# ELiRF at ClinAIS Task: Automatic Identification of Sections in Clinical Documents

Pere Marco[1,†], Maria Jose Castro-Bleda[1,*,†], Encarna Segarra[1,2,†] and Lluis Felip Hurtado[1,†]

[1]*VRAIN: Valencian Research Institute for Artificial Intelligence, Universitat Politècnica de València, Camí de Vera s/n, València, 46020, Spain*

[2]*ValgrAI: Valencian Graduate School and Research Network of Artificial Intelligence, Universitat Politècnica de València, Camí de Vera s/n, València, 46020, Spain*

**Abstract**

This paper presents our participation in the ClinAIS task of the IberLEF 2023. We approach the automatic identification of sections in unstructured Spanish clinical documents task as a word sequence classification problem, where the assigned label of each word determines the class of the segment to which it belongs. We use a large-scale biomedical Spanish language model that has been trained from scratch. During the fine-tuning phase, our system assigns to each word the label corresponding to the section to which it belongs. We apply a data augmentation technique based on back-translation in order to introduce variations in phrasing and word choice. We make a hyperparameter search following two different strategies. We present a total of 5 systems, which are the result of different combinations of hyperparameter search strategies and the utilization of data augmentation. The achieved results of our models are highly competitive, ranking us in the first position for this task.

**Keywords**

Natural Language Processing, Sequence Labelling, Transformers-based Models, Spanish Clinical Documents

## 1. Introduction

Identifying medical sections in the patient narratives documented in unstructured clinical documents can help with other processing tasks. For example, it could be applied to the recognition of biomedical named entities, which can be completely different depending on the section they are in. It could also help physicians find information easily, or support an information retrieval system to return specific information.

In the Pomares-Quimbaya et al. work [1] a systematic review of the approaches until 2018 to identify sections within clinical narratives from Electronic Health Records (EHR) was presented. The objective of this work was to report the results of a systematic review concerning approaches

*IberLEF 2023, September 2023, Jaén, Spain*

*Corresponding author.

†These authors contributed equally.

pmarco@vrain.upv.es (P. Marco); mcastro@dsic.upv.es (M. J. Castro-Bleda); esegarra@dsic.upv.es (E. Segarra); lhurtado@dsic.upv.es (L. F. Hurtado)

0009-0003-7026-3543 (P. Marco); 0000-0003-1001-8258 (M. J. Castro-Bleda); 0000-0002-5890-8957 (E. Segarra); 0000-0002-1877-0455 (L. F. Hurtado)

aimed at identifying sections in the narrative content of EHR, using both automatic and semi-automatic methods. Their analysis showed that the most popular Machine Learning methods were Conditional Random Fields (CRF) and Support Vector Machine (SVM). All these works rely on manually created training and test sets, at least partially. Zhou and Li reported in their work [2] a CRF model that combined both lexical and structural features to facilitate section identification for Information Extraction from Chinese Medical Literature. They reported experiments on a human-curated asthma dataset showing that their approach achieved better performance than SVM models.

Rosenthal et al. [3] proposed using sections from the medical literature (e.g., textbooks, journals, web content) that feature content similar to that found in EHR sections. Their approach used data from a different kind of source where labels were provided without the need of a time-consuming annotation effort. They used this data to train two models: a recurrent neural network model and a BERT-based model. They applied the learned models along with source data via transfer learning to predict sections.

More recently, Carrino et al. [4] presented the first large-scale biomedical Spanish language models trained from scratch, using a large biomedical corpus for a total of 1.1B tokens and an EHR corpus of 95M tokens. They fine-tuned the models on three clinical Named Entity Recognition (NER) tasks and compared them with both general-domain and other available Spanish clinical models. The results showed the superiority of their models across the NER tasks, making them competitive candidates for clinical Natural Language Processing (NLP) applications.

In our work, we approach the automatic identification of sections in unstructured Spanish clinical documents task as a word sequence classification problem, where the assigned label of each word determines the class of the segment to which it belongs. To implement our system, we take as a starting point the pretrained model 'PlanTL-GOB-ES/bsc-bio-ehr-es', created by Carrino et al. [4]. For fine-tuning, our system assigns to each word the label corresponding to the section to which it belongs. We made a hyperparameter search following two different strategies. We apply a data augmentation technique based on back-translation (translating the text into another language and then translating it back to the original language) in order to introduce variations in phrasing and word choice, helping the model learn different ways of expressing the same meaning. We present five systems that result from different combinations of hyperparameter search strategies and the use of data augmentation.

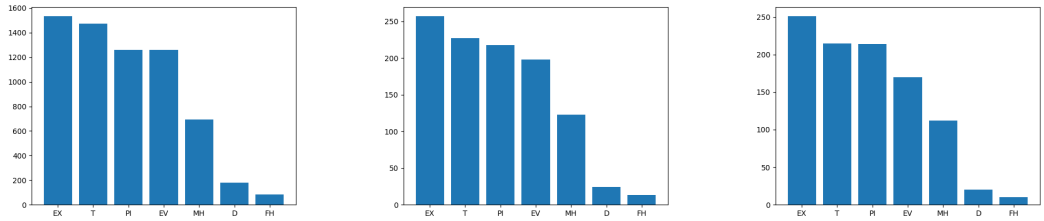
The rest of the paper is organized as follows: in Sections 2 and 3, we present the task, the dataset and the used evaluation metric; in Section 4, the concept of data augmentation; in Section 5, we discuss the proposed system. Then, Section 6 presents the results and error analysis. Finally, we conclude the work in Section 7 and describe what further has to be done.

## 2. Task Description

Labeling sequences is a common task in the domain of NLP and refers to the process of assigning specific labels or tags to individual elements or tokens within a sequence of text. This task is also known as sequence labeling or sequence tagging. The objective of the ClinAIS task [5], presented at IberLEF 2023 [6], is to address the challenge of automatically identifying sections

**Table 1**
Dataset division in splits.

| | Reports | | Sections | |
|---|---|---|---|---|
| Split | Number | (% Total) | Number | (% Total) |
| Train | 781 | (75.24%) | 6,476 | (75.94%) |
| Dev | 127 | (12.24%) | 1,060 | (12.43%) |
| Test | 130 | (12.52%) | 992 | (11.63%) |
| Total | 1,038 | (100%) | 8,528 | (100%) |




Training Development Test

**Figure 1:** Distribution of type of sections in the training, development, and test sets. Sections are ordered from left to right according to their frequency in the training set. Types of sections are: Exploration (EX), Treatment (T), Present Illness (PI), Evolution (EV), Past Medical History (MH), Derived from/to (D), and Family history (FH).

in unstructured Spanish clinical documents. This task is a combination of both segmentation and classification, where the goal is to segment the notes into different continuous sections and correctly classify them based on a predefined set of categories. The task focuses on identifying seven predefined medical sections: Present Illness (PI), Derived from/to (D), Past Medical History (MH), Family history (FH), Exploration (EX), Treatment (T), and Evolution (EV).

## 3. The Dataset and Evaluation Metric

The organizers provided a subset of the CodiEsp [7] corpus for the ClinAIS task. The CodiEsp is a collection of 1,000 unstructured Spanish clinical case reports from different medical specialties. An additional collection of 2,751 unannotated documents was also provided as a background set. The present corpus is a randomly-selected subset of the background CodiEsp corpus, consisting of 1,038 distinct reports. Table 1 and Figure 1 present some of its relevant statistics. As seen in the histograms of Figure 1, it is a very unbalanced dataset. A more detailed description of the dataset is presented in [8].

The task of identifying sections in unstructured clinical notes presents some characteristics that must be taken into account to establish its evaluation. For instance, since the end of one section is always connected to the beginning of another, commonly used evaluation methods would consider two sections as incorrect even if there is a single word error in one of the boundaries. Moreover, the sections are not delimited by paragraphs, lines, or phrases, meaning that a sentence may have more than one section, thus increasing the difficulty of the segmen-

**Table 2**
Levenshtein distance (mean and standard deviation) between the set of 6,476 pairs formed by the original text and the translated text. Number of zeros is the number of identical strings in the set of pairs. The average number of words per section is: Original=48.13, Translation1=47.77, and Translation2=33.73.

| | Mean | Standard deviation | Number of zeros (out of 6,476 pairs) | Percentage of zeros |
|---|---|---|---|---|
| Original - Translation1 | 18.80 | 28.81 | 293 | 4.52% |
| Original - Translation2 | 25.56 | 62.66 | 153 | 2.36% |

tation task. The organizers conducted a thorough analysis of existing metrics and designed the 'B2 evaluation metric', which is an adaptation of the 'boundary distance B' developed by C. Fournier [9], as a means of better evaluating the actual performance in the task.

B2 metric employs a variation of the editing distance with three operations (addition/deletion, substitution, and transposition) and is able to discern segment types. The main advantage is the introduction of the transpose operation, in which the boundary between two sections can be moved by a limited and configurable number of borders instead of performing an insert and a delete operation.

## 4. Data Augmentation

Data augmentation is a technique commonly employed in machine learning to artificially increase the size of a training dataset by applying various transformations or modifications to the existing data. The goal of data augmentation is to enhance the model's ability to generalize and improve its performance. NLP-specific techniques focus on modifying the text while preserving its meaning, coherence, and grammaticality. Common data augmentation techniques applied in NLP include synonym replacement, random word insertion, deletion or swapping, and also back-translation [10, 11]. We used this last technique, that is, translating the text into another language and then translating it back to the original language in order to introduce variations in the phrasing and word choice, helping the model learn different ways of expressing the same meaning.

We used two different automatic translators: Translation1 is performed by using DeepL (https://www.deepl.com/translator). Each section of the training dataset undergoes three translation steps: Spanish to American English, then English to German, and back to Spanish again. Translation2 is performed by using a set of bilingual OPUS-MT translators [12] trained from the Tatoeba Translation Challenge dataset [13]. In this case, each section undergoes two rounds of back-translation: Spanish to English, and back to Spanish; then Spanish to Catalan, and back to Spanish again.

We calculated the Levenshtein distance between the set of original sections in the training data and their corresponding translations as a measure of the dissimilarity between the two sets. The analysis, presented in Table 2, shows that Translation2 exhibits a higher level of differences and variability compared to the texts of Translation1.

# 5. System Description

## 5.1. Overview of the System

For generating our solution, we approach this text segmentation task as a word sequence classification problem, where the assigned label of each word determines the class of the segment to which it belongs.

To implement our system we decided to start from the pretrained model created by Carrino et al. [4], 'PlanTL-GOB-ES/bsc-bio-ehr-es'. For pretraining this model, the authors used two corpora of very different sizes and natures: an EHR corpus and a biomedical one. The 'EHR corpus' contains 95M tokens from more than 514K clinical documents (including discharge reports, clinical course notes and X-ray reports). The 'biomedical corpus' includes Spanish data from a variety of sources for a total of 1.1B tokens across 2,5M documents. The models presented in their work were pretrained from scratch employing a RoBERTa base architecture [14] with 12 self-attention layers.

For fine-tuning, our system assigns to each word the label corresponding to the section to which it belongs. For example, given the following word sequence:

> "Un paciente varón de 25 años miope magno es remitido con el diagnóstico de membrana neovascular subretiniana (MNVSR) en el ojo izquierdo (OI)."

The corresponding groundtruth output would be:

> "PI PI PI PI PI PI PI PI D D D D D D D D D D D D D D D"

Section class PI, denoting 'Present Illness', and class D, representing 'Derived from/to'. Finally, before evaluation, this output label sequence is converted to the output segmented format of the competition.

Since the model accepts inputs of length 512 tokens, we separated the documents into consecutive blocks of 512 tokens, without overlap nor excluding words of the document. After doing some preliminary tests using other approaches, such as considering some overlap between consecutive blocks, we discarded them since they increased the complexity of the problem without improving the results. Another important aspect to mention is the use of some heuristics to improve the results. First, expressions made up of 2 or 3 words with their own meaning and that constitute a section by themselves may appear in the documents, for this reason, we considered sections of length greater than or equal to 2 words, thus allowing the formation of these structures. However, we removed from the results the sections consisting of a single word that were considered as part of the previous section. Second, as the model tokenizer uses subwords, in some cases there are words that are assigned different labels. In these cases, the first label was assigned to this word.

## 5.2. Hyperparameter Optimization

For the model training, we made a hyperparameter search following two different strategies. In the first one, we made an exhaustive search going through all possible combinations among pre-established lists of values for different parameters and selecting the best performance based

**Table 3**
Different combinations of hyperparameter search strategies and data augmentation for the five systems.

| Parameter | System 1 | System 2 | System 3 | System 4 | System 5 |
|---|---|---|---|---|---|
| **Hyper parameter search data** | Original training set | Original training set | Original training set | Original training set + Translation1 | Original training set + Translation2 |
| **Training data** | Original training set | Original training set + Translation1 | Original training set | Original training set + Translation1 | Original training set + Translation2 |
| **Epochs** | 22 | 22 | 20 | 39 | 42 |
| **Learning rate** | 1e-04 | 1e-04 | 1.42e-04 | 1.17e-04 | 8.48e-05 |
| **Batch size** | 8 | 8 | 16 | 16 | 4 |
| **Optimizer** | Adamax | Adamax | AdamW | AdamW | AdamW |
| **Gradient accum. steps** | 1 | 1 | 4 | 16 | 2 |
| **Weight decay** | 0 | 0 | 6.37e-03 | 1.06e-03 | 3.73e-03 |
| **Lr scheduler** | - | - | Linear | Linear | Linear |

on the macro F1 metric results on the validation set. The parameters used were: Adam, SGD, and Adamax as optimizers; learning rates of 1e-4, 1e-5, 1e-6, and 1e-7; and batch sizes of 4, 8, and 16. For the SGD optmizer we assigned 0.9 as momentum value. The rest of parameters were left at default value of each optimizer (see https://pytorch.org/docs/stable/generated/torch.optim.Adam.html, https://pytorch.org/docs/stable/generated/torch.optim.SGD.html, https://pytorch.org/docs/stable/generated/torch.optim.Adamax.html).

In the second strategy, we decided to use Optuna [15], for the hyperparameter search based on the micro F1 results on the validation set. The parameters indicated to Optuna were: the number of epochs, from 10 to 60; the learning rate, from 1e-3 to 1e-7; batch size, among 4, 8, 16, and 32; gradient accumulation steps, among 2, 4, 8, 16, and 32; weight decay, from 1e-4 to 1.5e-2; and earning rate schedule type, between constant and linear. The remaining parameters were kept at the Hugging Face Trainer default values (see https://huggingface.co/docs/transformers/main_classes/trainer#transformers.TrainingArguments).

The unspecified parameters in both strategies were left at pretrained model default values for training (https://huggingface.co/PlanTL-GOB-ES/bsc-bio-ehr-es/blob/main/config.json). Finally, we present five systems that result from different combinations of hyperparameter search strategies and the use of data augmentation. The five systems are described in Table 3, where Systems 1 and 2 were obtained using the exhaustive search optimization, and Systems 3, 4, and 5 were obtained by Optuna hyperparameter search.

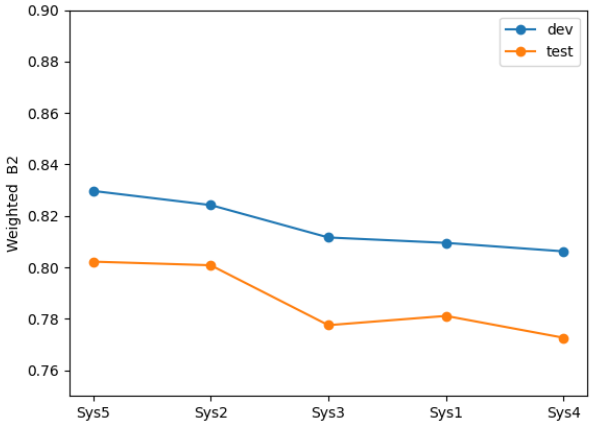



**Figure 2:** Score (weighted B2) on the development and test data of our systems, ordered from left to right according to the development score.

## 6. Experimental Results and Discussion

Our results are presented in Table 4. The best outcome is achieved by System 5, with a test score of 0.8022. The second system obtains a nearly identical score of 0.8008. The remaining three systems perform in a very similar manner. These results can be visualized in Figure 2, where the systems are ranked according to the scores obtained in the validation set. In all cases, the test results are slightly lower than the validation results, as the systems were optimized using the validation data. Furthermore, the positive influence of data augmentation can be observed, particularly when using the data obtained from the second translator. As identified in the analysis of Section 4, Translator 1 did not introduce the necessary difference between the original text and the translation.

Lastly, when conducting a study of the F1-score per class, the results shown in Table 5 are obtained. Systems 5 and 2 demonstrate substantial enhancements in the performance of the two least represented classes, ultimately establishing them as the top-performing systems.

**Table 4**
Score (weighted B2) on the development and test data of our systems.

| # | Data Aug. | Dev Score | Test Score |
|---|---|---|---|
| System 1 | - | 0.8095 | 0.7811 |
| System 2 | ✓ | 0.8242 | 0.8008 |
| System 3 | - | 0.8116 | 0.7775 |
| System 4 | ✓ | 0.8062 | 0.7726 |
| System 5 | ✓ | 0.8297 | 0.8022 |

**Table 5**
F1-score per type of section on the development and test data of our systems.

| # | Data Aug. | Dev F1-Score by class | | | | | | | Test F1-Score by class | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | EX | T | PI | EV | MH | D | FH | EX | T | PI | EV | MH | D | FH |
| System 1 | - | 0.94 | 0.50 | 0.69 | 0.77 | 0.51 | 0.39 | 0.33 | 0.91 | 0.43 | 0.65 | 0.73 | 0.52 | 0.39 | 0.36 |
| System 2 | ✓ | 0.94 | 0.48 | 0.66 | 0.76 | 0.51 | 0.39 | 0.33 | 0.92 | 0.46 | 0.69 | 0.74 | 0.50 | 0.57 | 0.40 |
| System 3 | - | 0.93 | 0.36 | 0.43 | 0.75 | 0.21 | 0.42 | 0.35 | 0.92 | 0.30 | 0.47 | 0.74 | 0.27 | 0.36 | 0.29 |
| System 4 | ✓ | 0.93 | 0.34 | 0.39 | 0.75 | 0.17 | 0.31 | 0.26 | 0.92 | 0.29 | 0.39 | 0.73 | 0.25 | 0.35 | 0.21 |
| System 5 | ✓ | 0.94 | 0.44 | 0.41 | 0.76 | 0.23 | 0.46 | 0.50 | 0.92 | 0.37 | 0.42 | 0.75 | 0.26 | 0.43 | 0.36 |

## 7. Conclusions and Future Work

This study presents our methodologies for automatically identifying sections within unstructured Spanish clinical documents. The task is approached as a word sequence classification problem, where each word is assigned a label to determine its corresponding segment class. To accomplish this, we utilized a pre-trained model consisting of a large-scale biomedical Spanish language model that was trained from scratch.

During the fine-tuning process, we conducted a hyperparameter search employing two distinct strategies. Additionally, a data augmentation technique based on back-translation was applied. We introduced five systems that were the outcome of various combinations of hyperparameter search strategies and the utilization of data augmentation. The performance of our systems yielded highly competitive results, placing us in the top position for this task.

The favorable outcomes obtained in this study showcase the feasibility and potential applicability of the proposed method within real-world scenarios. As a direction for future research, it would be worthwhile to explore the optimization of the hyperparameter search by incorporating the B2 metric.

## 8. Ethics Statement

We have not used additional data to those provided by the competition. The pretrained models used are obtained from HuggingFace models hub, under the Apache License 2.0.

## Acknowledgments

This work is partially supported by MCIN/AEI/10.13039/501100011033, by the "European Union and “NextGenerationEU/MRR”, and by “ERDF A way of making Europe” under grants PDC2021-120846-C44 and PID2021-126061OB-C41. It is also partially supported by the Generalitat Valenciana under project CIPROM/2021/023 and PROMETEO/2020/024, and by the Universitat Politècnica de València under the grant PAID-01-22 for pre-doctoral contracts for the training of doctors.

## References


[1] A. Pomares-Quimbaya, M. Kreuzthaler, S. Schulz, Current approaches to identify sections within clinical narratives from electronic health records: a systematic review, BMC Medical Research Methodology 19 (2019) 155. URL: https://doi.org/10.1186/s12874-019-0792-y. doi:10.1186/s12874-019-0792-y.

[2] S. Zhou, X. Li, Section Identification to Improve Information Extraction from Chinese Medical Literature, in: H. Chen, Q. Fang, D. Zeng, J. Wu (Eds.), Smart Health, Springer International Publishing, Cham, 2018, pp. 342–350. URL: https://doi.org/10.1007/978-3-030-03649-2_34.

[3] S. Rosenthal, K. Barker, Z. Liang, Leveraging Medical Literature for Section Prediction in Electronic Health Records, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 4864–4873. URL: https://aclanthology.org/D19-1492. doi:10.18653/v1/D19-1492.

[4] C. P. Carrino, J. Llop, M. Pàmies, A. Gutiérrez-Fandiño, J. Armengol-Estapé, J. Silveira-Ocampo, A. Valencia, A. Gonzalez-Agirre, M. Villegas, Pretrained biomedical language models for clinical NLP in Spanish, in: Proceedings of the 21st Workshop on Biomedical Language Processing, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 193–199. URL: https://aclanthology.org/2022.bionlp-1.19. doi:10.18653/v1/2022.bionlp-1.19.

[5] I. de la Iglesia, M. Vivó, P. Chocrón, G. de Maeztu, K. Gojenola, A. Atutxa, Overview of ClinAIS at IberLEF 2023: Automatic Identification of Sections in Clinical Documents in Spanish, Procesamiento del Lenguaje Natural 71 (2023).

[6] S. M. Jiménez-Zafra, F. Rangel, M. Montes-y Gómez, Overview of IberLEF 2023: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023), co-located with the 39th Conference of the Spanish Society for Natural Language Processing (SEPLN 2023), CEUR-WS.org, 2023.

[7] A. Miranda-Escalada, A. Gonzalez-Agirre, J. Armengol-Estapé, M. Krallinger, Overview of Automatic Clinical Coding: Annotations, Guidelines, and Solutions for non-English Clinical Cases at CodiEsp Track of CLEF eHealth 2020, in: L. Cappellato, C. Eickhoff, N. Ferro, A. Névéol (Eds.), Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020, volume 2696 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020. URL: https://ceur-ws.org/Vol-2696/paper_263.pdf.

[8] I. de la Iglesia, M. Vivó, P. Chocrón, G. de Maeztu, K. Gojenola, A. Atutxa, An Open Source Corpus and Automatic Tool for Section Identification in Spanish Health Records, Journal of Biomedical Informatics (2023).

[9] C. Fournier, Evaluating Text Segmentation using Boundary Edit Distance, in: ACL 2013 - 51st Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, volume 1, 2013.

[10] J. Wei, K. Zou, EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks, in: Proceedings of the 2019 Conference on Empirical Methods

in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 6382–6388. URL: https://aclanthology.org/D19-1670. doi:10.18653/v1/D19-1670.

[11] S. S. Al-Azzawi, G. Kovács, F. Nilsson, T. Adewumi, M. Liwicki, NLP-LTU at SemEval-2023 Task 10: The Impact of Data Augmentation and Semi-Supervised Learning Techniques on Text Classification Performance on an Imbalanced Dataset (2023). arXiv:2304.12847.

[12] J. Tiedemann, S. Thottingal, OPUS-MT — Building open translation services for the World, in: Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT), Lisbon, Portugal, 2020.

[13] J. Tiedemann, The tatoeba translation challenge – realistic data sets for low resource and multilingual MT, in: Proc. of the 5th Conference on Machine Translation, ACL, 2020, pp. 1174–1182. URL: https://aclanthology.org/2020.wmt-1.139.

[14] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A Robustly Optimized BERT Pretraining Approach, CoRR abs/1907.11692 (2019). arXiv:1907.11692.

[15] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: A Next-generation Hyperparameter Optimization Framework, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2019.