Lupus Alberto: A Transformer-Based Approach for SLE Information Extraction from Italian Clinical Reports

Livia Lilli[1,2,*], Laura Antenucci[1,2], Augusta Ortolan[3], Silvia Laura Bosello[3], Maria Antonietta D'Agostino[3], Stefano Patarnello[1], Carlotta Masciocchi[1] and Jacopo Lenkowicz[1]

[1]Real World Data Facility, Gemelli Generator, Fondazione Policlinico Universitario Agostino Gemelli IRCCS, Rome, 00168, Italy
[2]Catholic University of the Sacred Heart, Rome, 00168, Italy
[3]UOC di Reumatologia, Fondazione Policlinico Universitario A Gemelli IRCCS, 00168 Roma, Italy

Abstract

Natural Language Processing (NLP) is widely used across several fields, such as in medicine, where information often originates from unstructured data sources. This creates the need for automated systems, in order to classify text and extract information from Electronic Health Records (EHRs). However, a significant challenge lies in the limited availability of pre-trained models for less common languages, such as Italian, and for specific medical domains. Our study aims to develop an NLP approach to extract Systemic Lupus Erythematosus (SLE) information from Italian EHRs at Gemelli Hospital in Rome. We then introduce Lupus Alberto, a fine-tuned version of AlBERTo, trained for classifying categories derived from three distinct domains: Diagnosis, Therapy and Symptom. We evaluated Lupus Alberto's performance by comparing it with other baseline approaches, selecting from available BERT-based models for the Italian language and fine-tuning them for the same tasks. Evaluation results show that Lupus Alberto achieves overall F-Scores equal to 79%, 87%, and 76% for the Diagnosis, Therapy, and Symptom domains, respectively. Furthermore, our approach outperformed other baseline models in the Diagnosis and Symptom domains, demonstrating superior performance in identifying and categorizing relevant SLE information, thereby improving clinical decision-making and patient management.

Keywords

Natural Language Processing, Systemic Lupus Erythematosus, Text Classification, Italian Language

1. Introduction

Natural Language Processing (NLP) is used in many applications, such as in the medical domain, where the huge amount of unstructured data sources coming from Electronic Health Records (EHRs) generates the need to develop automated systems for text classification and information extraction. However, employing such methods is challenging due to the scarcity of pre-trained models in less common languages like Italian, and for specific medical domains.

In this study, we explored the Systemic Lupus Erythematosus (SLE), a complex pathology which involves different organ domains and can occur in patients at several levels of severity. For this reason, information about diagnoses, symptoms and therapies are used by physicians to characterize Lupus patients and to make better informed decisions about therapy changes or time for the next contact visit. However, these Lupic features are not always available in a structured format, then there is the need for NLP approaches in order to interpret clinical reports and extract the desired data. Based on the literature, large language models (LLMs) and transformer-based architectures represent the state-of-the-art for EHR classification tasks [1, 2, 3, 4].

This work aims to develop a transformer-based approach to identify SLE information from unstructured EHRs at the Italian Gemelli Hospital of Rome. We then propose Lupus Alberto, a fine-tuned version of Alberto [5], the available BERT-based model for the Italian language trained on Italian tweets. In order to assess the Lupus Alberto performance, we compare it with other baseline approaches, choosing among the BERT-based models available for the Italian language, always fine-tuned on the same tasks.

2. Background

Hospitals may not have structured data sources and often there is a need for advanced and automated approaches for the extraction of specific features from clinical reports. For this reason, there are several studies related to information extraction and text classification in the medical domain, in the context of different diseases and

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy
*Corresponding author.
†These authors contributed equally.
livia.lilli@policlinicogemelli.it (L. Lilli)
0009-0005-3319-7211 (L. Lilli); 0000-0002-4837-447X (S. L. Bosello); 0000-0002-5347-0060 (M. A. D'Agostino); 0009-0008-2765-5935 (S. Patarnello); 0000-0001-6415-7267 (C. Masciocchi); 0000-0002-8366-1474 (J. Lenkowicz)

languages.

Specifically for SLE, we found the work of Deng et al. [6], who applied rule-based and logistic regression to identify SLE patient population from unstructured EHRs in the English language. Also Turner et al. [7] investigated NLP techniques for SLE characterization from clinical notes, by using Bag-of-Words and cTakes to transform input EHR texts into features eligible for Machine Learning algorithms. They then used several models like Neural Networks, Random Forest, Support Vector Machines, Naïve Bayes and Word2Vec Bayesian inversion, for the final text classification. Furthermore, in the studies of Lilli et al. [8], Ortolan et al. [9], a rule-based approach combined with a Bert-based topic modelling, is proposed for the identification of longitudinal features in Italian EHRs of SLE patients.

We then found more recent techniques applied in other pathological contexts in the Italian language, and based on transformers and large language models. For example, the work of Paolo et al. [10] presented a NER transformer-based approach in the lung cancer domain, on Italian EHRs. Additionally, Crema et al. [11] delivered an Italian dataset for the neuropsychiatric domain, training a transformer-based model for NER tasks. About text classification, Torri et al. [12] exploited text classification models to extract relevant clinical variables, comparing rule-based, recurrent neural network and BERT-based models, in the ST-Elevation Myocardial Infarction domain, from an Italian hospital. Finally, Lilli et al. [13] proposed an ensemble of Llama with a Bert-based model, for metastasis classification of Italian EHRs in the Breast Cancer domain.

Based on the previous findings, our study aims to propose a transformer-based approach for the Italian language, specifically for SLE. To this scope, we searched for suitable methods to extract multiple Lupic features from the clinical reports of our Italian hospital. We based on the models delivered by Polignano et al. [5], who trained Albert [14] on Italian tweets, and by Buonocore et al. [15], who proposed transformer-based models, pre-trained on neural-machine translations of English resources and on natively Italian-written medical texts.

3. Methods

3.1. Data Corpus

In this paper, we used data from the SLE Data Mart of the Gemelli Hospital of Rome, which comprises an extensive collection of structured and unstructured data related to Lupus patients. We selected the outpatient clinical reports, considered by physicians as more informative for extracting information like diagnoses, therapies and symptoms. For their length, we also chose to treat EHRs at the paragraph level, complying with the token limit of the BERT models. The final classification was then aggregated on the entire report, through a logical-OR.

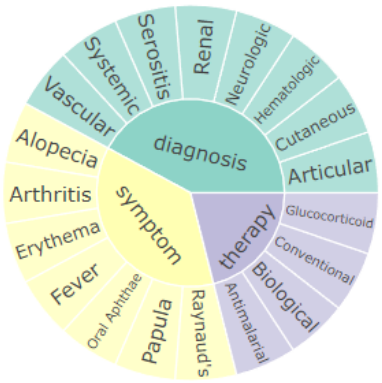


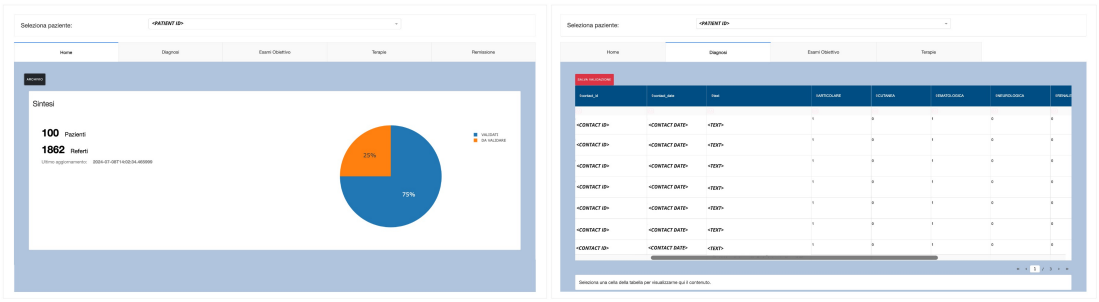
Figure 1: Diversity of the fine-tuned categories. The inner circle shows the three classification domains, while the outer circle represents the related categories.

3.2. Data Annotation

The training set for the fine-tuning consisted of a silver standard made up of annotations from a rule-based algorithm, developed ad hoc for the study [8]. In particular, we formulated rules and expressions for tagging each EHR paragraph with the presence of the categories shown in Figure 1, excluding the possible negations. Rules consist of personalized regex and checks on distances among words.

The gold standard for the evaluation was built by physicians, who annotated a set of EHRs in two steps. Manual annotation was performed by a first team of two physicians with medical knowledge in SLE, who annotated the reports of each patient with respect to the target information. A second team of two specialist rheumatologists reviewed the manual annotations, for the quality assessment. For labelling data, an interactive dashboard was developed ad hoc for the project, where the user assigned to each EHR the corresponding tags. The dashboard URL is accessible only from the hospital's internal network, then it's not sharable. However, Figure 2 provides a screen of the home and annotation pages.

The Inter Annotator Agreement (IAA) among the annotations of the two groups was also computed for a quality assurance measure of data and annotations [16]. For this purpose, we chose the Cohen's Kappa metric, which is a measure of the agreements of two annotators while considering the agreement that could occur by chance [17]:



(a) (b)

Figure 2: Annotation dashboard: (a) home page and (b) annotation page for Diagnosis domain.

$$k = \frac{p_0 - p_e}{1 - p_e} \tag{1}$$

In the Equation 1, p_0 is the observed agreement, while p_e is the expected agreement when both the annotators randomly assign labels, and it is estimated using a per-annotator empirical prior over the class labels [18].

3.3. Fine-Tuning and Classification

This study aimed to extract information about diagnoses, therapies and symptoms from the EHRs of the Gemelli Hospital of Rome. Our purpose was to identify, for each of the three domains, a set of categories provided by our team of rheumatologists, related to SLE. As explained in Figure 1, we then trained our model on 8 different types of diagnoses, 4 therapies, and 7 symptoms.

For this purpose, we fine-tuned AlBERTo [1], a BERT-based model for the Italian language proposed by Polignano et al. [5]. The fine-tuning was performed following the approach of Polignano et al. [5], by treating every category as a singular binary task, with its own training set of labelled texts, randomly sampled from the original data corpus. We then obtained multiple binary classifiers, one for each category to extract.

Fine-tuning and inference were implemented at the paragraph level and not at the entire reports, in order to comply with the token limit imposed by BERT models. The final evaluation was then applied at the overall EHR level, comparing the gold standard reports to the paragraphs' classification, combined at EHR level through a logical-OR. Then if at least a paragraph is positive to a specific category, the corresponding report is classified with that category.

4. Experiments

4.1. Dataset

For this study, we started from the SLE data mart of the Gemelli Hospital of Rome, by selecting among the 13299 available EHRs of outpatient visits.

For our training set, we sampled 1000 training texts for each binary category shown in Figure 1, balancing them among positive and negative samples, such that each category had 50% training samples labelled as positives. The training set was composed of EHR paragraphs, in order to comply with the token limit of 512 tokens imposed by BERT-models.

The gold standard set was composed of 750 EHRs randomly sampled from the data mart, verifying that their paragraphs were not already in the training set. Gold standard set was annotated by two groups of physicians through the annotation dashboard in Figure 2. The same set of gold standard reports were used for the evaluation of all the classification domains.

Details about the dataset are shown in Table 1, where some statistics are reported for each domain, distinguished by training set and gold standard. In particular, for each case are shown the number of categories to classify, the total of paragraphs processed during training and inference, the overall number of EHRs, and the mean of tokens and characters over the paragraphs. Tokens were computed through the BERT tokenizer[2] [19] available on Hugging Face [20].

For privacy reasons, the dataset used in this study is not publicly available. We then provided the descriptive summary metrics in Table 1.

4.2. Inter Annotator Agreement

In order to measure the Inter Annotator Agreement on the gold standards, we used the *cohen_kappa_score* func-

[1] https://github.com/marcopoli/AlBERTo-it

[2] google-bert/bert-base-uncased

Table 1
Statistics of the input dataset, distinguished by set types and domains.

Set Type	Domain	Categories	Paragraphs	EHRs	Mean Tokens	Mean Chars
Training Set	Diagnosis	8	8000	3093	140.9 ± 3.5	387.0 ± 9.7
	Therapy	4	4000	2452	118.9 ± 2.4	306.7 ± 6.5
	Symptom	7	7000	1562	141.5 ± 3.3	395.1 ± 9.2
Gold Standard	Diagnosis	8	6024	790	111.5 ± 2.6	303.7 ± 7.1
	Therapy	4	6024	790	111.5 ± 2.6	303.7 ± 7.1
	Symptom	7	6024	790	111.5 ± 2.6	303.7 ± 7.1

tion provided by the Python Scikit-Learn package [21]. As inputs to the function, we considered the arrays containing the binary annotations performed by the two groups of annotators respectively. Additionally, we performed the analysis grouping the annotations by the three domains: Diagnosis, Therapy and Symptom. Results are shown in Table 2. Staying on the grid proposed by Landis and Koch [22] for the interpretation of the coefficient, we have an almost perfect quality of annotation for the Diagnosis and Therapy domains ($k > 0.80$), and a substantial level for the Symptom case ($k = 0.69$). Although acceptable according to literature standards [16], the latter k score has a lower value than the others, because of the greater difficulty of identifying symptoms from text. Symptoms at current contact are in fact more complex concepts to identify, compared to therapies and diagnoses, which are usually mentioned in the EHR more explicitly. So, even if analyzed by clinical experts, the same report can present inconsistency of annotations, due to the poor quality of text semantics.

Table 2
The Inter Annotator Agreement (IAA) computed between the two groups of physicians, through the Cohen's Kappa metric, distinguished by the three classification domains.

Domain	Cohen's Kappa (k)
Diagnosis	0.88
Therapy	0.93
Symptom	0.69

4.3. Modeling

The AlBERTo fine-tuning was performed through the PyTorch Trainer of the Hugging Face Transformers library [20], using 10 epochs (for further implementation details, see Appendix A). Fine-tuning was performed for each of the 19 categories, in order to obtain a classifier for each binary task.

In order to assess the Lupus Alberto performance, we then compared the model to other baselines, always fine-tuned on the same binary tasks, choosing among several BERT-based models for text classification. Particularly, we considered the three models proposed by Buonocore et al. [15], BioBit[3], MedBit[4] and MedBIT-r3-plus[5], which are pre-trainings on the Italian language, in the medical context. Additionally, we also tried the two base versions of Albert[6] [14], that is the base model used by Polignano et al. [5] to release AlBERTo.

The inference for all the models was performed at the paragraph level instead of the whole report level, and the final classification was aggregated at the EHR level through a logical-OR. Then, if at least a paragraph is positive to the Articular Diagnosis, the overall EHR is classified as positive to that category.

4.4. Results and Discussion

For the evaluation, we compared Lupus Alberto to the other baseline models (fine-tuned on the same tasks), in terms of F-Score at the singular category level. Additionally, to quantify the overall performances, we also computed the mean F-Score for the Diagnosis, Therapy and Symptom domains.

As shown in Table 3, Lupus Alberto presents the highest F-score for the therapy domain, with a value of 87%. Then follow the Diagnosis and Symptom domains with overall metrics of 79% and 76% respectively. These performances reflect the IAA results in Table 2, which shows that Therapy presents a higher quality of annotations compared to Diagnosis and Symptom.

Concerning the baselines, Lupus Alberto outperforms the other experiments for Diagnosis and Symptom, while the Therapy domain presents the higher metric value with the fine-tuned MedBIT-r3-plus [15], whose score equals 88%.

At the singular category level, the Hematologic and Renal diagnoses present the highest performance metrics in their domain, with values of 98% and 94%, respectively. The Glucocorticoid is the therapy with the best F-Score, equal to 97%. Finally, Papula and Raynaud's Phenomenon

[3]IVN-RIN/bioBIT
[4]IVN-RIN/medBIT
[5]IVN-RIN/medBIT-r3-plus
[6]albert/albert-base-v1, albert/albert-base-v2

Table 3
F-Score reported for Lupus Alberto, compared to other baseline models, always fine-tuned on the same tasks. Results are computed for all the categories of the three classification domains. The metric is also reported at the overall domain, for all the experiments.

	lupus-alberto	albert-base-v2	albert-base-v1	bioBIT	medBIT	medBITplus
Diagnosis						
Articular	0,90	0,85	0,92	0,92	0,83	0,92
Cutaneous	0,87	0,80	0,81	0,88	0,92	0,90
Hematologic	0,98	0,96	0,94	0,93	0,96	0,90
Neurologic	0,86	0,57	0,86	0,81	0,79	0,88
Renal	0,94	0,85	0,92	0,85	0,90	0,85
Serositis	0,81	0,65	0,51	0,87	0,66	0,72
Systemic	0,29	0,28	0,07	0,12	0,13	0,07
Vascular	0,69	0,55	0,58	0,63	0,61	0,63
Overall	**0,79**	0,69	0,70	0,75	0,73	0,73
Therapy						
Antimalarial	0,93	0,96	0,94	0,96	0,95	0,93
Glucocorticoid	0,97	0,96	0,95	0,97	0,97	0,97
Conventional	0,91	0,85	0,77	0,9	0,83	0,87
Biological	0,66	0,34	0,45	0,49	0,49	0,73
Overall	0,87	0,78	0,78	0,83	0,81	**0,88**
Symptom						
Oral aphthae	0,66	0,6	0,54	0,47	0,48	0,57
Alopecia	0,65	0,14	0,63	0,74	0,68	0,42
Arthritis	0,83	0,2	0,81	0,77	0,72	0,79
Erythema	0,83	0,84	0,78	0,86	0,83	0,83
Raynaud's Phenomenon	0,87	0,18	0,91	0,19	0,78	0,19
Fever	0,57	0,38	0,52	0,48	0,55	0,54
Papula	0,89	0,76	0	0,94	0,84	0,73
Overall	**0,76**	0,44	0,60	0,64	0,67	0,58

are the best-performing symptoms, with a score equal to 89% and 87% respectively.

In all the three domains, the second version of Albert model present the lowest performance values, with F-Scores equal to 69%, 78% and 44% respectively, if compared to our Lupus Alberto and to the fine-tuned models of Buonocore et al. [15]. Then, as demonstrated from the above results, fine-tuning models specifically trained in the Italian language, improved the final classification performance.

5. Conclusion

This study aims to deliver a transformer-based approach to extract SLE information from real-world data of the Gemelli Hospital of Rome. The scarcity of available models for the Italian language, specialized in Lupus, prompted us to develop a solution to automate the extraction process of SLE information from Italian EHRs. We especially focused on identifying features in the domains of Diagnosis, Therapy and Symptom, reported as of interest for SLE. Our work shows that Lupus Alberto presents competitive performance if compared to other baseline methods, outperforming especially in the classification of information in the Diagnosis and Symptom domains, achieving F-Scores of 79% and 76%, respectively.

6. Limitations

While our proposed approach presents higher performances if compared to the baselines, many aspects could be investigated in future studies, in order to enhance the final performance. This includes the usage of a larger set of training data for the model fine-tuning. Additionally, new research could be conducted by extracting Lupus features through LLMs, and comparing the results with the traditional transformer-based classifiers. Finally, a first release of the Lupus Alberto could be implemented using differential privacy techniques to ensure the protection of data from inference risks [23].

Acknowledgments

For this study, the use of electronic health records was essential for training and testing our new technology.

However, these data contain sensitive patient information and it was fundamental adhering to strict privacy and confidentiality guidelines. To this purpose, the dataset used in this paper was fully de-identified and we received approval from our institution to conduct the presented research. Approval protocol number from the relevant Ethics Committee can be provided on request.

References

[1] Y. Li, S. Rao, J. R. A. Solares, A. Hassaine, R. Ramakrishnan, D. Canoy, Y. Zhu, K. Rahimi, G. Salimi-Khorshidi, Behrt: transformer for electronic health records, Scientific reports 10 (2020) 7155.

[2] V. Yogarajan, J. Montiel, T. Smith, B. Pfahringer, Transformers for multi-label classification of medical text: an empirical comparison, in: International Conference on Artificial Intelligence in Medicine, Springer, 2021, pp. 114–123.

[3] M. Rupp, O. Peter, T. Pattipaka, Exbehrt: Extended transformer for electronic health records, in: International Workshop on Trustworthy Machine Learning for Healthcare, Springer, 2023, pp. 73–84.

[4] Z. Yang, A. Mitra, W. Liu, D. Berlowitz, H. Yu, Transformehr: transformer-based encoder-decoder generative model to enhance prediction of disease outcomes using electronic health records, Nature communications 14 (2023) 7857.

[5] M. Polignano, P. Basile, M. De Gemmis, G. Semeraro, V. Basile, et al., Alberto: Italian bert language understanding model for nlp challenging tasks based on tweets, in: CEUR workshop proceedings, volume 2481, CEUR, 2019, pp. 1–6.

[6] Y. Deng, J. A. Pacheco, A. Ghosh, A. Chung, C. Mao, J. C. Smith, J. Zhao, W.-Q. Wei, A. Barnado, C. Dorn, et al., Natural language processing to identify lupus nephritis phenotype in electronic health records, BMC Medical Informatics and Decision Making 22 (2022) 348.

[7] C. A. Turner, A. D. Jacobs, C. K. Marques, J. C. Oates, D. L. Kamen, P. E. Anderson, J. S. Obeid, Word2vec inversion and traditional text classifiers for phenotyping lupus, BMC medical informatics and decision making 17 (2017) 1–11.

[8] L. Lilli, S. L. Bosello, L. Antenucci, S. Patarnello, A. Ortolan, J. Lenkowicz, M. Gorini, G. Castellino, A. Cesario, M. A. D'Agostino, et al., A comprehensive natural language processing pipeline for the chronic lupus disease, in: Digital Health and Informatics Innovations for Sustainable Health Care Systems, IOS Press, 2024, pp. 909–913.

[9] A. Ortolan, L. Lilli, S. Bosello, L. Antenucci, C. Masciocchi, J. Lenkowicz, P. Cerasuolo, L. Lanzo, S. Piunno, G. Castellino, et al., Pos1142 development and validation of a rule-based framework for automated identification of longitudinal clinical features about systemic lupus erythematosus patients from electronic health records, Annals of the Rheumatic Diseases 2024;83:1014 (2024).

[10] D. Paolo, A. Bria, C. Greco, M. Russano, S. Ramella, P. Soda, R. Sicilia, Named entity recognition in italian lung cancer clinical reports using transformers, in: 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 2023, pp. 4101–4107.

[11] C. Crema, T. M. Buonocore, S. Fostinelli, E. Parimbelli, F. Verde, C. Fundarò, M. Manera, M. C. Ramusino, M. Capelli, A. Costa, et al., Advancing italian biomedical information extraction with transformers-based models: Methodological insights and multicenter practical application, Journal of Biomedical Informatics 148 (2023) 104557.

[12] V. Torri, S. Mazzucato, S. Dalmiani, U. Paradossi, C. Passino, S. Moccia, S. Micera, F. Ieva, Structuring clinical notes of italian st-elevation myocardial infarction patients, in: Proceedings of the First Workshop on Patient-Oriented Language Processing (CL4Health)@ LREC-COLING 2024, 2024, pp. 37–43.

[13] L. Lilli, S. Patarnello, C. Masciocchi, V. Masiello, F. Marazzi, T. Luca, N. Capocchiano, Llamamts: Optimizing metastasis detection with llama instruction tuning and bert-based ensemble in italian clinical reports, in: Proceedings of the 6th Clinical Natural Language Processing Workshop, 2024, pp. 162–171.

[14] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, R. Soricut, Albert: A lite bert for self-supervised learning of language representations, arXiv preprint arXiv:1909.11942 (2019).

[15] T. M. Buonocore, C. Crema, A. Redolfi, R. Bellazzi, E. Parimbelli, Localizing in-domain adaptation of transformer-based biomedical language models, Journal of Biomedical Informatics 144 (2023) 104431.

[16] K. L. Soeken, P. A. Prescott, Issues in the use of kappa to estimate reliability, Medical care (1986) 733–741.

[17] J. Cohen, A coefficient of agreement for nominal scales, Educational and psychological measurement 20 (1960) 37–46.

[18] R. Artstein, M. Poesio, Inter-coder agreement for computational linguistics, Computational linguistics 34 (2008) 555–596.

[19] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[20] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Fun-

towicz, et al., Huggingface's transformers: State-of-the-art natural language processing, arXiv preprint arXiv:1910.03771 (2019).
[21] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., Scikit-learn: Machine learning in python, Journal of machine learning research 12 (2011) 2825–2830.
[22] J. R. Landis, G. G. Koch, The measurement of observer agreement for categorical data, biometrics (1977) 159–174.
[23] M. Miranda, E. S. Ruzzetti, A. Santilli, F. M. Zanzotto, S. Bratières, E. Rodolà, Preserving privacy in large language models: A survey on current threats and solutions, arXiv preprint arXiv:2408.05212 (2024).

A. Implementation Details

The fine-tuning was performed through the PyTorch Trainer[7] of the Hugging Face Transformers library [20], with a desktop GPU Nvidia RTX 5000 Graphics Processing with 16GB of RAM, on a machine with Ubuntu 20.04.3 LTS. The 20% of training set was used as `eval_dataset`, while the remaining was employed as `train_dataset`. The learning rate was set to 2e-5, the batch size to 16, and the weight decay to 0.01.

[7] https://huggingface.co/docs/transformers/main/en/training