BABEŞ-BOLYAI UNIVERSITY

Machine Learning based Solutions for Text Processing and Speech Synthesis

PHD THESIS SUMMARY

Ph.D. student:
Maria LEOTESCU (NUŢU)

Scientific Supervisor:
Prof. Univ. Dr. Horia F. POP

Faculty of Mathematics and Computer Science
Department of Computer Science

Cluj-Napoca
2023

Contents

Keywords

Machine learning, deep neural networks, Natural language processing, LSTM, CNN, sequence-to-sequence, Romanian, Automatic diacritics restoration, Lemmatization, Part-of-Speech tagging, MSD, CTAG, Speech synthesis, Expressive speech corpus, End-to-end systems, HTS, Merlin, Statistical-parametric synthesis, Limited data, post-filtering, Multi-classification model, Prediction, Communication style, Medical data, Psychological questionnaires,

List of Publications

All rankings are listed according to the UEFISCDI journal classification for financing of research results[1] and CORE classification of conferences in Computer Science[2]. For each article, we considered the classification valid in the year of publication

Publications in international journals and conferences

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.

   Rank B, 4 points.

2. [2] Adriana Mihaela Coroiu, Alina Delia Călin, and **Maria Nut¸tu**. 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

   Rank B, 4 points.

3. [3] Adriana Mihaela Coroiu, Alina Delia Călin, and **Maria Nu¸tu**. 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.

   Rank B, 4 points.

4. [4] Beáta Lőrincz, **Maria Nu¸tu**, 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.

   Rank C, 1 point.

5. [5] **Maria Nu¸tu**, 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.

   Rank C, 2 points.

6. [6] Beáta Lőrincz, **Maria Nu¸tu**, 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] https://uefiscdi.gov.ro/premierea-rezultatelor-cercetarii-articole-web-of-science-precisi

[2] Computing Research and Education Association of Australasia, https://portal.core.edu.au/conf-ranks

Rank C, 2 points.

7. [7] Adriana Stan, Beáta Lőrincz, **Maria Nu¸tu** 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,

   Rank D, 0.5 points.

Publications Score: 17.5 **points**

Citations of the published research paper (source: 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.

  Citations

  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 Nu¸tu** 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,

  Citations

  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 Nu¸tu**, 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

  Citations:

  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 Nu¸tu**, 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.

  Citations:

  1. [14] Stankevicˇius, L., Lukoševicˇius, M., Kapočiū tė-Dzikienė, J., Briedienė, M., & Krilavicˇius, 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., ONEAT, Ă, 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 Nu¸tu**, 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 Juricˇic´. *Oznacˇavanje vrsta rijecˇi pomoc´u neuronskih mreža*. Master thesis, University of Split, Faculty of Science. Department of Informatics, 2022.

- [3] Adriana Mihaela Coroiu, Alina Delia Călin, and **Maria Nu¸tu**. 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.

  Citations:

   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.

Keywords

Machine learning, deep neural networks, Natural language processing, LSTM, CNN, sequence-to-sequence, Romanian, Automatic diacritics restoration, Lemmatization, Part-of-Speech tagging, MSD, CTAG, Speech synthesis, Expressive speech corpus, End-to-end systems, HTS, Merlin, Statistical-parametric synthesis, Limited data, post-filtering, Multi-classification model, Prediction, Communication style, Medical data, Psychological questionnaires,

Introduction

The main objective of this thesis is focused on processing the text and on synthesising the speech. Going deeper, we enriched the text processing tools which automatically solves tasks for texts written in Romanian language, such as diacritics restoration, lemmatization and part-of-speech tagging. To gain more experience we went beyond the linguistic field in an attempt to cover certain gaps within the medical field, that could be automated. For the speech synthesis part we worked on methods to enrich the expressivity of the artificially created voice together with ways of improving its quality.

These two direction (text processing and speech synthesis) would eventually sum up in the near future, in order to obtain a solid tool that is able to produce a high quality expressive synthetic voice from Romanian texts.

Medical text data processing

In the era of Big Data, more and more information is available almost everywhere in any form (written, drawn, audio, video, etc.) and in a variety of communication styles: from formal (technical online courses, job descriptions, invitation to business or professional events, etc.) to informal (social networks, written blogs, etc.). Processing such large amounts of data can slow the daily activities, leading to fatigue or to exceed the deadline of daily tasks.

One of the domains which operates with the above mentioned large datasets is the medical domain. Medical physicians must not only make the right decisions based on the patient's history, but also fit in the time allocated to a person's consultation. Beyond analysing and correlating different aspects from the patient's life, the medical doctor should think on the spot a treatment scheme compatible with all the pre-existing ailments or diseases of the consulted patient. Having all these aspects in mind, the researchers investigated the impact of machine learning algorithms on developing better tools to analyze medical data. For example, machine learning algorithms can be used in medical imaging (namely X-rays or Magnetic resonance imaging -MRI- scans) using pattern recognition to search the patterns that indicate a particular disease [27]. Another application of machine learning in the medical domain is to gain insight in the written information of every patient. More precisely, using the topic modelling techniques, we can automatically find one person's diagnostic by analysing the personal medical records. Thus we not only ease the medical doctor's routine work, but we also avoid the effects of fatigue in making erroneous decisions. These automations do not suppress the human role, in the sense that there will always be the need of specialised human intervention, in order to offer a customised and contextualised interpretation, but at the same time, avoiding doing the repetitive and monotone work is priceless.

Starting from all the above mentioned aspects, we analysed medical records from a medical physician in order to find a topic modelling machine learning tool to automatically find one person's diagnostic. We processed an English written dataset, containing notes on patients' health conditions, manually gathered by a medical family doctor. Based on the state-of-the-art analysed in **Chapter** 3, we applied topic modelling techniques, namely the Latent Dirichlet Allocation (LDA) and the Latent Semantic Indexing, to cluster the medical documents based on the diagnostics described through similar symptoms. Our original results are described and discussed in **Chapter** 3 and published in the research paper [3].

When it comes to medical data, an important role is played by the written content obtained from questionnaires' responses (to determine different traits, communication styles, psychological personality, future trends in shopping or marketing, etc.). Therefore it becomes imperative necessarily to discover a methodology of automatically gaining insight from the collected data.

Enhancing the Romanian Text-to-Speech systems

Starting with the first years of life, the human specie learns how to communicate using words. Through speech, we express our needs, ideas, emotions or feelings. Thus, the speech synthesis or the process of generating spoken language from a written given text, earned its place on the top of artificial intelligence's researchers' interest. Nowadays, with the help of modern technologies and deep learning (Tacotron [28, 29], TransformerTTS [30]), we can obtain high quality artificial speech, close to the natural human speech. However, in most of the cases, the text-to-speech systems manage to transmit only the information comprised by the text, with no content about the speaker's emotions, characteristics or tones (sarcasm, irony, etc). This lead to a linear message, sometimes different in meaning to the original intended idea.

Maybe one of the most useful applications of speech synthesis is helping people diagnosed with severe illnesses that lead to voice loss (among which throat cancer and motor neuron disease), either by recreating their original voice using their older audio recordings, whenever it is possible, or by using an artificial voice output by a Text-to-Speech (TTS) system. A current and mundane example is the case of the American movie actor Val Kilmer who lost his voice after being diagnosed with throat cancer. When it comes to movies, the verbal communication is crucial, as acting involves transmitting a message both verbally and especially artistically, with different tones, intonations and emotions, leading to hidden meanings. Today Val Kilmer continues to play in movies by using an artificially produced voice[3].

Another famous example is that of the scientist Stephen Hawking[4], who lost his voice after falling ill from an early-onset slow-progressing form of the motor neuron disease, which slowly paralysed him, leading to the incapacity of speaking. In this case too the original scientist's voice could be recreated, as there are available many audio recordings with his voice, describing his scientific findings and research.

However, despite these two examples presented above, for the majority of patients, audio samples or recordings with their original voices are not always available. This implies creating an artificial voice with the available datasets tailored especially for this purposes [31, 32]. From this point, two questions arise:

1. How can we create voices that convey the speaker's emotions?

[3] Videos and samples of his reconstructed voice are available online: https://www.youtube.com/watch?v=OSMue60Gg6s

[4] More information is available here: https://www.hawking.org.uk

2. How can we create quality voices based on small data sets (for low-resourced languages), being known that current deep learning technologies require large input datasets for training?

Having those ideas in mind, many researchers focused their work on overcoming these aspects. We address these issues in the second part of this thesis. **Chapter** 4 presents the main ideas as well as a brief state of the art for both emotional TTS and speech synthesis for low-resourced languagess. Our original contributions are described in **Chapter** 5.

As a first step, we focused on ways to improve the quality of the obtained synthesized voice, since there are few large datasets available for the Romanian language [33], so necessary for synthesis processes. Therefore, we investigated different techniques of post-filtering the obtained synthesized voice in order to correct the artifacts that can appear after training the text-to-speech system with a limited set of input data. The results are presented in our original research paper [4].

Another step was to create MARA[5] [7], a data set with expressive data to be used in future research. Based on the newly created data set, we then analyzed different ways of artificially increasing the volume of expressive data, as well as the impact of this new data on subsequent syntheses. The results are presented in our original research [7].

Thesis as a whole

In order to obtain a more expressive voice within the text to speech synthesis process, we should model and control the prosody (intonation in speech) in a way close to natural speech. Prosody can be shaped both by the characteristics of the voice (intonation, stress, tonality, etc.) and by various annotations of the written text (accent, parts of speech, etc.). Therefore, as future work, we intend to create a software product which will integrate the both parts of the current thesis: Romanian Natural Language Processing (NLP) and Expressive TTS. More precisely, the input text, processed and annotated using the systems developed in [1,5,6] will be passed through a deep TTS system leading to a more expressive synthesised output.

On the other hand, when we applied the NLP mechanisms for data from the medical domain, we took in consideration only the texts written in English. When it comes to Romanian language, written text should obey certain rules. Diacritics play an important role in understanding the meaning of a given text. For instance, the written form *„peste"* without diacritics and no other contextual information, can mean both *pes,te (En. fish)* or *peste (En. over)*.The systems developed within our research [1,5,6] offer us the possibility to preprocess text written in Romanian, making it appropriate to be given as input to different machine learning classification learners. As future work, we intend to use the systems [1, 5, 6] to gain more insight from the medical Romanian texts.

Thesis structure

The present thesis is structured in two parts as we addressed two correlated domains, namely text processing and speech synthesis.

[5] The dataset is available online: https://speech.utcluj.ro/sped2021_mara/

I The first part of the thesis comes to offer a solution to automatize the text processing tasks, as follows:

- **Chapter 1** describes the theoretical background for the Natural Language Processing field. We presented the core areas and the main applications, together with correlated research papers.
- **Chapter 2** introduces our original contributions in solving linguistic problems using deep learning algorithms, such as restoring the diacritics for a written text [5] and finding the lemma [1] or the part of speech of certain given words [6]. All the experiments were conducted on texts written in Romanian Language. The results are intended to be used in correlation with the findings from **Chapter** 5.
- **Chapter 3** presents our personal contributions in processing the written text from the medical domain by applying the machine learning algorithms for two main tasks: identifying the medical diagnostic through the topic modelling techniques [3] and interpreting the psychological questionnaires results with the aid of the classification learners [2]. All the experiments are based on the texts written in English Language.

II The second part of the thesis focuses on improving the Romanian Text-to-Speech systems in terms of expressivity and speech quality.

- **Chapter 4** offers a brief theoretical background of the main speech processing aspects addressed within the current research and the state-of-the-art in the field of Text-to-Speech systems, using the Machine Learning methods. We focus on Expressive TTS and on Speech Synthesis for low resourced languages. The information collected in this chapter facilitates understanding the research published in [4] and [7].
- **Chapter 5** presents our personal contributions in improving the Romanian TTS systems by addressing two main aspects: improving the quality of the synthesised voice [4] and enhancing the expressivity of the TTS system's results [7]. The experiments were developed within a research project, supported by a grant of the Romanian Ministry of Research and Innovation, PCCDI – UEFISCDI, project number PN-III-P1-1.2-PCCDI-2017-0818/73, within PNCDI III. Our project is described in detail online at Sintero Project.

Original Contributions

The current PhD thesis derives from the theoretical and experimental research done in two main domains: Text Processing and Speech Synthesis.

I For the Natural Language Processing field, we have offered solutions to:

- automatically restore the diacritics for a text written in Romanian. We have compared 6 deep learning architectures trained using only parallel input-output pairs of texts, with and without diacritics. [5]
- automatically determine the lemma for Romanian words. We have analysed 24 systems based on Deep Neural Networks, trained on labelled pairs of words and the corresponding lemmas, using at most the part-of-speech tag as morphological information. [1]

- automatic Romanian Part of Speech tagging. We have analysed two types of architectures: 1. simple long short-term memory networks (LSTM) - based networks and 2. sequence-to-sequence architecture (seq2seq) based on LSTM layers - with different types of encodings for the input data (one hot encoding or letter encoding), resulting in 10 systems to be compared. [6]

II From the perspective of Speech Synthesis, myself along with the Sintero[6] colleagues have:

- created a large speech dataset containing more dynamic intonation patterns, the MARA Dataset[7] [7]
- trained and tested 6 deep learning TTS systems to improve the expressivity of the synthesised voice, in the context of lacking expressive datasets. The results are discussed in the original research paper [7].
- trained and tested 20 deep learning TTS systems in 3 postfiltering scenarios in order to evaluate the impact of each approach on the quality of the synthesised voice [4].

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

Theoretical insights into Natural Language Processing problems

Processing the Natural Language tasks before Deep Learning With a large amount of unlabelled data, one of the main challenges in solving NLP tasks is to learn a data representation from the inner data structure itself. This leads to Unsupervised Feature Learning, an approach to obtain a lower dimensional representation of the data from the higher-dimensional initial space. Techniques as Decision Tree Based Model, Support Vector Machine, Random Forest, Classification based on instances (k-NN), Logistic Regression or Principal Components Analysis have been successfully applied to solve NLP tasks as Topic Modelling, Sentiment Analysis, Text Classification.

In our research we also evaluated the above mentioned algorithms. The experiments were introduced in our research papers [2] and [3], described in **Chapter** 3.

However, during the last years, once with the revival of the neural networks, the traditional approaches have been almost totally replaced.

Deep Learning for Natural Language Processing An artificial intelligence (AI) goal might be to upgrade from generating, communicating and storing the data to processing the available data. With a daily increasing of the data volume, deep learning seems to be the solution of AI for analysing these large amounts of data. Deep learning consists in a set of mechanism which can generate optimal solutions given an appropriate input dataset. In most of the cases, these algorithms equal or even outperform the human capabilities. Although there is not a standardized definition accepted by all the researchers, a neural network with two or more hidden layers is called *deep neural network* (DNN). The main differences between the different types of neural networks consists in:

- the number of layers: sequential neural networks (Feed Forward Neural Network - FFNN) and deep neural networks (DNN)
- the way the nodes communicate between layers: horizontally sharing the weights (Convolutional Neural Networks - CNN), vertically sharing weights (Recurrent Neural Networks - RNN), skipping layers (Residuals Neural Networks), simply deactivating certain nodes (Dropout), forcing neurons to focus on certain pieces of input information (Attention Mechanism and Transformers) and adversarial learning (Generative Adversarial Networks - GAN).

Among this diversity of the neural networks, in our research studies [1, 5, 6] we focused on the Multilayer perceptron (MLP), the Recurrent Neural Networks (RNN), the Convolutional Neural Networks (CNN) and architectures which combine these three types of networks. The theoretical insights are depicted in **Chapter** 1, while the experimental results are described in **Chapter** 2. Moreover, we applied the CNN architectures for the Speech Synthesis tasks from [7], while the DNNs were successfully applied in the experiments run in [4]

Addressing linguistic problems using Machine Learning models

0.1 Automatic Diacritics Restoration applied for Romanian language

Automatic Diacritics Restoration (ADR) is the process of restoring the diacritic symbols in the orthographic texts. The Romanian language uses 5 diacritic letters: *a˘, â, î, s¸* and *¸t*. Although not all the words have alternative spellings with and without diacritics, in some cases, a missing diacritic could completely change a word's meaning (e.g *peste* = over vs. *pes¸te* = fish), while in other cases, the absence of the appropriate diacritic in the word's ending letter makes it impossible to discern between the definite or indefinite form of a noun (*mama˘* = a mother vs. *mama* = the mother).

In our work [5], we propose a deep learning approach to solve the ADR problem for Romanian using only grapheme sequences, without any expert linguistic knowledge. The sequence-to-sequence (seq2seq) [34] architecture is designed to handle input and output sequences with different lengths. Figure 1 presents a seq2seq model for the word "masa" as input and "masă" as output. The tags <SS> and <SE> mark the start and the end of the sequence. The most prevalent architectures behind the encoders/decoders are the *recurrent* and the *convolutional* neural networks.

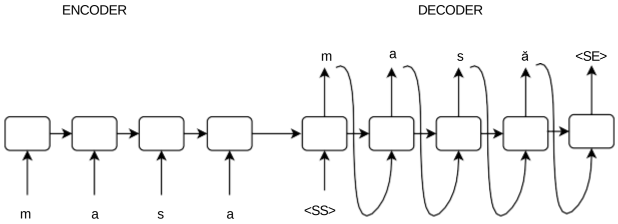

FIGURE 1: Sequence-to-sequence flow

Training Data For training and testing our models [5], we selected a subset of the CoRoLa corpus [35] which contain texts from the belletristic style. We subsequently split the dataset into disjoint training (80%) and testing (20%) sets, each of them being individually shuffled. A few pre-processing steps were performed and include the following operations: converting text to lowercase, striping the digits and punctuation, striping the diacritics, segmenting the text in trigrams, creating pairs of input-target sequences, appending a start-character ("\t") and an end-character ("\n") to the target trigram.

System architectures For our initial tests we selected two ADR systems [36, 37] previously applied for Romanian. The systems were retrained using our dataset, but preserving the original parameter values. Inspired by the architectures described in these two systems, we analyzed four other architectures with various *combinations of recurrent and convolutional layers*. For the implementation, we relied on Keras[8] with the TensorFlow[9] as backend. The networks' hyperparameters were tuned using a small development set. The results were reported in our original research paper [5].

Evaluation and discussion All the 6 system architectures introduced in [5] were evaluated using the *classification accuracy metric*, which is defined as the ratio between the correct predictions and the total number of samples. We computed the accuracy at three different levels: trigram, word and character level. At trigram and word level the accuracy reflects the number of correct predictions made by the system overall. At character level, we computed the accuracy only for the characters which may be written with diacritic symbols *(a, i, s, t)*. The **highest accuracy** 97% was obtained by the CNN based architecture, measured at word level. For our best performing system, we analysed the accuracy for the 4 ambiguous letter sets in Romanian (a-ă - â, i-î, s-s¸, t-¸t) were analyzed individually. For the **pair** *i-î* an **accuracy** of 99.44% was achieved.

0.2 Automatic Romanian lemmatization

Lemmatization is the process of determining the word's dictionary form, called lemma. In the linguistic fields, through lemmatization, all flexional forms of a word are grouped together to be analysed as a single entity. The lemmatization is language dependent and adheres to certain rules. For instance, in Romanian, the lemma of a noun the masculine singular nominative, while a verb's lemma is the infinitive form.

Experimental setup To train the systems presented in [1], we used two different datasets. The first dataset is the Romanian Explicative Dictionary[10](ID: **DEX**) which contains 1.158.194 word forms, each associated with its lemma and part of speech tag. The words are clustered in six major categories based on the part of speech: nouns, adjectives, verbs, pronouns, invariables (adverbs, proper names) and unique forms (interjections, archaic words, Latin names). The second dataset is the CoRoLa[11] corpus [35] which contains texts from different functional styles: belletristic, scientific, publicistic, official. In this work the belletristic subset was chosen. Each word form belongs to one of the thirteen POS categories: noun, verb, adjective, adverb, pronoun, apposition, numeral, conjunction, hyphen, abbreviation, determiner, article and particle. The neural networks were fed with pairs of (word-lemma) and the input data is one-hot encoded.

Ambiguous words. In order to solve the ambiguity problem of words with multiple lemmas, a dictionary of accepted lemmas was built for each dataset. More precisely, we paired each word with a set of lemmas. During evaluation, if the predicted lemma belongs to the given word's lemma dictionary then it was considered to be correct. The lemma dictionary is necessary even when POS context is added as ambiguity can exist within the same POS class when no additional morphological information is given (genre, case or verb tense). For instance:

[8] https://keras.io/

[9] https://www.tensorflow.org/

[10] https://dexonline.ro/

[11] http://CoRoLa.racai.ro/

Torturi	Birthday cakes (Noun. pl.)	→ **Lemma:** tort
Torturi	Torments (Noun pl.)	→ **Lemma:** tortura˘

Results and Discussions In [1] we analysed 24 systems based on deep neural networks in the context of the Romanian lemmatization. The systems were trained on labelled pairs of words and the corresponding lemmas, using at most the part-of-speech tag as morphological information. The input data was one-hot encoded and then passed through the encoder-decoder-based architectures, within a sequence-to-sequence approach. The **highest accuracy** obtained was 99.69% by the CNN based architecture trained with trigrams and with the dictioanry lemma for the ambiguous words. The scope of this study was to use as few lexical input information as possible, as the analysed language offers few corpora with completed annotated texts

0.3 Automatic Romanian Part of Speech tagging

Besides diacritics restoration and lemmatization, another important task in the NLP field is the part of speech tagging. It means to determine the part of speech of a given word, often enriched with morphological or syntactical information. Depending on the annotation level, in the Romanian language we can discriminate three types of tagsets RPOS, MSD and CTAG.

Exeprimental setup The systems presented in [6] were trained using three different datasets:

- The majority of the experiments were performed using the **WPT** [12] dataset [38], developed based on the DexOnline database[13] and Wikipedia[14]. For the RPOS prediction, we trained the systems using the first letter from the MSD tagset provided in **WPT**. For the words with multiple tags, we analysed two scenarios: firstly, we considered as a correct output any POS tag of the word, then we have looked only at the tag of the first occurrence of the word.
- We also used the **DEX** dataset (based on DexOnline database), which pairs each word with the root POS and a word frequency. We took into account only words with positive frequency.
- For the CTAG and the MSD prediction tasks we trained the systems with the **CoRoLa**[15] dataset.

In [6] we analysed the use of **LSTM networks** for the task of predicting the POS of a given word. The encoded input data is passed through a LSTM layer and the result is processed by two stacked dense layers. The second dense layer is the output layer and has as many nodes as the number of possible POS tags. Besides the LSTM architecture, a **sequence-to-sequence model** was implemented for the task of predicted the MSD tag. Both the encoder and the decoder are composed of LSTM layers. We trained the networks over various epochs (from 25 to 100) with different batch sizes (from 256 to 1024) and latent dimensions of the LSTM layers (from 64 to 1024).

[12] http://nlptools.infoiasi.ro/WebPosTagger

[13] https://dexonline.ro/

[14] https://ro.wikipedia.org/

[15] https://corola.racai.ro/

Results and discussions In [6] the performance of the systems was measured using the accuracy metric (as ratio of correct predicted output over the total number of samples). The **best accuracy** (99.18%) was achieved by the LSTM system with a dense layer, predicting only the first letter of the POS.

Medical text data processing

0.4 Topic modelling for identifying medical diagnostic

When it comes to evaluate patients' health, medical doctors analyse different aspects of the person's life, in the so called anamnesis process. Previous diagnostics, family antecedents of a certain illness, different symptoms declared by the patient, together with the personal background and lifestyle contribute to establish an appropriate diagnostic, thus to prescript an adequate medical treatment plan.

Experimental setup and results For the experiments run in [3] we used the medical records taken by a medical physician. The set contains 102 instances, each representing a patient with the clinical observation, the current and past treatments and the patient's response to the treatment. Thus, as type of data, we worked with text (the clinical observation and the prescribed treatment, both in English) and numbers (patient's response to treatment encrypted from 1 = non-responsive to 5 = very responsive). In order to comply with GDPR policies, all the patients' personal information (names, addresses) have been suppressed by the physician before giving us access to the data. Few pre-processing steps were required to prepare the dataset for machine learning algorithms. To generate the relevant textual features, we used word's frequencies: Term Frequency (**TF**), Inverse Document Frequency (**IDF**) and Term Frequency-Inverse Document Frequency (**TF-IDF**). To model the topics present in the analysed texts, we used the Latent Dirichlet Allocation (LDA) and the Latent Semantic Indexing. The first step was to classify the texts using the TF-IDF. The obtained model was fit using 80% of the data, while the predictions were made for the rest of 20% of the dataset. In order to apply the LDA for topic modelling, we created an input corpus and a dictionary, using the Natural Language Toolkit Python toolkit: we first matched each word with an unique ID and then each word ID was mapped with the word's frequency thus obtaining the desired corpus. We used the topic coherence score to determine the number of topics to be passed to the LDA algorithm.

0.5 Personal communication styles analysis

In the era of Big Data when written and recorded audio data are available almost everywhere (from the social networks to the official registered databases) it becomes imperative to address the issue of automatically gaining insights form the collected data either by interpreting the questionnaires responses (to determine different traits, communication styles, psychological personality, future trends in shopping or marketing, etc.) or by predicting or forecasting future events (diagnostics, illness's evolution or remission, stress levels related to contextual situations, suicidal intention, vulnerable categories of people in certain contexts, etc.)

Experimental setup and results In [2] we analysed six machine learning algorithms for the task of classification. More precisely, we used the data obtained from answering of a questionnaire for determining the communication style and try to link one person's stress level with the style of communication. Our dataset contains 220 instances with more than 60 variables. We chosen the questionnaire proposed by Marcus et al. in [39] which classifies a person's communication style into one out of the four communication styles: non-assertive, manipulator, aggressive and assertive by answering a set of 60 questions. Additionally, for the experiments described in [2] each person was asked to measure its stress level as low, medium or high, as the purpose of our study was to analyse the correlation between the stress level and the communication style. We passed the data to six classification algorithms: Decision Tree Based Model, Support Vector Machine, Random Forest, Classification based on instances (k-NN), Naive Bayes and Logistic Regression. We evaluated the learning processing by applying the cross-validation technique. We consequently divided the dataset into k subsets and repeatedly trained the systems using $k-1$ subsets for learning and the last k subset for validating. For each learning iteration, one different subsets was left outside for validation only.

We evaluated these algorithms in terms of accuracy, precision, sensitivity, and specificity. In terms of classifier metrics, the results obtained by our six learners fall within the limits accepted in the literature. If we analyse the results obtained by the accuracy metric, we can conclude that the Random Forest classifier best performs, obtaining an accuracy of **97%**, while the Naive Bayes obtained the poorest results, only **85%**.

Enhancing the Romanian Text-to-Speech Systems

Text-to-speech (TTS), also known as Speech Synthesis, represents a topic of interest for research due to the wide variety of applications in the industry. As TTS aims to create intelligible and natural speech by synthesising a given text, it requires knowledge from various disciplines: linguistics, acoustic, signal processing, machine learning.

0.6 Can synthesised speech data improve the speech expressivity?

In latest research, the naturalness of text-to-speech systems has grown due to the use of the deep learning models. However, the expressivity of the synthesized voices (which is dependent on the existence of expressive corpora) remains a field of interest, especially for the low resourced languages. Romanian is part of the languages with limited data in the field of expressivity, both for voices or datasets.

Experimental setup For our experiments conducted in [7] we created the *MARA*, an expressive dataset freely available online at http://speech.utcluj.ro/marasc/. Furthermore we aimed to analyse the impact of the synthesized speech data to the overall TTS expressivity. First, we divided the dataset into two subset based on the expressivity contained in data: MARA-Flat and MARA-Expr. We trained the TTS systems using only the narrative subset MARA-Flat. From the MARA-Expr audio samples, we extracted the phones duration and the F0 contour which, combined with the spectral parameters generated by the TTS systems for the same utterances, generated the synthetic waveform. To overcome the artefacts of the HTS synthesised data, we applied a postfiltering procedure. Consequently, the output of the HTS system trained with the entire MARA-Flat dataset was paired with the corresponding natural audio files in a voice- conversion manner. This voice-conversion architecture is meant to direct the original input towards the target voice, thus correcting the HTS systems' artefacts.

Evaluation and results The systems were evaluated both through objective and subjective methods. We concluded that no statistically significant differences were found between the systems' objective ratings. Moreover, the listening tests showed that although substituting the natural samples with synthesised copies of them in the training data of an end-to-end TTS system the network is capable of averaging out the spectral artefacts of these samples. However, the main gain of this study was introducing *MARA*, the newly obtained dataset with expressive audio data in Romanian.

0.7 Using Postfiltering to enhance the quality of TTS systems with limited data

For scarcely spoken languages it is difficult to obtain large datasets to train qualitative TTS. The most common approaches to overcome this disadvantage consists in fine-tuning or in adapting the pre-trained model's parameters using data from the target speaker or language, or in appending speaker or language embeddings to the acoustic/linguistic features in order to help the model to learn discriminative features from the training dataset.

Experimental setup In the original research paper [4] we evaluated the postfiltering approach to improve the synthesised voice. Consequently, we trained a TTS system (using various amount of data from different Romanian speakers) and we pass the resulted voice to a postfiltering network to overcome the limited data. We obtained 20 systems which were objectively analysed. Furthermore, we selected 7 systems for a listening test, subjectively analysed by native Romanian speakers. We structured the experiments into two steps:

1. train a DNN TTS with different amount of data
2. apply a postfiltering neural network to enhance the trained output

We used the the SWARA [33] Romanian speech corpus with two additional female voices (*MAR* and *BEA*) recorded for testing purposes, in similar recording conditions and using the same prompts. The input data was feed to the TTS systems are based on the Blizzard Challenge 2017 Merlin [40] setup. The amount of input training data varies from 50 utterances (approx. 5 minutes), 100 utterances (approx. 10 minutes) up to 500 utterances (approx. 50 minutes) and consists in pairs of linguistic and acoustic features. The postfiltering neural networks were based on voice conversion technique (targeting an initial voice to sound like a desired voice) and speaker adaptation (an eigen voice - trained over mixed data from multiple speakers - is targeted to the acoustic features of a certain speaker).

Results and interpretations The systems' performances were evaluated both from the objective and subjective measures perspective. If we analyse the results obtained by the postfiltering systems, we can observe that the MCD scores decreased with 5% to even 7.5%. Artificially doubling the data increased the systems' quality. Doubling the data for both training and postfiltering led to a decrease of 10% for the MCD values. However, doubling the data only for the postfiltering step slightly changed the results. If data for multiple speakers is available, speaker adaptation technique proved to be a solution, leading to lower MCD scores. In spite of these results, when we used multi speaker data only for the postfiltering step, the system obtains results comparable only with the speaker dependent filter.

Conclusions and Future Work

This thesis gathers together machine learning based solution for problems from text processing and speech synthesis.

For the Natural Language Processing (NLP) part, we focused on two directions. On the one hand, we applied several machine learning models to automatize the process of extracting relevant information from the medical records. We analysed both the supervised (text classification [2]) and the unsupervised (document clustering, topic modelling [3]) learning techniques. The experiments were run for written English. On the other hand, we addressed aspects from the text annotation field by applying neural networks based solutions in tasks like automatic diacritics restoration [5], automatic lemmatization [1] or POS tagging [6]. These latest experiments were run for texts written in Romanian.

In our experiments from [1, 5, 6] we trained our deep learning systems in a supervised manner, with labelled pairs of words and their corresponding annotations (lemma, diacritized form or POS tag, depending on the researched task). The input text was encoded and passed to a encoder-decoder architecture. As **future work**, we will focus on analysing different types of neural networks, such as bidirectional LSTMs, GRU, or only attention based architectures (transformers), which are already frequently applied in other text processing tasks. Nevertheless, enriching the input text with more context information may lead to better results in predicting the desired annotation. However, we have to mention that the scope of studies [1,5,6] was to automatically process the input text using the minimum context knowledge, due to the lack of large annotated corpora in Romanian.

Beside automatically annotating texts written in Romanian, we applied the machine learning NLP algorithms to ease the work of the medical physicians. We focused our work on two main directions: determine patients' medical diagnostic through topic modelling techniques [3] and interpreting the psychological questionnaires' results [2]. At first, we created a dataset consisting of the personal records of a family doctor, gathered during the consultations. The dataset contains 102 instances, consisting in written text, more precisely the clinical observations and the prescribed treatment, both in English, and numerical data for the patient's response to treatment. Another direction was working with personality data, by combining cognitive psychology and machine learning. We analysed over 200 instances with more than 60 variables, as each participant at the study was asked to answer to a 60 questions communication style questionnaire together with measuring the personal stress level (low, medium, high). For the obtained dataset and the current task, we trained and tested 6 machine learning classifiers.

As future work for the NLP tasks using medical data, described in [3] and [2] we intend to analyse the impact and the efficacy of other types of machine learning and deep learning algorithms. Secondly, it would be of interest to study medical data written in the Romanian language, having in mind not only the language particularities (diacritics, spelling, etc), but also the challenge of gathering the input data, as Romanian is a scarcely represented language. Moreover, we can apply the systems developed in [1,5,6] to automate the text's annotation and to correct its undiacritised written form.

For the Text to Speech Synthesis part, the aim was on increasing the quality and the expressivity of the synthesised text. We analysed different neural networks architectures using Romanian texts as input. Our approaches are novel in relation to the Romanian Speech Synthesis field and have been published in research articles within conferences proceedings [4] [7].

In our experiments from [4] we researched the potential of postfiltering techniques to enhances the quality of TTS systems with low resourced data input available. We split our approach in two parts: first we trained the TTS systems with different amounts of data, then we applied a postfiltering neural network to enhance the trained output. The amount of the input training data varies from 50 utterances (approx. 5 minutes), 100 utterances (approx. 10 minutes) up to 500 utterances (approx. 50 minutes) and consists in pairs of linguistic and acoustic features. Moreover, we trained two systems with doubled input data, by simply adding twice the initial data, in order to analyse if the quality of the output is influenced rather by the physical amount of data than by the data content. The postfiltering neural networks were based on voice conversion technique (targeting an initial voice to sound like a desired voice) and speaker adaptation (an eigen voice - trained over mixed data from multiple speakers - is targeted to the acoustic features of a certain speaker). These latest experiments were run for texts and audio samples in Romanian and the results are detailed in [4]. As this work was part of the *SINTERO* research project[16] we have to mention that I, in particular, was responsible for recording my own voice for testing (*MAR* voice) and for processing the resulting audio files. Moreover I also dealt with the systems which were trained based on the MAR voice and analysed their results obtained within the experiments from [4].

Research paper [7] analyses different ways to enrich the expressivity of the TTS systems in a low resourced emotional/expressive dataset context. With this purpose in mind, we first created an expressive dataset in Romanian, consisting in an audiobook (*Mara* - written by Ioan Slavici) which was manually segmented in smaller files, following the speaker's phrase break pauses. The written text was annotated through the RACAI Relate Platform with high-level linguistic information as described in Section **??**. Starting with MARA dataset we analysed the impact of synthesized speech data to the overall TTS expressivity. We trained 5 different TTS systems with various amount of expressive data as input: None, synthesised expressive data or all natural expressive data. The impact of expressive data and all the result are described in our research work [7].

As future work for the speech synthesis part, beside training more different TTS network architectures (attention based, transformers), it is of interest to examine other vocoders or to add more features to the postfilter network (lexical or speaker embeddings) in an attempt to enrich the synthesised speech quality. For the expressive TTS systems experiments, we intend to analyse other methods in order to obtain quality expressive synthesise data. Furthermore, we take into account the possibility of the inter-gender prosody transfer.

Thesis as a whole Having in mind all the experiments run for input data written in Romanian language, we plan to encapsulate all the resources and the systems described in the present work in a tool, aiming to help others researchers within their

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

work. More specifically, we intend to create a Romanian text-to-speech system, enriched with expressivity, which will be feed with the input written text preprocessed by the tools analysed in [1, 5, 6] and consisting on the TTS technologies introduced in [4, 7].

Bibliography

[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 Nut¸u. 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 Nu¸tu. 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 Nu¸tu, 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. Nu¸tu, 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 Nu¸tu, 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 Nut¸u, 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*, pages 1–26, 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 Stankevicˇius, Mantas Lukoševicˇius, Jurgita Kapočiū tė-Dzikienė, Monika Briedienė, and Tomas Krilavicˇius. Correcting diacritics and typos with a byt5 transformer model. *Applied Sciences*, 12(5):2636, 2022.

[15] Lukas Pakalniškis. *Giliuoju mokymusi gr˛istas diakritiniu˛ ženklu˛ atstatymas lietuviu˛ 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 ONEAT, Ă, 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 Juricˇic´. *Oznacˇavanje vrsta rijecˇi pomoc´u 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] 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.

[34] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

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

[36] 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.

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

[38] 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.

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

[40] 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).