BABEŞ-BOLYAI UNIVERSITY

Soluții bazate pe Algoritmi de Invătare Automată pentru Procesarea Textului și Sinteza de Voce

REZUMATUL TEZEI DE DOCTORAT

Doctorand:
Maria LEOTESCU (NUŢU)

Coordonator s, tiint, ific:
Prof. Univ. Dr. Horia F. POP

Facultatea de Mathematică și Informatică
Departamentul de Informatică

Cluj-Napoca
2023

Contents

Lista Publicat¸iilor

Clasamentul publicațiilor a fost realizat conform standardelor CNATDCU pentru domeniul Informatică, folosind clasificările UEFISCDI[1] pentru reviste și jurnale și clasificările CORE [2] pentru conferințe și volumele acestora. Pentru fiecare articol a fost luată în considerare clasificarea validă în anul publicației.

Publicat¸ii în jurnale internationale s¸i în conferint¸e

1. [1] **Maria Nu¸tu**. *Deep Learning Approach for Automatic Romanian Lemmatization*. In *2021 25th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES2021)*, Procedia Computer Science, Elsevier Publisher, vol. 192, pp. 49-58.

   Categoria B, 4 puncte.

2. [2] Adriana Mihaela Coroiu, Alina Delia Călin, and **Maria Nuţţu**. Communication Style - An Analysis from the Perspective of Automated Learning. In *Artificial Neural Networks and Machine Learning (ICANN)*, Cham Springer International Publishing, pp. 589–597, 2018.ISBN: 978-3-030-01418-6

   Categoria B, 4 puncte.

3. [3] Adriana Mihaela Coroiu, Alina Delia Călin, and **Maria Nuţu**. Topic Modeling in Medical Data Analysis. Case Study Based on Medical Records Analysis. In *2019 International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*, pp. 1–5, 2019.

   Categoria B, 4 puncte.

4. [4] Beáta Lőrincz, **Maria Nuţu**, Adriana Stan and Mircea Giurgiu. An Evaluation of Postfiltering for Deep Learning Based Speech Synthesis with Limited Data. In:2020 *IEEE 10th International Conference on Intelligent Systems (IS)*, pp. 437–442, 202, DOI:10.1109/IS48319.2020.9199932.

   Categoria C, 1 point.

5. [5] **Maria Nuţu**, Beáta Lőrincz and Adriana Stan Deep Learning for Automatic Diacritics Restoration in Romanian. In *2019 IEEE 15th International Conference on Intelligent Computer Communication and Processing (ICCP)*,IEEE Computer Society, pp. 235–240, 2019.

   Categoria C, 2 puncte.

6. [6] Beáta Lőrincz, **Maria Nuţu**, and Adriana Stan "Romanian Part of Speech Tagging using LSTM Networks". In *2019 IEEE 15th International Confer-ence*

[1] https://uefiscdi.gov.ro/premierea-rezultatelor-cercetarii-articole-web-of-science-precisi
[2] https://portal.core.edu.au/conf-ranks

on Intelligent Computer Communication and Processing (ICCP), IEEE Computer Society, pp. 223–228, 2019.

Categoria C, 2 puncte.

7. [7] Adriana Stan, Beáta Lőrincz, **Maria Nuţu** and Mircea Giurgiu, "The MARA corpus: Expressivity in end-to-end TTS systems using synthesised speech data," 2021 International Conference on Speech Technology and Human-Computer Dialogue (SpeD), Bucharest, Romania, 2021, pp. 85-90,

   Categoria D, 0.5 puncte.

Punctajul publicat¸iilor: 17.5 **puncte**

Citări ale lucrărilor de cercetare publicate (sursa: Google Scholar)

- [1] **Maria Nu¸tu** Deep Learning Approach for Automatic Romanian Lemmatization. In *2021 25th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES2021)*, Procedia Computer Science, Elsevier Publisher, vol. 192, pp. 49-58.

  Citări

  1. [8] Pratama, Angga, Raksaka Indra Alhaqq, and Yova Ruldeviyani. "Sentiment Analysis Of The Covid-19 Booster Vaccination Program as a Requirement for Homecoming During Eid Fitr In Indonesia." Journal Of Theoretical And Applied Information Technology , vol.101, No.1, ISSN: 1817-3195 (2023).

- [7] Adriana Stan, Beáta Lőrincz, **Maria Nuţu** and Mircea Giurgiu, "The MARA corpus: Expressivity in end-to-end TTS systems using synthesised speech data," 2021 International Conference on Speech Technology and Human-Computer Dialogue (SpeD), Bucharest, Romania, 2021, pp. 85-90,

  Citări

  1. [9] Ungureanu, D., Badeanu, M., Marica, G. C., Dascalu, M., and Tufis, D. I. (2021, October). Establishing a Baseline of Romanian Speech-to-Text Models. In 2021 International Conference on Speech Technology and Human-Computer Dialogue (SpeD) (pp. 132-138). IEEE.
  2. [10] Beáta Lőrincz, Elena Irimia, Adriana Stan, and Verginica Barbu Mititelu. "RoLEX: The development of an extended Romanian lexical dataset and its evaluation at predicting concurrent lexical information." Natural Language Engineering (2022): 1-26.

- [4] Beáta Lőrincz, **Maria Nuţu**, Adriana Stan and Mircea Giurgiu. An Evaluation of Postfiltering for Deep Learning Based Speech Synthesis with Limited Data. In: 2020 *IEEE 10th International Conference on Intelligent Systems (IS)*, pp. 437–442, 202, DOI:10.1109/IS48319.2020.9199932

  Citări:

1. [11] Eren, Eray, and Cenk Demiroglu. *Deep learning-based speaker-adaptive postfiltering with limited adaptation data for embedded text-to-speech synthesis systems*. Computer Speech & Language (2023): 101520.
2. [12] Beáta Lőrincz. Contributions to neural speech synthesis using limited data enhanced with lexical features. In Proc. 2021 ISCA Symposium on Security and Privacy in Speech Communication (pp. 83-85).
3. [13] Anas Fahad Khan et. al. *When Linguistics Meets Web Technologies. Recent advances in Modelling Linguistic Linked Open Data.* In The journal Semantic Web – Interoperability, Usability, Applicability, publisher IOS Press, ISSN: 1570-0844,

- [5] **Maria Nuţu**, Beáta Lőrincz and Adriana Stan. Deep Learning for Automatic Diacritics Restoration in Romanian. In *2019 IEEE 15th International Conference on Intelligent Computer Communication and Processing (ICCP)*,IEEE Computer Society, pp. 235–240, 2019.

  Citări:

  1. [14] Stankevičius, L., Lukoševičius, M., Kapočiūtė-Dzikienė, J., Briedienė, M., & Krilavičius, T. (2022). Correcting diacritics and typos with a ByT5 transformer model. Applied Sciences, 12(5), 2636.
  2. [15] Pakalniškis, L. (2022). Giliuoju mokymusi gr˛istas diakritiniu˛ ženklu˛ atstatymas lietuviu˛ kalbai (Doctoral dissertation, Kauno technologijos universitetas).
  3. [16] Stan, A., & Lőrincz, B. (2021). Generating the Voice of the Interactive Virtual Assistant. In Virtual Assistant. IntechOpen.
  4. [17] Hifny, Y. (2021, June). Recent Advances in Arabic Syntactic Diacritics Restoration. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 7768-7772). IEEE.
  5. [18] Náplava, J., Straka, M., & Straková, J. (2021). Diacritics Restoration using BERT with Analysis on Czech language.
  6. [19] Esmail, S., Bar, K., & Dershowitz, N. (2021). How Much Does Lookahead Matter for Disambiguation? Partial Arabic Diacritization Case Study. (Master thesis, Tel Aviv University, Blavatnik School of Computer Science)
  7. [20] Scott, K. M., Ashby, S., & Cibin, R. (2020, September). Implementing text-to-speech tools for community radio in remote regions of Romania. In Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers (pp. 123-126).
  8. [21] Al-Thubaity, A., Alkhalifa, A., Almuhareb, A., & Alsanie, W. (2020). Arabic diacritization using bidirectional long short-term memory neural networks with conditional random fields. IEEE Access, 8, 154984-154996.
  9. [22] IORDACHE, F., GEORGESCU, L., ONEAŢĂ, D., & CUCU, H. (2019). Romanian Automatic Diacritics Restoration Challenge. In Proceedings of the 14th international conference "linguistic resources and tools for natural language processing (pp. 64-74).

- [6] Beáta Lőrincz, **Maria Nuţu**, and Adriana Stan "Romanian Part of Speech Tagging using LSTM Networks". In *2019 IEEE 15th International Confer-ence on Intelligent Computer Communication and Processing (ICCP)*, IEEE Computer Society, pp. 223–228, 2019.

    1. [23] Shafahat Sardarov. *Development and Design of Deep Learning-based Parts-of-Speech Tagging System for Azerbaijani language*, Thesis for Master of Science in Engineering in Computer Science, 2022, Khazar University, Azerbaijan
    2. [24] Josipa Juričić. *Označavanje vrsta riječi pomoću neuronskih mreža*. Master thesis, University of Split, Faculty of Science. Department of Informatics, 2022.

- [3] Adriana Mihaela Coroiu, Alina Delia Călin, and **Maria Nuţu**. Topic Modeling in Medical Data Analysis. Case Study Based on Medical Records Analysis. In *2019 International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*, pp. 1–5, 2019.

    Citări:

    1. [25] Gupta, Aditi, and Hoor Fatima. "Topic Modeling in Healthcare: A Survey Study." NEUROQUANTOLOGY 20.11 (2022): 6214-6221.
    2. [26] Kenei, J., Opiyo, E., & Machii, J. Modeling and visualization of clinical texts to enhance meaningful and user-friendly information re-trieval. In Med. Sci. Forum, Vol. 1, February, 2022.

Cuvinte Cheie

Învățare automată, Rețele Neuronale Adânci, Procesarea Limbajului Natural, LSTM, CNN, sequence-to-sequence, Limba Română, Restaurarea Automată a Diacriticelor, Lematizare, Identificarea părții de vorbire, MSD, CTAG, Sinteza de Voce, Corpus Expresiv de date, Sisteme End-to-end, HTS, Merlin, Date reduse, postfiltrare, Clasificare, Predicție, Stiluri de comunicare, Date Medicale, Chestionare Psihologice.

Introducere

Obiectivul principal al acestei teze este axat pe prelucrarea textului și pe sintetizarea vocii. Mergând mai în profunzime, am îmbogățit instrumentele de procesare a textului care rezolvă automat sarcini pentru textele scrise în limba română, precum restaurarea diacriticelor, lematizare și determinarea părților de vorbire. Pentru a câștiga mai multă experiență am mers dincolo de domeniul lingvistic în încercarea de a acoperi anumite lacune din domeniul medical penru a automatiza diferite sarcini. Pentru partea de sinteză a vorbirii am lucrat la metode de sporire a expresivității vocii create artificial împreună cu modalități de imbunatatirea calității acesteia. Aceste două direcții (prelucrarea textului și sinteza vorbirii) vor fi unite în viitorul apropiat, pentru a obține un instrument solid și unitar, capabil să producă a voce sintetică expresivă de înaltă calitate pornind de la texte scrise în limba română.

În era Big Data, din ce în ce mai multe informații sunt disponibile aproape peste tot sub orice formă (scrisă, desenată, audio, video etc.) și într-o varietate de stiluri de comunicare: de la formal (cursuri tehnice online, fișe de post, invitație la evenimente de afaceri sau profesionale etc.) la informal (rețele sociale, bloguri scrise etc.). Prelucrarea unor cantități atât de mari de date poate încetini activitățile zilnice, ducând la oboseală sau la depășirea termenului limită al sarcinilor.

Unul dintre domeniile care operează cu seturile mari de date menționate mai sus este domeniul medical. Medicii trebuie nu numai să ia deciziile corecte pe baza istoricului medical al pacientului, ci și să se încadreze în timpul alocat consultației unei persoane. Dincolo de analiza si corelarea diferitelor aspecte din viata pacientului, medicul ar trebui să gândeasca pe loc o schemă de tratament compatibilă cu toate afectiunile sau bolile preexistente ale pacientului consultat.

Având în vedere toate aceste aspecte, cercetătorii au investigat impactul algoritmilor de învățare automată asupra dezvoltării unor instrumente mai bune pentru analiza datelor medicale. De exemplu, algoritmii de învățare automată pot fi utilizați în imagistica medicală (și anume raze X sau scanări prin rezonanță magnetică - RMN) folosind recunoașterea modelelor pentru a căuta modelele care indică o anumită boală [27]. O altă aplicație a învățării automate în domeniul medical este de a extrage informații utile din fișa medicală a fiecărui pacient. Mai exact, folosind tehnicile de *Topic modelling*, putem găsi automat diagnosticul unei persoane prin analiza dosarelor medicale personale. Astfel nu numai că ușurăm munca de rutină a medicului, dar evităm și efectele oboselii în luarea deciziilor eronate. Aceste automatizări nu înlocuiesc rolul ființei umane (medicul), în sensul că va fi întotdeauna nevoie de intervenție umană specializată, pentru a oferi o interpretare personalizată și contextualizată. Însă integrarea proceselor automate în activitatea zilb=nică a medicilor duece la evitarea sarcinilor repetitive și monotone, ceea ce este de neprețuit.

Pornind de la toate aspectele menționate mai sus, am analizat înregistrări medicale pentru a dezvolta un instrument de detectare automată a diagnosticului unui pacient, folosind metode de *topic modelling*. Am procesat un set de date scris în limba engleză, care conțin date despre starea de sănătate a pacienților, colectate manual de

un medic de familie. Pe baza studiului tendințelor actuale realizat în **Capitolul** 3, am aplicat tehici de *topic modelling*, și anume Latent Dirichlet Allocation (LDA) și Latent Indexare semantică, pentru a grupa documentele medicale pe baza diagnosticelor descrise prin simptome similare. Rezultatele noastre au fost puclicate în lucrarea de cercetare originală [3].

Când vine vorba despre date medicale, un rol important îl joacă conținutul scris obținut din răspunsurile colectate prin intermediul aplicării unor chestionare (pen- tru a determina diferite trăsături, stiluri de comunicare, personalitate psihologică, tendinț e viitoare în cumpărături sau marketing etc.). Prin urmare, devine imperios necear să descoperim o metodologie de exatrgere automată a informațiilor/conținutului din datele colectate.

Îmbunătățirea sistemelor de sinteză a vocii pentru limba română

Începând cu primii ani de viață, specia umană învață să comunice folosind cuvinte. Prin vorbire, ne exprimăm nevoile, ideile, emoțiile sau sentimentele. Prin urmare, sinteza vorbirii sau procesul de generare a limbajului vorbit dintr-un text scris dat, și-a câștigat locul în topul interesului cercetătorilor de inteligență artificială. În zilele noastre, cu ajutorul tehnologiilor moderne și a Învățării Profunde ((Tacotron [28, 29], TransformerTTS [30]) putem obține voci sintetizate de înaltă calitate, aproapiate de vorbirea naturală a omului. Cu toate acestea, în majoritatea cazurilor, sistemele text-to-speech reușesc să transmită doar informațiile cuprinse în text, fără emoțiile, caracteristicile sau tonurile vocii vorbitorului (sarcasm, ironie etc.). Astfel, obținem un mesaj liniar, uneori diferit ca înțeles față de ideea transmisă inițial.

Poate că una dintre cele mai utile aplicații ale sintezei vorbirii este ajutorul oferit oamenilor diagnosticați cu boli severe care duc la pierderea vocii (printre care cancerul laringian și boala neuronului motor), fie prin recrearea vocii lor originale folosind înregistrări audio mai vechi, ori de câte ori acets lucru este posibil, sau prin utilizarea unei voci artificiale, sintetizată cu ajutorul unui sistem Text-to-Speech (TTS). Un exemplu actual și monden este cel al actorului american Val Kilmer, care și-a pierdut vocea după ce a fost diagnosticat cu cancer laringian. Când vine vorba de filme, comunicarea verbală este crucială, așa cum actoria presupune transmiterea unui mesaj atât verbal cât și mai ales artistic, cu tonuri, intonații și emoții diferite, ducând la semnificații ascunse si diverse. Astăzi, Val Kilmer continuă să joace în filme folosind o voce produsă artificial [3].

Un alt exemplu cunoscut este cel al omului de știință Stephen Hawking[4] care și-a pierdut-o vocea după ce a fost diagnosticat cu o formă cu debut precoce și evoluție lentă a bolii neuronului motor, care l-a paralizat treptat, ducând la incapacitatea de a vorbi. Și în acest caz, vocea originlă a omului de știință a putut fi recreată, deoarece sunt disponibile multe înregistrări audio cu vocea sa, care descriu descoperirile și cercetările sale științifice.

Totuși, în ciuda celor două exemple prezentate mai sus, pentru majoritatea pacienților mostrele audio sau înregistrările cu vocile lor originale nu sunt întotdeauna disponibile. Aceasta implică crearea unei voci artificiale cu seturile de date disponibile adaptate mai ales în acest scop [31, 32]. F Din acest punct, apar două întrebări:

1. Cum putem crea voci care să transmită emoțiile vorbitorului?

[3] videoclipuri s,i mostre ale vocii reconstruite sunt disponibile online: https://www.youtube.com/watch?v=0SMue60Gg6s

[4] Mai multe informat,ii sunt disponibile online: https://www.hawking.org.uk

2. Cum putem crea voci de calitate pornind de la seturi mici de date (pentru limbile cu resurse reduse), fiind cunoscut faptul că tehnologiile actuale de învățare profundă necesită seturi mari de date de intrare pentru antrenare?

Pornind de la aceste idei, mulți cercetători și-au concentrat munca pe rezolvarea acestor aspecte. Abordăm aceste probleme în partea a doua a acestei teze. **Capitolul** 4 prezintă ideile principale, precum și o scurtă prezentare a studiilor recente atât pentru sistemele de sinteza a vocii axate pe transmiterea emoțiilordar și pentru sinteza vorbirii pentru limbi cu resurse reduse. Contribuțiile noastre originale sunt descrise în **Capitolul** 5.

într-o primă etapă ne-am concentrat pe modalități de îmbunătățire a calității vocii sintetizate obținut, deoarece există puține seturi mari de date disponibile pentru limba română, atât de necesare proceselor de sinteză. Prin urmare, am investigat diferite tehnici de postfiltrare a vocii sintetizate în vederea corectării artefactelor care pot apărea în urma antrenării sistemuelor text-to-speech cu un set limitat de date de intrare. Rezultatele sunt prezentate în lucrarea de cercetare originală [4].

Un alt pas a fost să creăm MARA[5] [7], un set cu date expresive, pentru a fi utilizat în cercetările viitoare. Pe baza setului de date nou creat, am analizat apoi diferite modalități de creștere artificială a volumului de date expresive, precum și a impactului acestor noi date asupra sintezelor ulterioare. Rezultatele sunt prezentate în cercetarea noastră originală [7].

Teza în ansamblu Pentru a obține o voce expresivă în urma procesului de sinteză, ar trebui să modelăm și să controlăm prozodia (intonația în vorbire) într-un mod apropiat vorbirii naturale. Prozodia poate fi modelată atât de caracteristicile vocii (intonație, tonalitate etc.) cât și de diverse adnotări ale textului scris (accent, părți de vorbire etc.). Prin urmare, în lucrările noastre viitoare, intenționăm să creăm un software care va integra ambele părți ale tezei: Procesarea Limbajului Natural (NLP) în limba română și sinteza text-voce pentru obținerea unei voci expresive. Mai exact, textul introdus, procesat și adnotat folosind sistemele dezvoltate și prezentate în lucrările [1,5,6] vor fi folosite ca date de intrare pentru un sistem TTS bazat pe rețele neuronale adânci, care va genera o voce sintettică expresivă.

Pe de altă parte, când am aplicat mecanismele NLP pentru datele din domeniul medical, am luat în considerare doar textele scrise în limba engleză. Când vine vorba de limba romana, textul scris ar trebui sa respecte anumite reguli. Diacritice joacă un rol important în înțelegerea sensului unui text dat. De exemplu, forma scrisă „peste" fără semne diacritice și fără alte informații contextuale, poate însemna fie *pește, (En. fish)* fie *peste (En. over)*. Sistemele dezvoltate în cadrul cercetării noastre [1, 5, 6] ne oferă posibilitatea de a preprocesa text scris în limba română, pentru a corespunde normelor limbii, fiind dat ca intrare diferiților algoritmi de clasificare din cadrul învățării automate. Ca lucrări viitoare, intenționăm să folosim sistemele [1, 5, 6] pentru a obține/extrage mai multe informații din textele medicale scrise în limba română.

Structura tezei Teză este structurată în două părți, întrucât am abordat două domenii corelate, și anume procesarea textului și sinteza vorbirii.

I Prima parte a tezei vine să ofere o soluție de automatizare a sarcinilor de procesare a textului, după cum urmează:

[5]Setul de date este dinponibil online: https://speech.utcluj.ro/sped2021_mara/

- **Capitolul** 1 descrie fundalul teoretic pentru procesarea limbajului natural. Am prezentat domeniile de bază alături de principalele aplicații, împreună cu lucrările de cercetare corelate.
- **Capitolul** 2 introduce contribu¸tiile noastre originale în rezolvarea problemelor lingvistice folosind algoritmi de învățare profundă, cum ar fi restaurarea semnelor diacritice ale textul scris [5] și determinarea lemei [1] sau a părții de vorbire a cuvintelor [6]. Toate experimentele au fost efectuate pentru texte scrise în Limba română. Rezultatele sunt destinate a fi utilizate în corelare cu constatările din **Capitolul** 5.
- **Capitolul** 3 prezintă contribuțiile proprii în prelucrarea textului scris provenind din domeniul medical, prin aplicarea algoritmilor de învățare automată pentru două sarcini principale: identificarea diagnosticului medical prin intermediul tehnici de *topic modelling* [3] și interpretarea chestionarelor psihologice cu ajutorul algoritmilor de clasificare [2]. Toate experimentele sunt efectuate pe texte scrise în limba engleză.

II A doua parte a tezei se concentrează pe îmbunătățirea sistemelor text-to-speech în limba română, din pumct de vedere al expresivității și al calității vocii.

- **Capitolul** 4 oferă o descriere teoretică succintă a conceptelor principale din domeniul sintezei de voce bazate pe metode ale învătării aurtomate. Ne concentrăm pe siteme TTS expresive și pe sisteme TTS pentru limbi cu resurse reduse. Informațiile colectate în acest capitol facilitează înțelegerea cercetării publicate în [7] și [4].
- **Capitolul** 5 prezinta contributiile proprii în îmbunătățirea sistemelor de sinteză a vocii pentru limba română prin abordarea a două aspecte principale: îmbunătățirea calității vocii sintetizate [4] și sporirea expresivității vocii sintetizate [7]. Experimentele au fost dezvoltate în cadrul unui proiect de cercetare, finanțat printr-un grant al Ministerului Cercetării și Inovației din România, PCCDI – UEFISCDI, numarul proiectului PN-III-P1-1.2-PCCDI-2017-0818/73, în cadrul PNCDI III. Proiectul nostru este descris în detaliu online la Sintero Project.

Contribuții orignale Actuala teză de doctorat a rezultat din cercetările teoretice și experimentale efectuate în două domenii principale: Procesarea textului și Sinteza vorbirii.

I Pentru domeniul procesării limbajului natural, am oferit soluții pentru aspecte ca:

- restaurarea automată a diacriticelor pentru un text scris în limba română. Am comparat 6 arhitecturi de deep learning instruite folosind doar perechi de text cu și fără semne diacritice. [5]
- determinarea automată a lemei pentru cuvintele romanesti. Noi am analizat 24 de sisteme bazate pe rețele neuronale adânci, instruite folosind perechi de cuvinte și lemele corespunzătoare, folosind cel mult partea de vorbire ca informație morfologică suplimentară. [1].
- identificarea automată a părții vorbirii pentru cuvinte din limba română. Am analizat două tipuri de arhitecturi: 1. rețele neuroale recurente (LSTM) și 2. modele secvență-la-secvență (seq2seq) folosind codificatoare si deocdoare recurente, rezultând 10 sisteme de comparat. [6]

II Din perspectiva Sintezei vorbirii, alături de colegii din proiectul Sintero[6] am realizat următoarele:

- am creat un set de date audio-text imbogățite cu expresivitate: *MARA*[7] [7]
- antrenat și testat 6 sisteme TTS bazate pe rețele neuronale de tip deep learning cu scopul de a îmbunătăți expresivitatea vocii sintetizate, în contextul lipsei unor seturi de date expresive. Rezultatele sunt discutate în lucrarea originală de cercetare [7].
- antrenat și testat 20 de sisteme TTS bazate pe rețele neuronale de tip deep learning, cu scopul de a analiza metode de postfiltrare si impactul acestor abordări asupra calității vocii sintetizate [4].

[6] https://speech.utcluj.ro/sintero/
[7] https://speech.utcluj.ro/corpora/mara.html

Aspecte teoretice din domeniul Procesării Limbajului Natural

Procesarea Limbajului Natural înainte de apariția metodelor de tip Deep Learning Cu o cantitate mare de date neetichetate, una dintre principalele provocări în rezolvarea sarcinilor ale Procesării Limbajului Natural este de a învăta o reprezentare a datelor folosind structura internă a datelor. Acest lucru duce la *Unsupervised Feature Learning* pentru obținerea unei reprezentărea datelor într-un spațiu mai mic dimensional. Modelele bazate pe arbori de decizie, Suport Vector Machine, Random Forest, Clasificare bazată pe instanțe (k-NN), Regresia logistică sau Analiza Componentelor Principale (En. *Principal Components Analysis*) au fost aplicate cu succes pentru a rezolva sarcini NLP precum *topic modelling*, analizarea sentimentelor, clasificarea textului. În cercetarea noastră am evaluat algoritmii menționați mai sus. Experimentele au fost introduse în lucrările noastre de cercetare [2, 3] și descrise în **Capitolul** 3. Cu toate acestea, în ultimii ani, odată cu renașterea rețelelor neuronale, abordările tradiționale au fost aproape în totalitate înlocuite.

Procesarea Limbajului Natural în era Deep Learning Unul dintre obiectivele Inteligenței artificiale (AI) este evoluția generarea, comunicarea și stocarea datelor la prelucrarea datelor disponibile. Cu o creștere zilnică a volumului de date, metodelor de tip Deep Learning par a fi soluția AI pentru analiza acestor cantități mari de date. Deep Learning constă într-un set de mecanisme care pot genera soluții optime, folosindu-se un un set de date adecvat. În majoritatea cazurilor, acești algoritmi sunt ating sau chiar depășesc performanțele umane. Deși nu există o definiție standardizată, acceptată de întreaga comunitate științifică, o rețea neuronală cu două sau mai multe straturi ascunse este numită rețea neuronală adâncă (DNN). Principalele diferențe dintre tipurile de rețele neuronale constau în:

1. numărul de straturi: rețele neuronale secvențiale (Feed Forward Neural Network - FFNN) și rețele neuronale adânci (DNN)

2. comunicarea dintre straturi: distribuirea ponderilor pe orizontală (rețele neuronale convoluționale - CNN), distribuirea ponderilor pe verticală (rețele neuronale recurente - RNN), sărind peste straturi (rețele neuronale reziduale), dezactivarea anumitor noduri (Dropout), forțarea neuronilor de a se concentra asupra anumitor părți a informației (modele bazate pe atenție sau Transformeri), învățare adversativă (En. *(Generative Adversarial Networks*)

Din varietatea rețelelor neuronale, în lucrările noastre de cercetare [1, 5, 6] neam concentrat asupra perceptronului multistrat (MLP), rețelelor neuronale recurente (RNN), rețelelor neuronale convoluționale (CNN) și asupra arhitecturilor care se combină aceste trei tipuri de rețele. Perspectivele teoretice sunt prezentate în **Capitolul** 1, în timp ce rezultatele experimentelor noastre sunt descrise în **Capitolul** 2.

Mai mult, am aplicat arhitecturile convoluționale pentru sarcinile de sinteză a vorbirii din [7], în timp ce DNN-urile au fost aplicate cu succes în experimentele derulate în [4].

Tratarea unor probleme lingvistice folosind metode ale Învățării Automate

Restaurarea automată a diacriticelor pentru texte scrise în Limba Română

Restaurarea automată a diacriticelor (ADR) este procesul de restaurare a simbolurilor diacritice în textele ortografice. Limba română folosește 5 litere diacritice: *ă, â, î, ș, ț*. Deși nu toate cuvintele au scrieri alternative cu și fără semne diacritice, în unele cazuri, un semn diacritic lipsă ar putea schimba complet sensul unui cuvânt (de ex. *peste* vs. *pește*), în timp ce în alte cazuri, absența semnului diacritic pentru ultima literă a cuvântului duce la imposibilitatea de a decide dacă substantiv respectiv este sau nu articulat (*mama* vs. *mamă*). În lucrarea noastră [5], propunem o abordare de tip deep learning pentru a rezolva problema ADR în limba română, folosind numai secvențe grafice, fără alte informații lingvistice suplimentare. Arhitectura secvență-la-secvență (seq2seq) este proiectată pentru a gestiona secvențe de intrare și de ieșire cu lungimi diferite. Figura 1 ilustrează un model seq2seq pentru cuvântul „masa" ca intrare și „masă" ca ieșire. Etichetele <SS> și <SE> marchează începutul și sfârșitul secvenței. Cele mai răspândite arhitecturi din spatele codificatoarelor și decodificatoarelor sunt rețelele neuronale recurente și convoluționale.

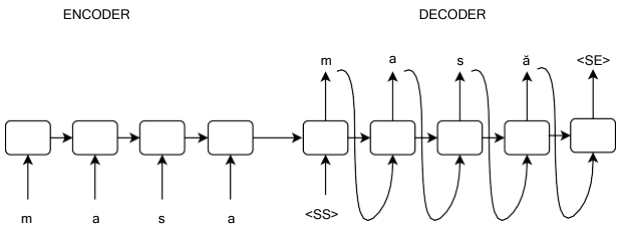

FIGURE 1: Sequence-to-sequence flow

Date de intrare Pentru instruirea și testarea modelelor noastre [5], am selectat un subset al corpusului CoRoLa [33], care conține texte din stilul beletristic. Ulterior, am împărțit setul de date în două seturi disjuncte pentru antrenare (80%) și pentru testare (20%), fiecare dintre ele fiind amestecate individual. Au fost efectuați anumiți pași de preprocesare care presupun următoarele operații: conversia textului în litere mici, eliminarea cifrelor și a semnelor de punctuație, eliminarea semnelor diacritice, împărțirea textului în trigrame, crearea de perechi de secvențe de intrare-ieșire, adăugând un caracter de început ("\t") și un caracter de final ("\n") pentru trigramul de ieșire.

Arhitectura sistemelor Pentru testele noastre inițiale am selectat două sisteme ADR [34, 35] aplicate anterior pentru limba română. Sistemele au fost reinstruite folosind setul nostru de date, dar păstrând valorile inițiale ale parametrilor. Inspirat de arhitecturile descrise folosite de aceste două sisteme, am analizat alte patru arhitecturi folosind combinații de straturi recurente și convoluționale. Pentru implementare, ne-am bazat pe Keras[8] cu TensorFlow[9] ca backend. Hiperparametrii rețelelor au fost ajustați folosind un set mic de dezvoltare. Rezultatele au fost raportate în lucrarea noastră originală de cercetare [5].

Evaluare și discuții Toate cele 6 arhitecturi introduse în [5] au fost evaluatedin punct de vedere al acurateții, definite ca raportul dintre predicțiile corecte și numărul total de mostre. Am calculat acuratețea la trei niveluri diferite: la nivel de trigram, de cuvânt și de caracter. La nivel de trigram și la nivel de cuvânt acuratețea reflect numărul de predicții corecte făcute de sistem per total. La nivel de caracter, am calculat acuratețea doar pentru caracterele care pot fi scrise cu simboluri diacritice (ă,â,î,ș,ț). Cea mai bună acuratețe (97%) a fost obținută de arhitectura bazată pe CNN și măsurată la nivel de cuvânt. Pentru acest sistem, am calculat și acuratețea celor 4 seturi ambigue de litere în limba română (a-ă-â, i-î, s-ș, t-ț). Perechea i-î a obținut o acuratețe de 99,44% a fost atins.

Determinarea automată a lemei unui cuvânt

Lematizarea este procesul de determinare a formei de dicționar a cuvântului, numită *lema*. În domeniile lingvistice, prin lematizare, toate formele flexionale ale unui cuvânt sunt grupate pentru a fi analizate ca o singură entitate. Lematizarea este dependentă de limbaj și respectă anumite reguli. De exemplu, în limba română, lema unui substantiv este forma de nominativ masculin singular, în timp ce lema unui verb este forma infinitivă a acestuia.

Experimental setup Pentru a antrena sistemele prezentate în [1], am folosit două seturi de date dfiferite. Primul set de date este Dicționarul EXplicativ al Limbii Române (ID: DEX) care conține 1.158.194 de forme ale cuvintelor, fiecare fiind asociată cu lema și cu partea de vorbire corespunzătoare. Cuvintele sunt grupate în șase mari categorii, în funcție de partea de vorbire: substantive, adjective, verbe, pronume, invariabile (adverbe, nume proprii) și forme unice (interjecții, cuvinte arhaice, nume latine). Al doilea set de date este Corpusul CoRoLa[10] [33] care conține texte din diferite stiluri funcționale: beletristic, științific, publicistic, oficial. În această lucrare s-a ales subsetul beletristic. Fiecare forma a cuvântului aparține uneia dintre 13 categorii de părți de vorbire: substantiv, verb, adjectiv, adverb, pronume, apoziție, numeral, conjuncție, cratimă, abreviere, determinant, articol și particule. Rețelele neuronale au primit la intrare perechi de forma (cuvânt-lemă), iar datele de intrare sunt codificate one-hot.

Cuvinte ambigue Pentru a rezolva problema ambiguității cuvintelor cu mai multe leme, pentru fiecare set de date, a fost construit un dicționar de leme acceptate. Mai exact, am asociat fiecare cuvânt cu un set de leme. În timpul evaluării, dacă lema

[8] https://keras.io/
[9] https://www.tensorflow.org/
[10] http://CoRoLa.racai.ro/

prezisă aparține dicționarului de leme al cuvântului respectiv, atunci este luată în considerare a fi corectă. Dicționarul lemelor este necesar chiar și atunci când este precizată și partea de vorbire a cuvântului, deoarece ambiguitatea poate exista în cadrul aceleiași categorii POS, atunci când nu sunt oferite informații morfologice suplimentare (gen, caz sau timp verbal). De exemplu:

Torturi Prăjituri (subst. pl.) → **Lema:** tort
Torturi Chinuri (subst pl.) → **Lema:** tortură

REzultate și discuții În [1] am analizat 24 de sisteme bazate pe rețele neuronale adânci pentru determinarea automată a lemei cuvintelor din limba română. Sistemele au fost instruite folosind perechi etichetate de cuvinte și lemele corespunzătoare, furnizând cel mult partea de vorbire ca informație morfologică suplimentară. Datele de intrare au fost codificate one-hot și apoi oferite ca input unor arhitecturi de tip codificator-decodor, în cadrul unui model secvență-la-secvență. Cea mai mare acuratețe (99,69%) a fost obținută de către arhitectura CNN, antrenată cu trigrame și cu un dicționar de leme pentru cuvintele ambigue. Scopul acestui studiu a fost de a folosi la intrare cât mai puține informații de lexicale, întrucât limba analizată oferă puține corpusuri cu texte adnotate complet.

Identificarea automată părții de vorbire

Alături de restaurare diacriticelor și de lematizare, o altă sarcină importantă în procesarea limbajului natural este identificarea părții de vorbire. În funcție de gradul de adnotare, în limba română avem trei tipuri de seturi de etichete RPOS, MSD și CTAG

Experimental setup Sistemele prezentate în [6] au fost antrenate folosind trei seturi de date diferite:

- Majoritatea experimentelor au fost efectuate folosind setul de date WPT [11] [36], dezvoltat folosind baza de date DexOnline [12] și Wikipedia[13]. Pentru p redicția RPOS, am antrenat sistemele folosind prima literă de la MSD set de etichete furnizat în WPT. Pentru cuvintele cu mai multe etichete, am analizat două scenarii: în primul rând, am considerat ca o ieșire corectă orice etichetă POS a cuvântului, apoi ne-am uitat doar la eticheta corspunzătoare primei apariții a cuvântului.
- Am folosit, de asemenea, setul de date DEX (bazat pe baza de date DexOnline), care asociază fiecărui cuvânt rădăcina POS corespunzătoare și o anumită frecvență. Am luat în calcul doar cuvinte cu frecvență pozitivă.
- Pentru sarcinile de predicție CTAG și MSD am antrenat sistemele cu Setul de date CoRoLa[14].

În [6] am analizat utilizarea rețelelor LSTM pentru sarcina de a prezice eticheta POS a unui cuvânt dat. Datele de intrare codificate sunt trecute printr-un strat LSTM, iar rezultatul este procesat de două straturi dense suprapuse. Al doilea strat dens

[11] http://nlptools.infoiasi.ro/WebPosTagger
[12] https://dexonline.ro/
[13] https://ro.wikipedia.org/
[14] https://corola.racai.ro/

este stratul de ieșire și are tot atâtea noduri câte etichete POS sunt posibile. Pe lângă arhitectura LSTM, un model secvență-la-secvență a fost implementat pentru sarcina de a prezice eticheta MSD. Atât codificatorul cât și decodorul sunt compuse din straturi LSTM. Am antrenat rețelele pentru un număr diferit de epoci (de la 25 la 100) cu diferite dimensiunile ale batch-ului (de la 256 la 1024) și diferite dimensiunile latente ale straturilor LSTM (de la 64 la 1024).

Rezultate și discuții În [6], performanța sistemelor a fost măsurată la nivel de acuratețe (ca raport dintre predicțiile corecte prezisă și numărul total de instațe). Cea mai bună precizie (99,18%) a fost obținută de sistemul LSTM cu un strat dens, prezicând doar prima literă a POS-ului.

Procesarea datelor din domeniul medical

Diagnosticarea pacient,ilor folosind metode de *Topic Modelling*

Când vine vorba deespre evaluarea stării de sănătate a pacienților, medicii analizează diferite aspecte ale vieții persoanei, în cadrul procesului de anamneză. Diagnosticele anterioare, antecedente familiale ale unei anumite boli, diferite simptome declarate de pacient, împreună cu mediul personal și cu stilul de viață contribuie la stabilirea unui diagnostic adecvat, astfel încât să poată fi prescris un plan de tratament adecvat.

Descrierea experimentelor și a rezultatelor Pentru experimentele derulate în [3] am folosit fișele medicale colectate de un medic de familie. Setul conține 102 de instanțe, fiecare reprezentând un pacient cu observația clinică, tratamentele curente și anterioare și răspunsul pacientului la tratament. Astfel, ca tip de date, am lucrat cu text (observarea clinică și tratamentul prescris, ambele în limba engleză) și numere (răspunsul pacientului la tratament criptat de la 1 = neresponsibil la 5 = foarte receptiv). Pentru a respecta politicile GDPR, toate informațiile personale ale pacienților (nume, adrese) au fost anonimizate de către medic înainte de a ne oferi acces la date. Au fost necesari câțiva pași de preprocesare pentru a pregăti set de date pentru algoritmii de învățare automată aleși. Pentru a genera caracteristicile textuale relevante, am folosit frecvențele cuvântului: Frecvența Termenului (TF), Frecvența Documentului invers (IDF) și Frecvența pe termen - Frecvența inversă a documentelor (TF-IDF). Pentru a modela subiectele prezente în textele analizate, am folosit alocarea de dirichlet latent (LDA) și indexarea semantică latentă. Primul pas a fost clasificarea textelor folosind TF-IDF. Modelul obținut a fost potrivit utilizând 80% din date, în timp ce predicțiile au fost realizat pentru restul de 20% din setul de date. Pentru a aplica LDA pentru modelarea subiectelor, am creat un corpus de intrare și un dicționar, folosind limbajul natural Toolkit Python toolkit: mai întâi am asociat fiecare cuvânt cu un ID unic și apoi fiecare ID-ul cuvântului a fost mapat cu frecvența cuvântului, obținând astfel corpusul dorit. Am folosit *topic coherence score* pentru a determina numărul de subiecte folosite pentru algoritmul LDA.

Analizarea stilului personal de comunicare

În era Big Data când datele audio și scrise sunt disponibile aproape peste tot (de la rețelele sociale până la bazele de date înregistrate oficiale) devine imperios necesar să abordăm problema obținerii automate a informațiilor din datele colectate fie prin interpretarea răspunsurilor la chestionare (pentru a determina diferite trăsături, stiluri de comunicare, personalitate psihologică, tendințe viitoare în cumpărături sau marketing etc.) sau prin prezicerea evenimentelor viitoare (diagnostice, evoluția

sau remisiunea bolii, niveluri de stres legate de situații contextuale, intenție suicidară, categorii vulnerabile de persoane în anumite contexte etc.).

Descrierea experimentelor și a rezultatelor În [2] am analizat șase algoritmi de învățare automată pentru sarcina de clasificare. Mai exact, am folosit datele obținute din răspunsurile oferite la un chestionar pentru determinarea stilului de comunicare și am încercat să corelăm nivelul de stres al unei persoane cu stilul de comunicare. Setul nostru de date conține 220 de instanțe cu mai mult de 60 de variabile. Am ales chestionarul propus de Marcus şi colab. în [37] care clasifică stilul de comunicare al unei persoane într-unul din cele patru stiluri de comunicare: non-asertiv, manipulator, agresiv și asertiv, prin răspunsul la un set de 60 de întrebări. În plus, pentru experimentele descrise în [2] fiecare persoană a fost rugată să-și măsoare nivelul de stres (scăzut, mediu sau ridicat). Scopul studiului nostru a fost de a analiza corelația dintre nivelul de stres și stilul de comunicare. Am transmis datele la șase algoritmi de clasificare: modele bazate pe arbori de decizie, suport vector machine, Random forest, clasificare bazată pe instanțe (k-NN), Naive Bayse și regresie logistică. Am evaluat învățarea prin aplicarea tehnicii de validare încrucișată. Prin urmare, ne-am împărțit setul de date în k subseturi și am antrenat în mod repetat sistemele folosind k - 1 subseturi pentru învățare și ultimul subset k pentru validare. Am evaluat acești algoritmi în termeni de acuratețe, precizie, sensibilitate și specificitate. Rezultatele obținute de cei șase clasificatori se află în limitele acceptate în literatura de specialitate. Dacă analizăm rezultatele obținute de către metrica de precizie, putem concluziona că Random Forest are cele mai bune performanțe, obținând o precizie de 97%, în timp ce Naive Bayes au obținut cele mai slabe rezultate, doar 85%.

Îmbunătățirea sistemelor de sinteză a vocii pentru limba română

Text-to-speech (TTS), cunoscut și sub numele de Sinteza vorbirii, reprezintă un subiect de interes pentru cercetare datorită varietății mari de aplicații din industrie. TTS urmărește să creeze vorbire inteligibilă și naturală prin sintetizarea unui text dat și necesită cunoștințe din diverse discipline: lingvistică, acustică, procesare a semnalului, învățare automată.

Pot datele de vorbire sintetizate să îmbunătățească expresivitatea vorbirii?

Conform celor mai recente cercetări, naturalețea sistemelor text-to-speech a crescut datorită utilizării modelelor de deep learning. Cu toate acestea, expresivitatea vocilor sintetizate (care depinde de existența corpusurilor expresive) rămâne un domeniu de interes, în special pentru limbajele cu resurse reduse cantitativ. Limba română face parte din categoria limbilor cu date limitate în domeniul expresivității, atât pentru voci, cât și pentru seturi de date.

Experimental Setup Pentru experimentele noastre efectuate în [7] am creat MARA, un set de date expresive disponibil online la http://speech.utcluj.ro/marasc/. În plus, ne-am propus să analizăm impactul datelor obținute în urma procesului de sinteză asupra expresivitate generale a sistemului TTS. În primul rând, am împărțit setul de date în două subseturi pe baza expresivității conținute în date: MARA-Flat și MARA-Expr. Am antrenat sistemele TTS care folosesc numai subsetul narativ MARA-Flat. Din datele audio MARA-Expr am extras durata fonemelor și conturul F0 care, combinate cu parametrii spectrali generați de sistemele TTS pentru aceleași enunțuri, au generat forma de undă. Pentru a suprima artefactele obținute în urma sintetizate datelor folosind HTS, am aplicat o procedură de postfiltrare. În consecință, ieșirea sistemului HTS antrenat cu întregul set de date MARA-Flat a fost asociat cu fișiere audio naturale într-o manieră de conversie de voce. Această arhitectură de conversie a vocii este menită să direcționeze intrarea inițială către vocea țintă, corectând astfel artefactele rezultate.

Evaluare și rezultate Sistemele au fost evaluate atât prin metode obiective, cât și subiective. Am ajuns la concluzia că apar diferențe semnificative statistic între caracteristicile sistemele antrenate cu date expresive obtinute prin sinteză. Mai mult, testele de ascultare au arătat că, deşi substituind probele naturale cu copiile lor sintetizate în datele de antrenament ale unui sistem TTS end-to-end, rețeaua este capabilă să facă o medie a artefactelor spectrale pentru acste mostre audio. Cu toate acestea, principalul scop al acestui studiu a fost introducerea setului de date *MARA*, cu date audio expresive în limba română.

Utilizarea postfiltrarii pentru a îmbunătăți calitatea sistemelor TTS cu date limitate

Pentru limbile cu puțini vorbitori nativi, este dificil să obținem seturi mari de date pentru a antrena un sistem TTS calitativ. Cele mai comune abordări pentru a suprima acest dezavantaj constau în înajustarea sau în adaptarea parametrilor modelului pre-antrenat folosind date de la un anumit vorbitor sau dintr-o anumită limbă țintă, sau prin adăugarea unor embedding-uri ale vorbitorului la caracteristicile acustice sau lingvistice pentru a ajuta modelul să învețe discriminativ caracteristici din setul de date de antrenament.

Experimental setup În lucrarea originală de cercetare [4] am evaluat utilizarea tehnicilor de postfiltrare pentru a îmbunătăți vocea sintetizată. În consecință, am antrenat un sistem TTS (folosind diverse cantități de date de la diferiți vorbitori în limba română) și am rafinat vocea obținută folosind o rețea de postfiltrare, pentru a depăși dezavantajul datelor limitate. Am obținut 20 de sisteme care au fost analizate folosind metode obiectiv de evaluare. În plus, am selectat 7 sisteme pentru o evaluare prin metoda subiectivă a testelor de ascultare, aplicate unor vorbitori nativi de limba română. Experimentele au fost structurate în două etape:

1. am antrenat un sistem TTS bazat pe deep learning, folosind cantități diferite de date
2. am folosit o rețea neuronală de postfiltrare pentru a îmbunătăți vocea sintetocă obținută

Am folosit corpusul de vorbire în limba română SWARA [38] cu două voci feminine suplimentare (MAR și BEA) înregistrate în scopuri de testare, în condiții similare de înregistrare și folosind aceleași prompturi. Datele de intrare au fost transmise sistemelor TTS se bazează pe configurația Blizzard Challenge 2017 Merlin [39]. Cantitatea de datele intrare variază de la 50 de enunțuri (aprox. 5 minute), 100 de enunțuri (aprox. 10 minute) până la 500 de enunțuri (aprox. 50 de minute) și constau în perechi de caracteristici lingvistice și acustice. Rețelele neuronale de postfiltrare s-au bazat pe tehnica conversiei de voce (En. *voice conversion*) și adaptarea vorbitorului (o voce eigen - antrenată pe date mixte de la mai mulți vorbitori - este direcționată apoi către caracteristicile acustice ale unui anumit vorbitor).

Rezultate și interpretări Performanțele sistemelor au fost evaluate atât din perspectiva măsurilor obiective cât şi subiective. Dacă analizăm rezultatele obţinute prin aplicarea sistemelor de postfiltrare, putem observa că scorurile MCD au scăzut cu 5% până la 7,5%. Dublarea artificială a datelor a crescut calitatea sistemelor. Dublare datele atât pentru antrenament, cât și pentru postfiltrare au condus la o scădere de 10% a valorilor MCD. Cu toate acestea, dublarea datelor numai pentru etapa de postfiltrare a schimbat ușor rezultatele. Dacă sunt disponibile date pentru mai mulți vorbitori, tehnica de adaptare a vorbitorilor s-a dovedit a fi o soluție, ceea ce a condus la scoruri MCD mai mici.

Concluzii și direcții viitoare de cercetare

Această teză reunește soluții bazate pe învățarea automată pentru problemele pentru procesarea textului și sinteza vorbirii.

Pentru partea de procesare a limbajului natural (NLP), ne-am concentrat pe două direcții. Pe de o parte, am aplicat mai multe modele de învățare automată pentru a automatiza procesul de extragere a informațiilor relevante din fișele medicale. Am analizat atât tehnicile de învățare supravegheate (clasificarea textului [2]), cât și cele nesupravegheate (gruparea documentelor, topic modelling [3]). Experimentele au fost efectuate pentru texte scrise în limba engleză. Pe de altă parte, am abordat aspecte ale adnotării textului prin aplicarea soluțiilor bazate pe rețele neuronale în sarcini precum restaurarea automată a diacriticelor [5], lematizarea automată [1] sau etichetarea POS [6]. Aceste ultime experimente au fost rulate pentru texte scrise în limba română.

În experimentele noastre de la [1,5,6] am antrenat sistemele de învățare profundă într-o manieră supervizată, cu perechi de cuvinte etichetate și adnotările corespunzătoare. (lemă, formă diacritizată sau etichetă POS, în funcție de sarcina cercetată). Textul de intrare a fost codificat și folosit de o arhitectură codificator-decodor. Ca direcție viitoare de cercetare, ne vom concentra pe analizarea și altor tipuri de rețele neuronale, cum ar fi rețelel recurente bidirecționale, GRU sau arhitecturi bazate doar pe atenție (transformers), care sunt deja aplicate frecvent în alte sarcini de procesare a textului. Cu toate acestea, îmbogățind textul de intrare cu mai multe informații de context poate duce la rezultate mai bune în prezicerea adnotării dorite. Cu toate acestea, trebuie să menționăm că studiile [1,5,6] au avut ca scop procesarea automată a textului de intrare folosind cunoștințele minime legate de context, mai ales în lipsa unor corpusuri mari adnotate în limba română.

În paralel cu adnotarea automată a textelor scrise în limba română, am aplicat algoritmii NLP de învățare automată pentru a ușura munca medicilor. Ne-am concentrat cercetarea pe două direcții principale: determinarea diagnosticului medical al pacienților prin tehnici de topic modelling [3] și interpretarea rezultatelor chestionarelor psihologice [2]. La început, am creat un set de date format din înregistrările personale ale unui medic de familie, adunate în cadrul consultațiilor. Setul de date conține 102 instanțe, constând în text scris, mai precis observațiile clinice și tratamentul prescris, limba engleză, cât și date numerice reprezentând răspunsul pacientului la tratament. O altă direcție de cercetare a combinat psihologiea cognitivă și învățarea automată: interpretarea utomată a răspunsurilor obținute în urma aplicării unui chestionar psihologic. Am analizat peste 200 de instanțe cu mai mult de 60 de variabile, deoarece fiecare participant la studiu a fost rugat să răspundă la un 60 întrebări ale unui chestionar cu scopul de a determina stilul de comunicare. Pentru setul de date obținut și sarcina curentă, am instruit și testat 6 clasificatori de învățare automată. Ca direcții viitoare de cercetare din domeniul NLP folosind date medicale, intenționăm să analizăm impactul și eficacitatea altor algoritmi de

învățare automată, dar și algoritmi de învățare profundă. În al doilea rând, ar fi interesant să studiem date medicale scrise în limba română, având în vedere nu doar particularitățile limbii (diacritice, ortografie etc.), ci și provocarea de a culege datele de intrare, având în vedere ca limba română este puțin reprezentată. Mai mult, putem aplica sistemele dezvoltat în [1, 5, 6] pentru a automatiza prelucrarea textelor medicale scrise în limba română.

Pentru partea de sinteza text în vorbire, scopul a fost sporirea calității și expresivității vocii sintetizate. Am analizat diferite arhitecturi de rețele neuronale folosind texte în limba română ca date de intrare. Abordările noastre sunt noi în raport cu domeniul Sintezei vorbirii în limba română și au fost publicate în articole de cercetare în cadrul lucrărilor conferințelor [4, 7].

În experimentele noastre din [4] am cercetat potențialul tehnicilor de postfiltrare pentru a îmbunătăți calitatea sistemelor TTS cu cu resurse reduse. Ne-am împărțit abordarea în două părți: mai întâi am antrenat sistemele TTS cu cantități diferite de date, apoi am aplicat o rețea neuronală de postfiltrare pentru a îmbunătăți vocea sintetizată. Cantitatea de date de intrare la antrenare variază de la 50 de enunțuri (aprox. 5 minute), 100 de enunțuri (aprox. 10 minute) până la 500 de enunțuri (aprox. 50 de minute) și constau în perechi de caracteristici lingvistice și acustice. În plus, am antrenat două sisteme cu date de intrare dublate, prin simpla adăugare de două ori a datelor inițiale, pentru a analiza dacă calitatea rezultatului este influențată mai degrabă de cantitatea fizică de date decât de conținutul datelor. Rețelele neuronale de postfiltrare s-au bazat pe tehnica de conversie a vocii și adaptarea către un vorbitor (o voce eigen - antrenată pe date mixte de la mai multe vorbitori - este orientată către caracteristicile acustice ale unui anumit vorbitor). Aceste ultime experimente au fost efectuate pentru texte și mostre audio în limba română și rezultatele sunt detaliate în lucrarea de cercetare [4]. Întrucât această lucrare a făcut parte din proiectul de cercetare SINTERO [15] trebuie să menționăm că eu, în particular, am fost responsabilă pentru înregistrarea propriei voci pentru testare (voce MAR) și pentru procesarea fișierelor audio rezultate. Mai mult, m-am ocupat de sistemele care au fost instruite pe baza vocii MAR și am analizat rezultatele lor obținute în cadrul experimentelor din [4].

Lucrarea de cercetare [7] analizează diferite moduri de a îmbogăți expresivitatea sistemeor TTS în contextul unor seturi de date emoțional/expresiv cu resurse reduse. Cu acest scop în minte, am creat mai întâi un set de date expresiv în limba română, pornind de la o carte audio (Mara - scrisă de Ioan Slavici) care a fost segmentată manual în fișiere mai mici, după pauzele de de vorbire ale cititorului. Textul scris a fost adnotat prin Platforma RACAI Relate cu informații lingvistice de nivel înalt. Pornind de la setul de date MARA, am analizat impactul datelor de vorbire sintetizate asupra expresivității generale a sistemelor TTS. Am antrenat 5 sisteme TTS diferite cu diverse cantități de date expresive ca intrare. Toate rezultatele sunt descrise în munca noastră de cercetare [7].

Ca direcții viitoare de cercetare pentru partea de sinteză a vorbirii, în plus față de analizarea mai multor tipuri de arhitecturi TTS (bazate pe atenție, transformatoare), este interesant de examinat alte vocodere sau de a adăuga mai multe caracteristici (lexicale sau embedinguri ale vorbitorilor) la rețeaua de postfiltrare în încercarea de a îmbogăți calitatea vorbirii sintetizate. Pentru experimentele cu sisteme expresive TTS, ne propunem să analizăm și alte metode pentru a obține date de sinteză expresive de calitate. În plus, luăm în considerare posibilitatea transferului de prozodie de la feminim către masculin și viceversa.

[15] https://speech.utcluj.ro/sintero/

Teza în ansamblu Având în minte toate experimentele rulate pentru datele de intrare scrise în limba română, intenționăm să încapsulăm toate resursele și sistemele descrise în lucrarea de față într-un singur instrument, cu scopul de a ajuta alți cercetători. Mai precis, ne propunem să creăm un sistem text-to-speech românesc, îmbogățit cu expresivitate, care să fie alimentat cu textul scris de intrare preprocesat folosind instrumentele analizate în [1, 5, 6] și constând în tehnologiile TTS introduse în [4, 7].

Bibliografie selectivă

[1] Maria Nu¸tu. Deep learning approach for automatic romanian lemmatization. *Procedia Computer Science*, 192:49–58, 2021. Knowledge-Based and Intelligent Information and Engineering Systems: Proceedings of the 25th International Conference KES2021.

[2] Adriana Mihaela Coroiu, Alina Delia Călin, and Maria Nuțu. Communication style - an analysis from the perspective of automated learning. In Věra Kůrková, Yannis Manolopoulos, Barbara Hammer, Lazaros Iliadis, and Ilias Maglogiannis, editors, *Artificial Neural Networks and Machine Learning – ICANN 2018*, pages 589–597, Cham, 2018. Springer International Publishing.

[3] Adriana Mihaela Coroiu, Alina Delia Călin, and Maria Nuţu. Topic modeling in medical data analysis. case study based on medical records analysis. In *2019 International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*, pages 1–5, 2019.

[4] Beáta Lőrincz, Maria Nuţu, Adriana Stan, and Mircea Giurgiu. An evaluation of postfiltering for deep learning based speech synthesis with limited data. In *2020 IEEE 10th International Conference on Intelligent Systems (IS)*, pages 437–442, 2020.

[5] M. Nuţu, B. Lőrincz, and A. Stan. Deep learning for automatic diacritics restoration in romanian. In *2019 IEEE 15th International Conference on Intelligent Computer Communication and Processing (ICCP)*, pages 235–240, 2019.

[6] Beáta Lőrincz, Maria Nuţu, and Adriana Stan. Romanian part of speech tagging using lstm networks. In *2019 IEEE 15th International Conference on Intelligent Computer Communication and Processing (ICCP)*, pages 223–228. IEEE, 2019.

[7] Adriana Stan, Beáta Lőrincz, Maria Nuțu, and Mircea Giurgiu. The mara corpus: Expressivity in end-to-end tts systems using synthesised speech data. In *2021 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 85–90, 2021.

[8] ANGGA PRATAMA, RAKSAKA INDRA ALHAQQ, and YOVA RULDEVIYANI. Sentiment analysis of the covid-19 booster vaccination program as a requirement for homecoming during eid fitr in indonesia. *Journal of Theoretical and Applied Information Technology*, 101(1), 2023.

[9] Dan Ungureanu, Madalina Badeanu, Gabriela-Catalina Marica, Mihai Dascalu, and Dan Ioan Tufis. Establishing a baseline of romanian speech-to-text models. In *2021 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 132–138. IEEE, 2021.

[10] Beáta Lőrincz, Elena Irimia, Adriana Stan, and Verginica Barbu Mititelu. Rolex: The development of an extended romanian lexical dataset and its evaluation at

predicting concurrent lexical information. *Natural Language Engineering*, pages1–26, 2022.

[11] Eray Eren and Cenk Demiroglu. Deep learning-based speaker-adaptive post-filtering with limited adaptation data for embedded text-to-speech synthesis systems. *Computer Speech and Language*, page 101520, 2023.

[12] Beáta Lorincz. Contributions to neural speech synthesis using limited data enhanced with lexical features. In *Proc. 2021 ISCA Symposium on Security and Privacy in Speech Communication*, pages 83–85, 2021.

[13] Anas Fahad Khan, Christian Chiarcos, Thierry Declerck, Daniela Gifu, Elena González-Blanco García, Jorge Gracia, Maxim Ionov, Penny Labropoulou, Francesco Mambrini, John P McCrae, et al. When linguistics meets web technologies. recent advances in modelling linguistic linked open data. *Semantic Web*, 14, 2022.

[14] Lukas Stankevičius, Mantas Lukoševičius, Jurgita Kapočiūtė-Dzikienė, Monika Briedienė, and Tomas Krilavičius. Correcting diacritics and typos with a byt5 transformer model. *Applied Sciences*, 12(5):2636, 2022.

[15] Lukas Pakalniškis. *Giliuoju mokymusi grįstas diakritiniu̧ ženklu̧ atstatymas lietuviu̧ kalbai*. PhD thesis, Kauno technologijos universitetas, 2022.

[16] Adriana Stan and Beáta Lőrincz. Generating the voice of the interactive virtual assistant. In *Virtual Assistant*. IntechOpen, 2021.

[17] Yasser Hifny. Recent advances in arabic syntactic diacritics restoration. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7768–7772. IEEE, 2021.

[18] Jakub Náplava, Milan Straka, and Jana Straková. Diacritics restoration using bert with analysis on czech language. *arXiv preprint arXiv:2105.11408*, 2021.

[19] Saeed Esmail, Kfir Bar, and Nachum Dershowitz. How much does lookahead matter for disambiguation? partial arabic diacritization case study. *Computational Linguistics (2022) 48 (4): 1103–1123.*, 2022.

[20] Kristen M Scott, Simone Ashby, and Roberto Cibin. Implementing text-to-speech tools for community radio in remote regions of romania. In *Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers*, pages 123–126, 2020.

[21] Abdulmohsen Al-Thubaity, Atheer Alkhalifa, Abdulrahman Almuhareb, and Waleed Alsanie. Arabic diacritization using bidirectional long short-term memory neural networks with conditional random fields. *IEEE Access*, 8:154984–154996, 2020.

[22] FLORIN IORDACHE, LUCIAN GEORGESCU, DAN ONEAȚĂ, and HORIA CUCU. Romanian automatic diacritics restoration challenge. In *Proceedings of the 14th international conference "linguistic resources and tools for natural language processing*, pages 64–74, 2019.

[23] Shafahat Sardarov. *Development and Design of Deep Learning-based Parts-of-Speech Tagging System for Azerbaijani language*. PhD thesis, Khazar University,Azerbaijan, 2022.

[24] Josipa Juričić. *Označavanje vrsta riječi pomoću neuronskih mreža*. PhD thesis, University of Split. Faculty of Science. Department of Informatics, 2022.

[25] Aditi Gupta and Hoor Fatima. Topic modeling in healthcare: A survey study. *NEUROQUANTOLOGY*, 20(11):6214–6221, 2022.

[26] Jonah Kenei, E Opiyo, and J Machii. Modeling and visualization of clinical texts to enhance meaningful and user-friendly information re-trieval. In *Med. Sci. Forum*, volume 1. The 2nd International Electronic Conference on Healthcare, 2022.

[27] Alexander Selvikvåg Lundervold and Arvid Lundervold. An overview of deep learning in medical imaging focusing on mri. *Zeitschrift für Medizinische Physik*, 29(2):102–127, 2019. Special Issue: Deep Learning in Medical Physics.

[28] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous. Tacotron: Towards End-to-End Speech Synthesis. In *Proc. Interspeech*, 2017.

[29] Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ Skerry-Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Fei Ren, Ye Jia, and Rif A Saurous. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. *arXiv preprint arXiv:1803.09017*, 2018.

[30] Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, and Ming Liu. Neural speech synthesis with transformer network. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press, 2019.

[31] Keith Ito and Linda Johnson. The lj speech dataset. https://keithito.com/ LJ-Speech-Dataset/, 2017.

[32] Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. Libritts: A corpus derived from librispeech for text-to-speech. *arXiv preprint arXiv:1904.02882*, 2019.

[33] Verginica Barbu Mititelu, Elena Irimia, and Dan Tufis. Corola—the reference corpus of contemporary romanian language. In *LREC*, pages 1235–1239, 2014.

[34] Jakub Náplava, Milan Straka, Pavel Straňák, and Jan Hajic. Diacritics restoration using neural networks. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, 2018.

[35] Horia Cristescu. Romanian diacritic restoration with neural nets.

[36] Radu Simionescu. Graphical grammar studio as a constraint grammar solution for part of speech tagging. In *The Conference on Linguistic Resources and Instruments for Romanian Language Processing*, volume 152, 2011.

[37] Stroe Marcus, Teodora David, and Adriana Predescu. *Empatia s¸i rela¸tia profesor-elev*. Editura Academiei Republicii Socialiste Romania, 1987.

[38] Adriana Stan, Florina Dinescu, Cristina Tiple, Serban Meza, Bogdan Orza, Magdalena Chirila, and Mircea Giurgiu. The SWARA Speech Corpus: A Large Parallel Romanian Read Speech Dataset. In *Proceedings of the 9th Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, Bucharest, Romania, July, 6-9 2017.

[39] Zhizheng Wu, Oliver Watts, and Simon King. Merlin: An open source neural network speech synthesis system. In *proceedings of the 9th International Speech Communication Association (ISCA) Speech Synthesis Workshop: SSW 2016*, pages 202–207, Sunnyvale, United States, 2016. International Speech Communication Association (ISCA).