom research agenda. *Language Teaching Research, 10*, 1–24.

Watson, W., & Watson, S. L. (2007). An argument for clarity: What are learning management systems, what are they not, and what should they become. *TechTrends, 51*(2), 28-34.

Weidlich, J., & Bastiaens, T. J. (2019). Designing sociable online learning environments and enhancing social presence: An affordance enrichment approach. *Computers & Education, 142*, 16.

Witte, S. P., & Faigley, L. (1981). Coherence, cohesion, and writing quality. *College composition and communication, 32*(2), 189-204.

Zeno, S. M., Ivens, S. H., Millard, R. T., & Duvvuri, R. (1995). *The educator's word frequency guide.* Brewster, NY: Touchstone Applied Science Associates, Inc.