LlamaMTS: Optimizing Metastasis Detection with Llama Instruction Tuning and BERT-Based Ensemble in Italian Clinical Reports

Livia Lilli[1,2], Stefano Patarnello[1], Carlotta Masciocchi[1], Valeria Masiello[3], Fabio Marazzi[3], Luca Tagliaferri[3], Nikola Dino Capocchiano[1]

[1] Real World Data Facility, Gemelli Generator, Gemelli Hospital of Rome
[2] Catholic University of the Sacred Heart of Rome
[3] Department of Diagnostic Imaging, Radiation Oncology and Hematology,
UOC of Radiation Oncology, Gemelli Hospital of Rome
livia.lilli@policlinicogemelli.it

Abstract

Information extraction from Electronic Health Records (EHRs) is a crucial task in healthcare, and the lack of resources and language specificity pose significant challenges. This study addresses the limited availability of Italian Natural Language Processing (NLP) tools for clinical applications and the computational demand of large language models (LLMs) for training. We present LlamaMTS, an instruction-tuned Llama for the Italian language, leveraging the LoRA technique. It is ensembled with a BERT-based model to classify EHRs based on the presence or absence of metastasis in patients affected by Breast cancer. Through our evaluation analysis, we discovered that LlamaMTS exhibits superior performance compared to both zero-shot LLMs and other Italian BERT-based models specifically fine-tuned on the same metastatic task. LlamaMTS demonstrates promising results in resource-constrained environments, offering a practical solution for information extraction from Italian EHRs in oncology, potentially improving patient care and outcomes.

1 Introduction

Electronic health records (EHRs) represent the principal data source for hospital centers, housing invaluable information regarding medical histories, treatments, examinations, disease progression and symptoms of a patient. However, efficiently extracting this data with high accuracy and minimal computational resources presents a growing challenge, particularly in the context of the Italian language. While solutions specialized in the clinical domain are readily available for the English language (Lee et al., 2020; Luo et al., 2022; Labrak et al., 2024; Wang et al., 2024), the exploration of similar solutions for the Italian language remains limited, with only a handful of alternatives (Buonocore et al., 2023). Consequently, our objective is to investigate novel approaches that could be implemented in real-world clinical contexts, to extract specific outcomes from Italian textual data. For this purpose, we searched for methods that enable fine-tuning of large language models for specific tasks while minimizing computational resource consumption. Recent studies have showcased the effectiveness of implementing instruction-tuning on pre-trained large language models (Wei et al., 2021; Chung et al., 2022; Liu et al., 2024; Wang et al., 2022), also leveraging techniques such as LoRA (Hu et al., 2022). Through this approach, the number of trainable parameters is reduced, and the model is trained to respond to specific instructions provided during training.

In this paper, we introduce LlamaMTS (Figure 1), a fine-tuned Llama model, through the LoRA instruction tuning technique. Our model is designed to identify the presence of tumoral metastasis by analyzing EHRs from patients diagnosed with breast cancer. Llama was fine-tuned by using as base model Camoscio (Santilli and Rodolà, 2023), which is a Llama adapter for the Italian language, trained on the Italian translation of the Stanford Alpaca Dataset (Taori et al., 2023). To further enhance model performance, we employed an ensemble approach by incorporating a BERT-based model fine-tuned on the same classification task. Additionally, to allow the model to learn from entire EHRs (which may exceed the maximum token limit allowed by Llama during training), we implemented text summarization on both the training and testing datasets. This enabled information extraction from shorter and more concise texts, reducing the noise that long texts may cause.

To evaluate LlamaMTS performances, we compare it with several benchmarks, exploring zero-shot LLMs configurations and fine-tuning known BERT-based model for text classification. Results show that our approach, which leverages instruction-tuning and model ensembling, outper-

Proceedings of the 6th Clinical Natural Language Processing Workshop, pages 162–171
June 21, 2024 ©2024 Association for Computational Linguistics

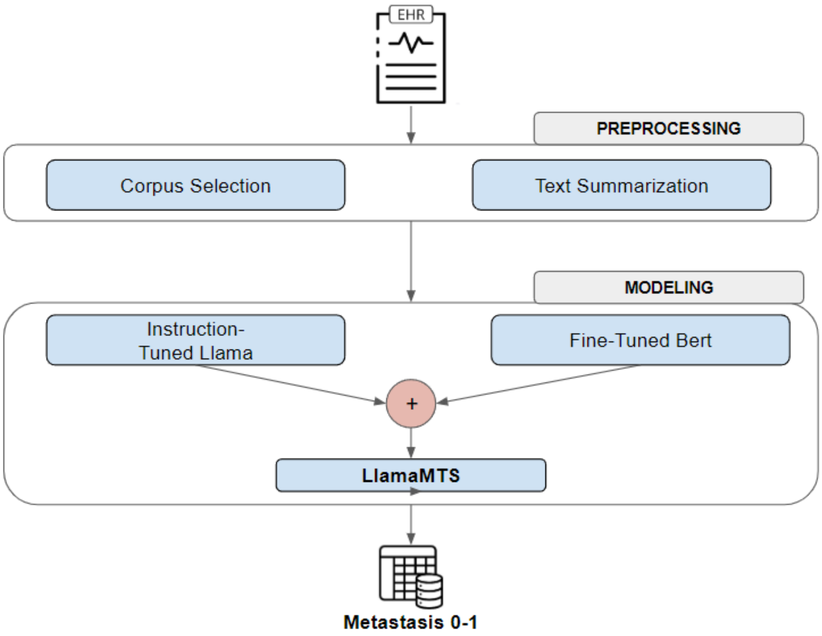

Figure 1: LlamaMTS framework

forms all the other baselines on our metastatic classification task.

2 Background

2.1 Clinical Text Classification

In the Italian language domain, the availability of pre-trained language models for text classification, especially in the clinical field, is currently limited. Notable mentions include BioBit, MedBit and MedBIT-r3-plus, which are different versions of pre-trainings on Italian clinical texts, proposed by Buonocore et al. (2023). In particular, BioBit relies on Italian translations of PubMed abstracts, MedBit is trained on medical textbooks originally written in Italian, while MedBIT-r3-plus is trained on Italian textbooks augmented with web-crawled data. Other works for the Italian language of interest for our study are: AlBERTo (Polignano et al., 2019), an Italian version of BERT (Devlin et al., 2018) trained on Italian tweets, GePpeTto (De Mattei et al., 2020), an Italian fine-tune version of GPT-2 base (117 million parameters), IT5 (Sarti and Nissim, 2022), a T5 model tailored for Italian and BART-IT (La Quatra and Cagliero, 2022), an Italian variant of BART (Lewis et al., 2019). Finally Abdaoui et al. (2020) proposed a set of multilingual models (including the Italian language), pretrained on a reduced number of parameters.

2.2 Instruction Tuning

Recent works demonstrated the efficacy of implementing instruction-tuning on a pre-trained large language model, to increase the downstream performances (Wei et al., 2021; Chung et al., 2022; Liu et al., 2024; Wang et al., 2022). A first step in this direction was made by Taori et al. (2023), who presented Stanford Alpaca, an instruction-tuned version of Llama in the English language. Following this approach, further instruction-tuned Llama models have been trained with LoRA (Hu et al., 2022), as the English Alpaca Lora (Wang, 2023), the Portuguese Cabrita (Larcher et al., 2023) and the Italian Camoscio (Santilli and Rodolà, 2023). In addition to Camoscio, Bacciu et al. (2023) presented Fauno, a language model trained on a corpus of self-chat performed by ChatGPT. Compared to Camoscio, Fauno is a conversational agent for the Italian language. Similarly, Michael (2023) released Stambecco, an instruction-tuned version of LLaMA on a translation to Italian of the GPT-4-LLM dataset (Peng et al., 2023).

This study is inspired by the approach of Hromei et al. (2023), implementing the LoRA instruction-tuning on the Italian Camoscio adapter of Santilli and Rodolà (2023). In this study, the output is represented by extremITLLaMa, a fine-tuning on the EVALITA task (Lai et al., 2023).

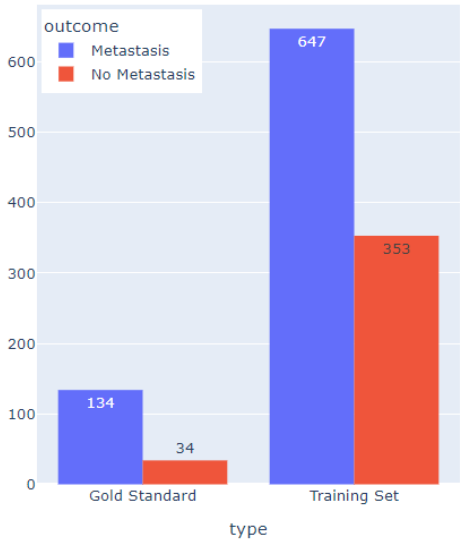

Figure 2: Distribution of metastasis outcome overall the data, distinguished by set type.

2.3 Ensemble

Ensemble is an approach widely used to improve model performance in medical applications, especially in the case of raw data (Nilashi et al., 2022; Doppala et al., 2022; Dutta et al., 2022) and images (Khamparia et al., 2020; Tasci et al., 2021). However, recent works have applied these techniques to the domain of natural language processing, (Yang et al., 2023; Abdennour et al., 2023; Chen et al., 2023; Zhou et al., 2023). In our work, we adopt the approach of Zhou et al. (2023) which combined the BERT predictions with the generated tokens of a large language model to obtain the final ensemble output. We also compared this with the average voting approach of Dutta et al. (2022), where the predicted probability of a class, is a weighted average over all the models.

3 Method

Our methodology involves 1) the selection of the data corpus for fine-tuning, 2) The summarization of EHRs to obtain shorter texts, 3) the instruction tuning of an existing large language model on the metastatic classification task and 4) the ensemble of the obtained instruction-tuned model with a BERT-based model fine-tuned on the same task.

3.1 Data Corpus

In this study, we used EHRs from a data mart consisting of a collection of structured and textual data referencing patients diagnosed with Breast Cancer and being treated at the Italian Gemelli Hospital of Rome.

We selected all the data sources for extracting information relating to tumour metastasis. Guided by a team of physicians, we chose data on clinical diaries, medical histories, and radio-diagnostic reports, because these texts typically contain past and current information about the patient's health status and examination results. We extract all the relevant EHRs for this study from the Gemelli Breast data mart (Marazzi et al., 2021).

3.2 Text Summarization

The EHR length distribution was highly varied and a large portion of the data would risk not being fully processed, due to limits in maximum number of tokens allowed by many large language models.

Additionally, text semantics of clinical reports can be very complex, with relevant information (in this case, the presence or absence of metastasis) not always explicitly reported.

For this reason, we decided to use text summarization methodologies to include data with a reasonable range of tokens, written in a simpler form.

For this purpose, we chose to use Mixtral 8x7B, a pretrained generative Sparse Mixture of Experts language model (Jiang et al., 2024), which outperforms Llama 2 70B (Touvron et al., 2023b) on many benchmarks. To safeguard the confidentiality of clinical reports, we chose to employ locally executable models like Mixtral, thereby excluding the use of GPT (Achiam et al., 2023).

Finally, we formulated an Italian prompt meant to generate a summary of a few words of the input report, retaining all the information relevant to metastasis. We also provided a list of synonymous terminologies as instruction to the model, ensuring a more accurate topic detection. The final prompt was written as follows: *Dato il seguente referto, restituisci una sintesi coincisa in lingua italiana di poche parole, mantenendo tutte le informazioni inerenti a metastasi, lesioni, noduli, attività metabolica o staging:* `{EHR Text}`.

For the implementation of the Mixtral model, we leveraged the Ollama Python library[1].

3.3 Instruction Tuning

During the instruction tuning phase, we leveraged the Camoscio language model proposed by Santilli and Rodolà (2023), who fine-tuned the smallest

[1] https://github.com/ollama/ollama-python

version of Llama (Touvron et al., 2023a) on the Italian translation of Alpaca instruction-tuning dataset (Taori et al., 2023), using the LoRA technique (Hu et al., 2022). Then, following the methodology of Hromei et al. (2023), we merged the Camoscio adapter[2] to the original Llama model and fine-tuned it on our Italian classification task.

The dataset we used has the `instruction`, `input` and `output` fields, where `input` contains the summarized EHRs, the `output` is the binary information about the presence or absence of metastasis, and the `instruction` is written as follows: *Dato il seguente referto medico in italiano, indica con 1 presenza di metastasi e con 0 assenza di metastasi*.

These fields are then put together, for generating the final prompt; we used the same prompter template of Camoscio.

3.4 Ensemble

In order to enhance the final classification performance, we adopted an ensemble approach by combining our instruction-tuned LLM, with the BERT-based model having the best performance among our experiments.

Our approach takes inspiration from Zhou et al. (2023), where the final ensemble prediction corresponds to the one with the highest confidence among the two models, as shown in Equation 1.

$$pred_{ENS} = \begin{cases} pred_{LLM} & \text{if } prob_{LLM} > prob_{BERT} \\ pred_{BERT} & \text{if } prob_{LLM} < prob_{BERT} \end{cases} \quad (1)$$

For the BERT-based model, the confidence $prob_{BERT}$ is the prediction probability related to the predicted class. While for our instruction-tuned model, we considered p_{LLM} as the predicted probability of the generated tokens. Thus, the final ensemble prediction corresponds to the most confident prediction produced by either the two models.

To compare different methods, we also applied a further ensemble technique, the average approach used by Dutta et al. (2022). In this approach, given $M = 2$ models and C=2 classes, we considered: the model output $\mathrm{Y}_j \in \mathbb{R}^C$ for each j^{th}-model, and the confidence values $P_i \in \mathbb{R}^M$ for each i^{th} class with $i \in \{0, 1\}$. So, the final ensemble confidence for a given class k is defined as a weighted combination of all the models:

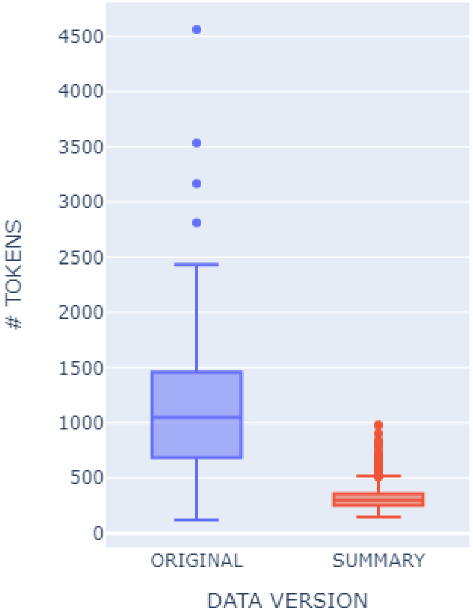

Figure 3: Distribution of Llama tokens for the EHR data.

$$P_k^{ens} = \frac{\sum_{j=1}^{2} P_{kj} \times W_j}{\sum_{i=0}^{1} \sum_{j=1}^{2} P_{ij} \times W_j} \quad (2)$$

In the above equation, W_j is the weight of the j^{th} classifier. Once we have the output $\mathrm{Y} \in \mathbb{R}^C$, which contains the confidence values $P_i^{ens} \in [0, 1]$ computed on the unseen data X, the final prediction will be the i-class, such that: $\arg\max_i \mathrm{Y}(X)$.

4 Experiments

We started by generating the instruction-tuned model on the metastatic classification task, using the summarized EHRs as training data.

We then compared the performance of the instruction-tuned model with several baseline methods, including BERT-based approaches fine-tuned on our classification task and large language models implemented in a zero-shot environment.

Finally, we applied the ensemble techniques between the instruction-tuned model and the BERT-based fine-tuned model with the highest performance in order to obtain the final LlamaMTS classifier.

[2] https://github.com/teelinsan/camoscio

Model	Precision	Recall	Accuracy	AUC	F-Score
Zero-Shot LLM					
Mixtral 8x7B	92,3	71,6	72,6	**74**	80,6
LLaMa2 7B	79,8	1	79,8	50	**88,7**
Camoscio	79,8	1	79,8	50	88,7
BERT-Based Fine-Tuning					
dbmdz BERT	84,5	89,5	78,6	62,4	86,9
BioBIT	84,8	86,6	76,7	62,4	85,6
MedBIT	88,7	88	81,5	**72**	**88,4**
MedBIT-r3-plus	86,3	85	77,4	66	85,7
mBERT-Ita	85	88	78	63,1	86,4
DistilBERT	84,1	87,3	76,8	61,3	85,7
RoBERTa-Ita	79,5	98,5	78,6	49,3	88
BERT-Tiny-Ita	79,5	98,5	78,6	49,3	88
Instruction Tuning					
Instruction-Tuned LlamaMTS	80,2	1	80,4	**51,5**	**89**

Table 1: Results of the intruction-tuned LlamaMTS, compared with the zero-shot large language models and the BERT-based fine-tuning experiments, on the metastatic classification task.

4.1 Data and Summarization

Starting from our selected data corpus, we focused on a subsample of 1168 EHRs, randomly selected from three different data sources, clinical diaries, medical histories, and radio-diagnostic reports.

A total of 168 EHRs (14% of the available data) were annotated by a team of physicians and used as the gold standard for the final evaluation. In contrast, the remaining 1000 EHRs (86% of the overall data) were used as a training set for the model fine-tuning. As shown in Figure 2, the 80% of gold standards (which corresponds to 134 of the 168 EHRs) were positive to the presence of metastasis, while training set had the 65% of positive-labeled samples (that are 647 of the overall 1000 train reports).

We then analyzed the number of tokens in the final texts, using the model tokenizer. Figure 3 shows that the original EHR data has a median of 938 tokens, with first and third quartiles equal to 577 and 1351 respectively and with a maximum value that achieves 4453 tokens.

Considering the maximum number of tokens supported by Llama (2048) (Touvron et al., 2023a), we adopted approaches to reduce the size of the input texts used in our instruction-tuning environment. For this reason, we opted for the text summarization approach using Mixtral 8 x7B (Jiang et al., 2024), which returned a summarized version of the original data, whose tokens' distribution has a median of 301.5, with a first and third quartiles respectively equal to 255 and 360 tokens (as shown in Figure 3). For privacy reasons, we do not report practical examples of summaries, but we provide summary metrics.

4.2 Instruction-Tuned Model

Our first experiment concerns the instruction tuning of the smallest version of Llama (Touvron et al., 2023a) through the Italian adapter of Santilli and Rodolà (2023). Following the Camoscio repository[3], we set up the fine-tuning by first preparing the input base model. We then merged the adapter checkpoints with the original Llama model and then selected 10 epochs for training, using the 1000 summarized clinical texts described in the above paragraph as inputs. We also set the cutoff length at the maximum value supported by Llama, i.e. 2048 tokens.

In the inference phase, we forced the maximum number of generated tokens to 1. We also prefixed the generation of tokens in order to output binary values for classification.

The resulting model represents our instruction-tuned LlamaMTS that will be ensembled with the best-performing BERT-based model to create the final LlamaMTS classifier.

4.3 BERT-Based Fine-Tuning

As baseline experiments, we considered several BERT models available on Hugging Face (Wolf

[3] https://github.com/teelinsan/camoscio

Model	Precision	Recall	Accuracy	AUC	F-Score
Ensemble *Max Method*	88,1	88,8	81,5	72,3	88,8
Ensemble *AUC-Weighted Avg*	88,7	88	81,5	72	88,4
Ensemble *F1-Weighted Avg*	88,8	88,8	**82,1**	**72,3**	**88,8**

Table 2: Results of the ensemble between the instruction-tuned LlamaMTS and the best BERT-based model. The third approach, about the F1-Weighted Average, represents our final LlamaMTS classifier.

et al., 2020) for the Italian language, fine-tuning them on our classification task. The fine-tuning was performed for 10 epochs and the models we chose are pre-trained in the Italian language.

We focused on the work of Buonocore et al. (2023), using their three models (BioBit[4], MedBit[5] and MedBIT-r3-plus[6]), which are different versions of pre-trainings on Italian clinical texts.

Additionally, we explored the work of Abdaoui et al. (2020), fine-tuning their multilingual models[7], pre-trained on a reduced number of parameters.

Finally, we applied other available models trained in the Italian language[8], for further comparisons.

4.4 Zero-Shot LLM

As additional baselines, we considered the classification capability of conversational large language models, forcing the answers to be binary values (meaning presence or absence of metastasis). We chose the two best-performing open-source models, Llama2 (Touvron et al., 2023b) and Mixtral (Jiang et al., 2024), using the Ollama Python library[9], with a prompt in the Italian language. The prompt asks to return an integer number for the given task, where the task is to output a binary value indicating the presence or absence of metastasis in the given text. The final prompt was written as follows: *'Per il seguente task, restituisci solo un numero come risposta, senza ulteriore testo. Dato il seguente referto, rispondi con "1" se è indicata presenza di metastasi, altrimenti rispondi con "0":* `{EHR Text}`.

Whenever other strings are returned in addition to the binary output, then a regex search of the desired values is performed on the generated response, to produce the appropriate binary value.

Moreover, to show the advantage of performing the Llama instruction-tuning, we also applied the Camoscio checkpoints on the same metastatic classification task, with the same inference configuration previously discussed for the instruction-tuned LlamaMTS in subsection 4.2. We chose to focus just on Llama2, as it was the only version available in the Ollama library.

4.5 Ensembling Models

The instruction-tuned LlamaMTS was then combined with the best-performing BERT-based fine-tuned model, to achieve improvements in the final performance metrics. We implemented two different ensemble approaches, as described in subsection 3.4, and considered the ensemble with the best performances as our final LlamaMTS classifier.

In the ensemble experiments, we didn't consider the LLM-based models, because we couldn't compute the corresponding predicted probabilities. For this reason, these models are only used as a baseline benchmark, for a first comparison of the results.

The ensemble results consist of three experiments, where the first one leverages on the approach described by Equation 1, while the second and the third implementations are based on Equation 2, using AUC and F-Score as weights respectively.

4.6 Results and Discussion

Results are measured through the Python Scikit-Learn package (Pedregosa et al., 2011) by computing the typical scores for classification tasks: Precision, Recall, Accuracy, F-Score, and AUC. For the evaluation of the models' performances, we focus on the F-Score and on the AUC metrics, which are typically preferred to Accuracy when the test set is not perfectly balanced among classes. In our case, gold standards present the 80% of positive metastatic samples overall the 168 EHRs.

Table 1 shows that our instruction-tuned LlamaMTS presents the best performances in terms of F-Score, which is 89%. In particular, it presents good sensitivity, that is approximately 100%, and

[4] IVN-RIN/bioBIT

[5] IVN-RIN/medBIT

[6] IVN-RIN/medBIT-r3-plus

[7] Geotrend/bert-base-it-cased, Geotrend/distilbert-base-it-cased

[8] osiria/roberta-base-italian, mascIT/bert-tiny-ita

[9] https://github.com/ollama/ollama-python

a precision of over the 80%. However, we got an AUC of 51.5%, which is lower when compared with the other models. As far as computational resources are concerned, the Llama instruction-tuning spent about 6h 57m 47s, by using an Nvidia RTX 5000 Graphics Processing Unit (GPU) and 16GB of Random Access Memory (RAM).

Among the BERT-based fine-tuned models, MedBIT shows the best metrics in terms of both AUC and F-Score, which are equal to 72% and 88.4% respectively. All the BERT-based experiments present F-Scores over 85%, but an AUC that ranges between 49% and 72%.

With the zero-shot learning of generative large language models, we got the highest results in terms of AUC with Mixtral (74%). Llama2 does not perform well in terms of AUC, that is 50%, though it has a higher F-Score when compared to Mixtral, with a value of 88.7%. Moreover, Llama2 and Camoscio present identical results: this suggests that the adaptation of Llama1 to Italian in Camoscio does not yield superior results compared to the advancements achieved by Llama2, which involved pre-training on a larger Italian corpus.

Additionally, Table 2 shows the results for the three ensemble experiments, performed combining the instruction-tuned LlamaMTS with the fine-tuned MedBit. The approach based on the selection of the highest confident prediction, and the average approach weighted by the F-Score, present the best performances, both having AUC and F-Score equal to 72.3% and 88.8%. Moreover, the F1-average approach has also a higher accuracy of 82.1% (if compared to the 81.5% of the first technique). Then this last method returns the final LlamaMTS classifier, with an AUC that is higher if compared to the instruction-tuned model and MedBit, and with an F-Score that is halfway between the values obtained from the two ensembled models.

5 Conclusions

The instruction-tuning allowed us to specialize an existing large language model on a medical classification task in an optimized fine-tuning environment, using the LoRA approach. Our study shows that LlamaMTS, which is a fne-tuned LLM using LoRA, has higher performance metrics when compared to the base model Camoscio and to other existing approaches that involve conversational LLMs and BERT-Based checkpoints (Table 1). Indeed, the instruction-tuned classifier tends to identify well all the existing positives, even if with low performances in distinguishing the negative samples. This is reflected in the high F-score of 89% and the low 51.5% AUC. We then applied the ensemble technique, combining the classification capability of the instruction-tuned model, with the best-performing BERT-based fine-tuned model. Thus we obtained our final LlamaMTS classifier, which outperforms both the models in terms of AUC, achieving a value of 72.3%, and with an F-Score of 88.8%, close to that of instruction-tuned model.

With this work, we extended advanced NLP techniques on clinical EHR data, automating processes through the usage of powerful language models, trained in the Italian language, on a specific classification task, for the extraction of the tumor metastasis information from EHRs. We proposed an approach that is easily portable to other kinds of outcomes, for extracting information not necessarily available in a structured format, from textual EHRs. Furthermore, the instruction-tuning approach enables fine-tuning large language models in reasonable time frames, leveraging mid-range computational resources.

Limitations

While our study presents promising results for metastasis classification in Breast cancer patients, several limitations may be investigated in future research. These include the application of the model to new outcomes beyond metastasis and its adaptation to both binary and multi-classification tasks. Additionally, new work could be focused on testing the portability of the model by evaluating its performance on EHRs from new hospitals. Furthermore, improvements in model performance could be explored through extended fine-tuning on additional epochs and training data.

Ethics Statement

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.

Acknowledgements

This study received partial funding from Italian Ministry for University and Research (MUR) under the Program PON "Research and Innovation" supporting the development of artificial intelligence platform Gemelli Generator at Policlinico Universitario A. Gemelli IRCCS".

References

Amine Abdaoui, Camille Pradel, and Grégoire Sigel. 2020. Load what you need: Smaller versions of multilingual bert. *arXiv preprint arXiv:2010.05609*.

Ghada Ben Abdennour, Karim Gasmi, and Ridha Ejbali. 2023. Ensemble learning model for medical text classification. In *International Conference on Web Information Systems Engineering*, pages 3–12. Springer.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Andrea Bacciu, Giovanni Trappolini, Andrea Santilli, Emanuele Rodolà, Fabrizio Silvestri, et al. 2023. Fauno: The italian large language model that will leave you senza parole! In *IIR 2023*. Pisa.

Tommaso Mario Buonocore, Claudio Crema, Alberto Redolfi, Riccardo Bellazzi, and Enea Parimbelli. 2023. Localizing in-domain adaptation of transformer-based biomedical language models. *Journal of Biomedical Informatics*, 144:104431.

Chao-Yi Chen, Kao-Yuan Tien, Yuan-Hao Cheng, and Lung-Hao Lee. 2023. Ncuee-nlp at semeval-2023 task 7: Ensemble biomedical linkbert transformers in multi-evidence natural language inference for clinical trial data. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 776–781.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Lorenzo De Mattei, Michele Cafagna, Felice Dell'Orletta, Malvina Nissim, and Marco Guerini. 2020. Geppetto carves italian into a language model. *arXiv preprint arXiv:2004.14253*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Bhanu Prakash Doppala, Debnath Bhattacharyya, Midhunchakkaravarthy Janarthanan, Namkyun Baik, et al. 2022. A reliable machine intelligence model for accurate identification of cardiovascular diseases using ensemble techniques. *Journal of Healthcare Engineering*, 2022.

Aishwariya Dutta, Md Kamrul Hasan, Mohiuddin Ahmad, Md Abdul Awal, Md Akhtarul Islam, Mehedi Masud, and Hossam Meshref. 2022. Early prediction of diabetes using an ensemble of machine learning models. *International Journal of Environmental Research and Public Health*, 19(19):12378.

CD Hromei, D Croce, V Basile, R Basili, et al. 2023. Extremita at evalita 2023: Multi-task sustainable scaling to large language models at its extreme. In *CEUR WORKSHOP PROCEEDINGS*, volume 3473, pages 1–9.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Aditya Khamparia, Aman Singh, Divya Anand, Deepak Gupta, Ashish Khanna, N Arun Kumar, and Joseph Tan. 2020. A novel deep learning-based multi-model ensemble method for the prediction of neuromuscular disorders. *Neural computing and applications*, 32:11083–11095.

Moreno La Quatra and Luca Cagliero. 2022. Bart-it: An efficient sequence-to-sequence model for italian text summarization. *Future Internet*, 15(1):15.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains. *arXiv preprint arXiv:2402.10373*.

Mirko Lai, Stefano Menini, Marco Polignano, Valentina Russo, Rachele Sprugnoli, Giulia Venturi, et al. 2023. Evalita 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for italian. In *Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR. org, Parma, Italy*.

Celio Larcher, Marcos Piau, Paulo Finardi, Pedro Gengo, Piero Esposito, and Vinicius Caridá. 2023. Cabrita: closing the gap for foreign languages. *arXiv preprint arXiv:2308.11878*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Fei Liu, Xi Lin, Qingfu Zhang, Xialiang Tong, and Mingxuan Yuan. 2024. Multi-task learning for routing problem with cross-problem zero-shot generalization. *arXiv preprint arXiv:2402.16891*.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*, 23(6):bbac409.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft.

Fabio Marazzi, Luca Tagliaferri, Valeria Masiello, Francesca Moschella, Giuseppe Ferdinando Colloca, Barbara Corvari, Alejandro Martin Sanchez, Nikola Dino Capocchiano, Roberta Pastorino, Chiara Iacomini, et al. 2021. Generator breast datamart—the novel breast cancer data discovery system for research and monitoring: Preliminary results and future perspectives. *Journal of Personalized Medicine*, 11(2):65.

Michael. 2023. Stambecco: Italian instruction-following llama model. https://github.com/mchl-labs/stambecco.

Mehrbakhsh Nilashi, Rabab Ali Abumalloh, Behrouz Minaei-Bidgoli, Sarminah Samad, Muhammed Yousoof Ismail, Ashwaq Alhargan, and Waleed Abdu Zogaan. 2022. Predicting parkinson's disease progression: evaluation of ensemble methods in machine learning. *Journal of healthcare engineering*, 2022.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.

Marco Polignano, Pierpaolo Basile, Marco de Gemmis, Giovanni Semeraro, and Valerio Basile. 2019. AlBERTo: Italian BERT Language Understanding Model for NLP Challenging Tasks Based on Tweets. In *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019)*, volume 2481. CEUR.

Andrea Santilli and Emanuele Rodolà. 2023. Camoscio: An italian instruction-tuned llama. *arXiv preprint arXiv:2307.16456*.

Gabriele Sarti and Malvina Nissim. 2022. It5: Large-scale text-to-text pretraining for italian language understanding and generation. *arXiv preprint arXiv:2203.03759*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Erdal Tasci, Caner Uluturk, and Aybars Ugur. 2021. A voting-based ensemble deep learning method focusing on image augmentation and preprocessing variations for tuberculosis detection. *Neural Computing and Applications*, 33(22):15541–15555.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

E. J. Wang. 2023. Alpaca-lora. https://github.com/tloen/alpaca-lora.

Xidong Wang, Nuo Chen, Junyin Chen, Yan Hu, Yidong Wang, Xiangbo Wu, Anningzhe Gao, Xiang Wan, Haizhou Li, and Benyou Wang. 2024. Apollo: Lightweight multilingual medical llms towards democratizing medical ai to 6b people. *arXiv preprint arXiv:2403.03640*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Han Yang, Mingchen Li, Yongkang Xiao, Huixue Zhou, Rui Zhang, and Qian Fang. 2023. One llm is not enough: Harnessing the power of ensemble learning for medical question answering. *medRxiv*, pages 2023–12.

Weipeng Zhou, Dmitriy Dligach, Majid Afshar, Yanjun Gao, and Timothy A Miller. 2023. Improving the transferability of clinical note section classification models with bert and large language model ensembles. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2023, page 125. NIH Public Access.

A Implementation Details

The LlamaMTS model is trained with the LoRA Parameter-efficient Finetuning technique (Hu et al., 2022), using the Hugging Face Transformers and PEFT libraries (Wolf et al., 2020; Mangrulkar et al., 2022) and the Camoscio repository[10]. Specifically, our model is trained for 10 epochs on a desktop GPU Nvidia RTX 5000 Graphics Processing with 16GB of RAM, on a machine with Ubuntu 20.04.3 LTS. Training is implemented with batches of dimension 8 and gradient accumulation to obtain a final "virtual batch" of 128. The maximum length used for training is 2048 tokens. The learning rate is set to 3 x 10-4 with AdamW (Loshchilov and Hutter, 2018) and a total of 100 warmup steps are performed. We used a lora_r (i.e., the dimensionality of the low-rank update of the matrices) equal to 16. As base model, we merged the Camoscio adapter to the LLaMA 7 billion checkpoint[11]. In the evaluation we limited the max_new_tokens parameters to 1, forcing values to be binary through the prefix_allowed_tokens_fn parameter.

The BERT-based models are fine-tuned by using 10 epochs, 16 batches and a learning rate of 2 x 10-5.

[10] https://github.com/teelinsan/camoscio

[11] decapoda-research/llama-7b-hf