

Leveraging Large Language Models to Geolocate Linguistic Variations in Social Media Posts

Davide Savarro^{1,*}

Davide Zago^{1,*}

Stefano Zoia^{1,*}

¹Department of Computer Science, University of Turin,
Corso Svizzera 185, 10149 Torino, Italy
{davide.savarro, davide.zago, stefano.zoia}@unito.it

Abstract

Geolocalization of social media content is the task of determining the geographical location of a user based on textual data, that may show linguistic variations and informal language. In this project, we address the GeoLingIt challenge of geolocalizing tweets written in Italian by leveraging large language models (LLMs). GeoLingIt requires the prediction of both the region and the precise coordinates of the tweet. Our approach involves fine-tuning pre-trained LLMs to simultaneously predict these geolocalization aspects. By integrating innovative methodologies, we enhance the models' ability to understand the nuances of Italian social media text to improve the state-of-the-art in this domain. This work is conducted as part of the Large Language Models course at the Bertinoro International Spring School 2024. We make our code publicly available on GitHub¹.

1 Introduction

GeoLingIt is the first shared task focused on geolocating linguistic variation in Italy using social media posts with non-standard Italian language. Part of the EVALITA evaluation campaign, it aims to advance natural language processing (NLP) for non-standard Italian and provide sociolinguistic insights through quantitative analysis.

Social media offers a unique opportunity to study informal language use across sociolinguistic dimensions, particularly diatopic variation (variation across geographic space). Italy's linguistic diversity includes numerous local languages, dialects, and regional varieties of Standard Italian. Online, Italians often use local language elements to signal social identities.

GeoLingIt seeks to understand linguistic variation in Italy by developing methods to predict

the locations of Twitter posts based solely on linguistic content. Unlike other geolocation tasks, it filters posts for non-standard Italian, focusing on linguistic rather than lexical variations. Variations in GeoLingIt data may include local words, code-switching, or entire posts in a local language or dialect.

GeoLingIt includes two subtasks. Coarse-grained geolocation, named subtask A, consists in the classification of the region of provenance of a given text. Fine-grained geolocation, named subtask B, is a double regression task, in which the algorithm must predict the coordinates of the given social media post.

Our approach involves fine-tuning pre-trained large language models (LLMs) to solve both sub-tasks in a single generation step. Our methodology is inspired by ExtremITA (Hromei et al., 2023), where a multi-task approach was used to train LLMs to solve all EVALITA tasks. We fine-tune and compare three decoder-only LLMs on GeoLingIt and analyze the obtained results.

2 Data and models

In this section we describe the GeoLingIt dataset and we outline the characteristics of the three LLMs used in our experiments. We also discuss the pre-processing steps and the methodologies employed for fine-tuning these models.

2.1 Dataset description

The GeoLingIt dataset contains 15039 samples and it is divided in a specific train-evaluation-test split of 13669 (90.09%), 552 (3.67%) and 818 (5.44%) samples, to enable a fair comparison between the different approaches. Furthermore, the dataset is divided into 2 subsets that correspond respectively to subtask A and B. Since our approach tackles both tasks simultaneously, we join the two portions into a new merged dataset and we use it during

*Equal contribution.

¹<https://github.com/dawoz/geolingit-biss2024>

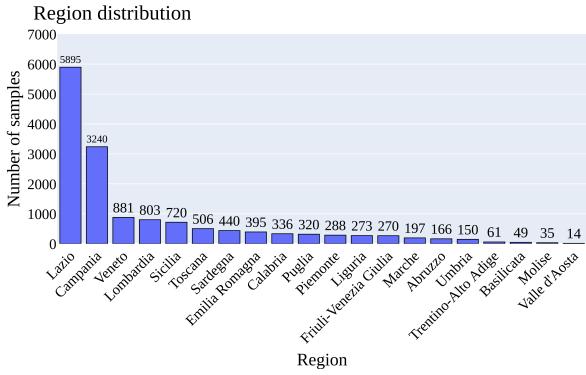


Figure 1: Bar chart representing the number of posts for each label (region of provenance). The labels (x-axis) are sorted by decreasing frequency (y-axis).

training. Table 1 shows some of the samples of the resulting set of data.

Regarding the prediction of the region of provenance (subtask A), the dataset contains 20 labels, which correspond to all the regions of Italy. The distribution of such labels is very nonuniform, as shown in Figure 1: 39.20% of the samples are labeled with the *Lazio* region, and 21.54% with the *Campania* region, the two subsets alone constituting the 60.74% of the dataset.

The locations of the social media posts are displayed in the map in Figure 2a, in which different regions are filled with different colors. Each point corresponds to a specific latitude-longitude pair associated to a sample. From the figure it is evident that the frequency of posts increase in highly populated areas, like in the regional capitals. Figure 2b shows in the geolocation map the density of posts in each region. As a rule of thumb, regions with higher number of posts are regions with higher population, but not all densely populated regions are proportionally represented in the dataset (e.g. *Lombardia*, which has the highest population).

2.2 Models description

In recent years pre-trained LLMs have been effectively used to solve varied problems in natural language processing. In our setting, we fine-tune and compare three different pre-trained LLMs on the Italian language: Camoscio-7B (Santilli and Rodolà, 2023), ANITA-8B (Polignano et al., 2024) and Minerva-3B (Orlando et al., 2024).

Camoscio is an Italian instruction-tuned 7 billion parameters model based on LLaMA (Touvron et al., 2023). The training of the model follows the one of (Taori et al., 2023) with Low-Rank Adaptation (LoRA) (Hu et al., 2021). We use a slightly

modified version of that model implemented by the ExtremITA (Hromei et al., 2023) team.

ANITA is a 8 billion parameter instruction-tuning of the LLaMAntino family (Basile et al., 2023). This model was obtained by fine-tuning LLaMA 3 (AI@Meta, 2024) with Direct Preference Optimization (DPO) (Rafailov et al., 2024). ANITA aims to be a multilingual model to be used for further fine-tunings on Italian language tasks.

Minerva is the first family of LLMs pre-trained from scratch on the Italian language. This set includes three model sizes of 350 millions, 1 billion and 3 billions parameters respectively. We include Minerva in our study to test the capability of a pre-trained LLM on Italian, and we use the largest (3 billion) version to obtain a comparison that is as fair as possible with the other larger models.

3 Experiments and result analysis

We will now briefly present the hardware used, the experiments performed and we will analyze the metrics and the results we obtained.

3.1 Hardware

All our experiments are run on a single machine equipped with a *Tesla T4* GPU with 16 GB of VRAM. Due to the limited amount of compute power, we had to make some compromises when choosing how to train and test our models (see subsection 3.2).

3.2 Experiment settings

All the previously mentioned models have been fine-tuned and tested on the GeoLingIt dataset following the ExtremITA instruction encoding:

```
<instruction> <post_content> [region] <region> [geo] <lat> <long>
```

In this way, both subtasks A and B have been solved with a single model fine-tuning with the language modeling objective.

The upper panel of Table 2 shows the hyperparameters we used for fine-tuning across all the experiments except for Minerva, for which we used a larger batch and minibatch size. Due to our limited disposal of GPU memory of 16 GB, we used gradient accumulation to simulate larger batch sizes and stabilize training.

To further reduce the memory footprint we used 4-bit quantized versions of the three considered

Id	Text	Region	Latitude	Longitude
280	[USER] A suma bin ciapa'! meglio alleggerire un attimo	Piemonte	45.0729	7.6758
286	[USER] Ce ripigliamm tutt chell ch è o nuost	Campania	40.8541	14.2435
500	[USER] Sta bon, vecio!	Veneto	46.1572	12.2865

Table 1: Some samples of the GeoLingIt dataset. The column *Text* includes the content of the post. This can contain urls, user tags and images that are replaced with placeholders. The region of provenance is annotated under the column *Region*, and the coordinates under the columns *Latitude* and *Longitude*.

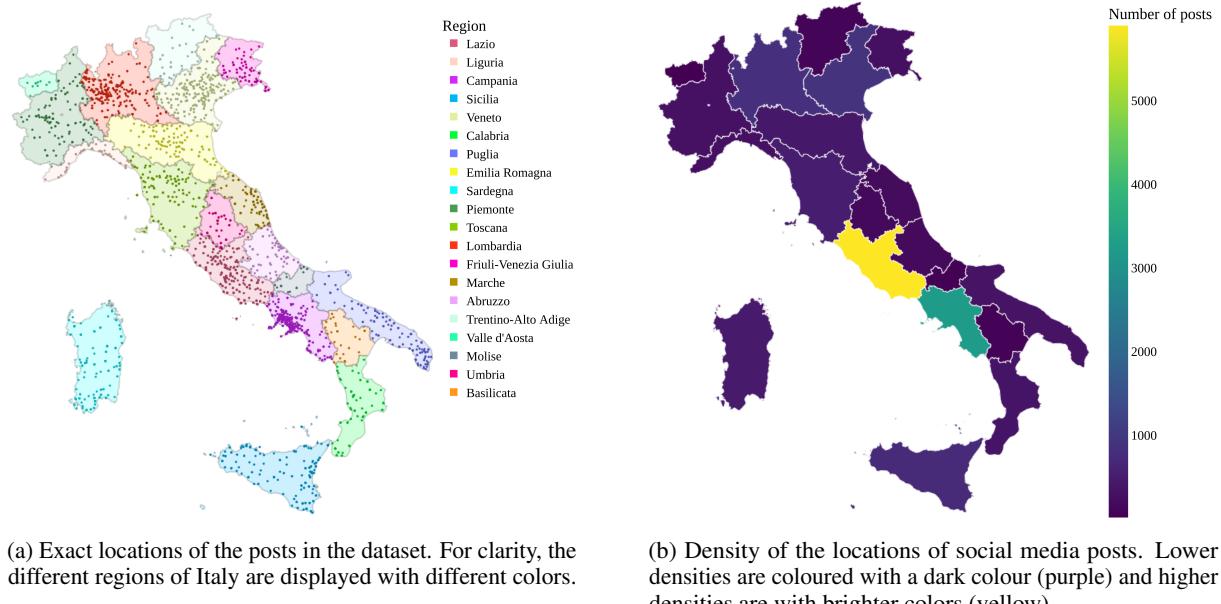


Figure 2: Geographical distribution of the social media posts in the GeoLingIt dataset.

models, and we used LoRA to reduce the number of training parameters and therefore the training time. Although smaller models such as Minerva (3 billion parameters) would have benefited from a more conservative LoRA configuration (e.g. higher rank values or greater bit precision), those same settings would not have been viable for Camoscio and ANITA with our computational resources. For these reasons, we kept the same configuration for all the tested models. The lower panel of Table 2 sums up the parameters we used for LoRA with quantization.

It is worth noting that we performed 10 epochs of training for each of the models analyzed in order to compensate for the aggressive LoRA settings. In this way we were able to still get some noteworthy results when comparing them with the ones presented in (Hromei et al., 2023).

3.3 Metrics and result analysis

We choose to use the same metrics that were adopted for the EVALITA 2023 campaign to evaluate our models and enable a fair comparison with the other competitors: macro (unweighted) F1-

	Parameter	Value
Fine-tuning	N Epochs	10
	Batch Size	32 (64*)
	M. Batch Size	8 (32*)
	Learning Rate	$3 \cdot 10^{-4}$
	Warmup Ratio	0.1
QLoRA	R	8
	Alpha	16
	Dropout	0.05
	N. Bits	4

Table 2: Hyperparameters used for fine-tuning (* parameters used for the Minerva model). The parameters used for quantized LoRA are shown in the bottom part of the table.

Model	F1-score (macro)	Avg Km
Camoscio-7B	0.4935	124.35
ANITA-8B	0.5411	103.06
MINERVA-3B	0.4704	125.35
W_EVALITA23	0.6630	97.74

Table 3: Experiment results, compared with the best-performing models in EVALITA 2023 (Ramponi and Casula, 2023)

score for subtask A, and average distance error in km for subtask B.

The average distance error is a precise and intuitive metric when comparing the regressive capabilities of the models in predicting the right coordinates. On the other hand, we argue that the macro F1-score is not the best metric for evaluating region classification capabilities in our setting. Given the fact that the GeoLingIt dataset is highly unbalanced (see subsection 2.1 for details), we expect that most represented classes would show lower prediction error, and, on the contrary, samples of less represented classes would be more often misclassified. We believe that the F1-score weighted by class cardinality (also known as "micro") is a more appropriate choice of metric, and envision additional experiments with this considerations.

Analyzing the results reported in Table 3 it is interesting to see how a 1 year evolution of LLM research actually influences the performances. We can assert the performance of Minerva is comparable with Camoscio even though Minerva is less than half the size of Camoscio. Moreover, ANITA produced very strong results, almost matching the ones obtained by the top performing models of the EVALITA 2023 campaign.

4 Error analysis

For subtask A (region classification), the models showed varying performance in classifying the region of provenance, particularly struggling with less represented regions in the dataset (e.g. *Trentino-Alto Adige*). Due to space reasons we include the confusion matrices of the test set classification in the Appendix. Figures 3a, 3b and 3c reveal that the most frequent errors occurred in predicting regions with fewer samples. As expected, the models tend to favor regions with higher representation, such as *Lazio* and *Campania*, leading to biased predictions. It is still interesting to see that often misclassifications happen between nearby re-

gions (e.g. *Umbria* and *Lazio* or *Piemonte* and *Lombardia*) highlighting linguistic similarities that are difficult for the LLM to distinguish.

In subtask B (coordinate regression), predicting precise coordinates posed a challenge due to the fine-grained nature of the task. Figures 4b, 4d and 4f show the average distance errors in Italian provinces. It is evident from the figures that while the models could approximate the general area, pinpointing exact locations was difficult. ANITA-8B outperformed the others, likely due to its larger parameter size and advanced training techniques.

Observing the sum of distance errors evidences how much the models predicted distant values despite the number of samples, as shown in figures 4a, 4c and 4e. From these figures we can see that non negligible amounts of error also come from frequently represented areas (see Figure 2b).

5 Conclusions

This project tackled the challenge of geolocating social media posts written in non-standard Italian using large language models (LLMs). By fine-tuning pre-trained LLMs, we aimed to predict both the region and precise coordinates of tweets, addressing the GeoLingIt subtasks simultaneously. Our experiments with Camoscio-7B, ANITA-8B, and Minerva-3B demonstrated encouraging results in geolocation accuracy. Despite the dataset's non-uniform distribution posing challenges, our experiments showed that recent advancements in LLMs significantly enhance geolocation performance, with the ANITA-8B model achieving the best results among the tested models.

Future work could explore advanced pre-processing techniques (e.g. including additional information in the prompt, like the coordinate centroid of a given region) and different models to further improve geolocation accuracy. Additionally, class imbalance issues could be addressed with data augmentation or re-sampling techniques.

Overall, this project contributes to the research on using LLMs for sociolinguistic analysis and geolocation tasks, offering a foundation for developing more accurate and sophisticated models.

Acknowledgements

We thank the Department of Computer Science of the University of Turin for providing the access to HPC4AI with the necessary computational resources for running the experiments of this work.

References

AI@Meta. 2024. Llama 3 model card.
[https://github.com/meta-llama/llama3/
blob/main/MODEL_CARD.md](https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md).

Pierpaolo Basile, Elio Musacchio, Marco Polignano, Lucia Siciliani, Giuseppe Fiameni, and Giovanni Semeraro. 2023. Llamantino: Llama 2 models for effective text generation in italian language. *arXiv preprint arXiv:2312.09993*.

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. Mirko Lai, Stefano Menini, Marco Polignano, Valentina Russo, Rachele

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Riccerdo Orlando, Luca Moroni, Pere-Lluis Huguet Cabot, Simone Conia, Edoardo Barba, and Roberto Navigli. 2024. Minerva. <https://nlp.uniroma1.it/minerva/>.

Marco Polignano, Pierpaolo Basile, and Giovanni Semeraro. 2024. Advanced natural-based interaction for the italian language: Llamantino-3-anita. *Preprint*, arXiv:2405.07101.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

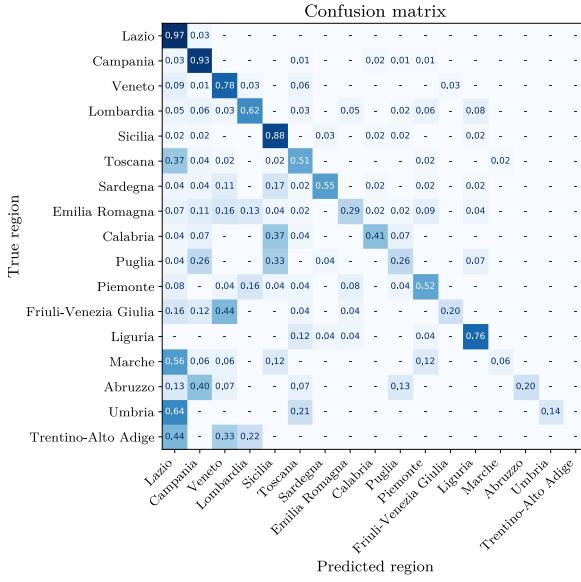
Alan Ramponi and Camilla Casula. 2023. Geolingit at evalita 2023: Overview of the geolocation of linguistic variation in italy task. In *International Workshop on Evaluation of Natural Language and Speech Tools for Italian*.

Andrea Santilli and Emanuele Rodolà. 2023. Camoscio: an italian instruction-tuned llama. *Preprint*, arXiv:2307.16456.

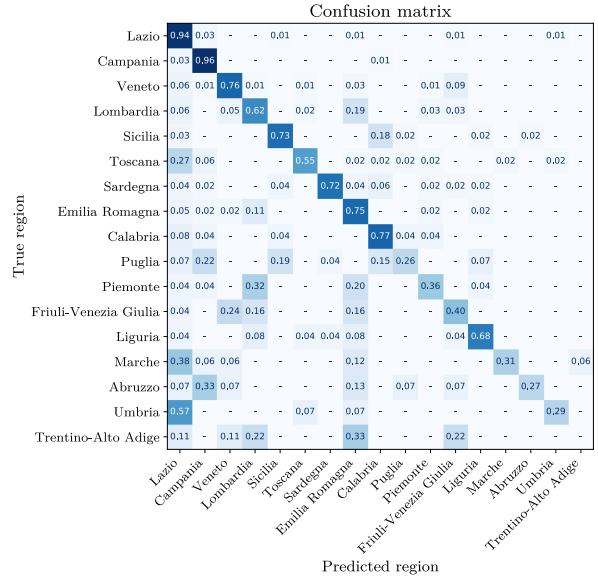
Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

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

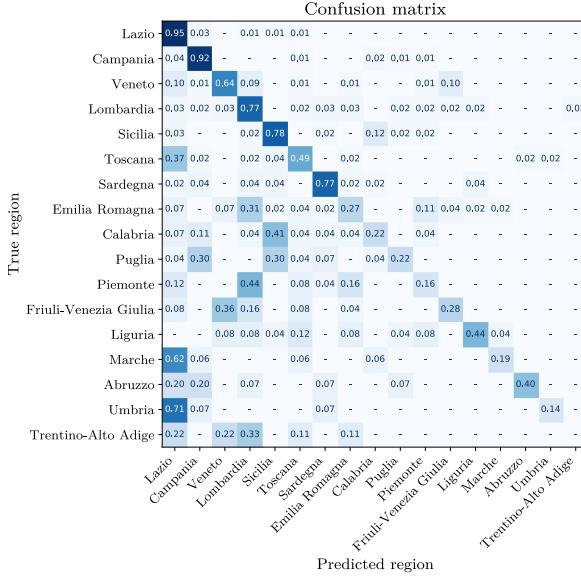
Appendix



(a) Confusion matrix for the classification with Camoscio.

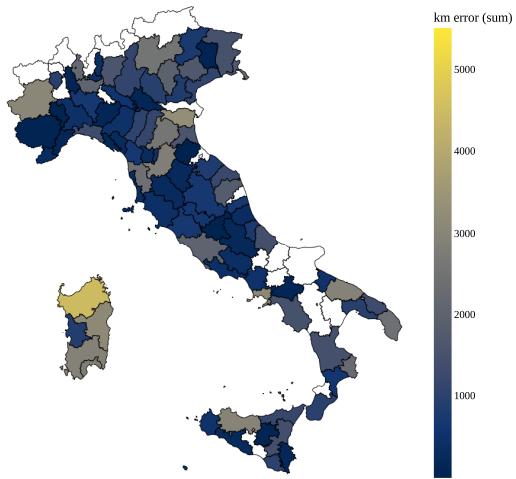


(b) Confusion matrix for the classification with ANITA.

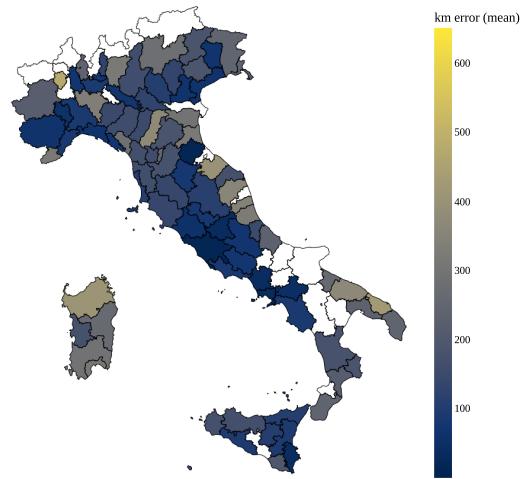


(c) Confusion matrix for the classification with Minerva.

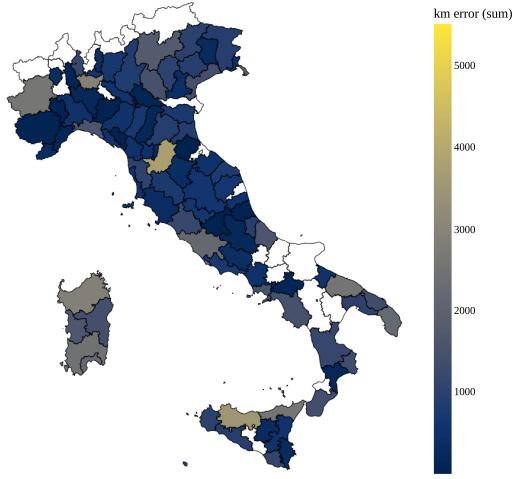
Figure 3: Confusion matrices for the classification of the samples in the test set for all tested models: Camoscio (3a), ANITA (3b) and Minerva (3c). The classes on the x and y axis include only the classes present in the test set, which are a subset of Italian regions. The numbers in each cell (c_{pred}, c_{true}) correspond to the frequency of samples with class c_{true} classified as c_{pred} and normalized by the total number of samples of the true class (row). Cells containing "-" mean zero frequency of classified samples. Darker colors highlight higher frequencies, and a darker main diagonal on the matrix implies strong classification performance.



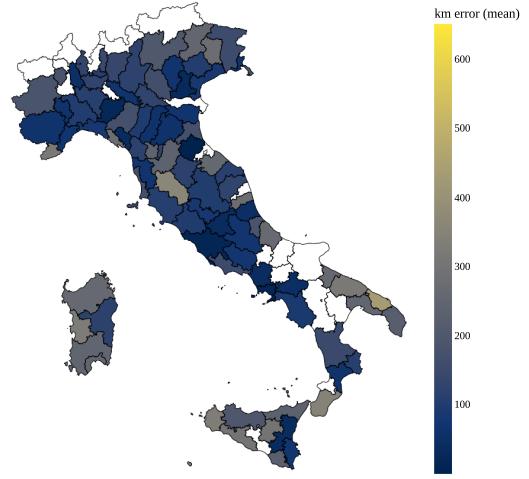
(a) Heatmap of the **sum** of the regression error (in km) over Italian provinces for Camoscio.



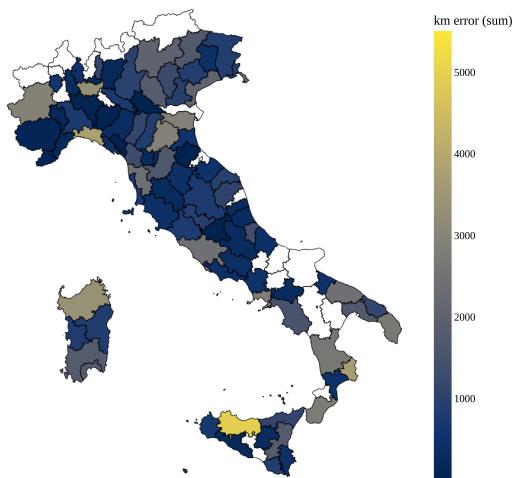
(b) Heatmap of the **mean** of the regression error (in km) over Italian provinces for Camoscio.



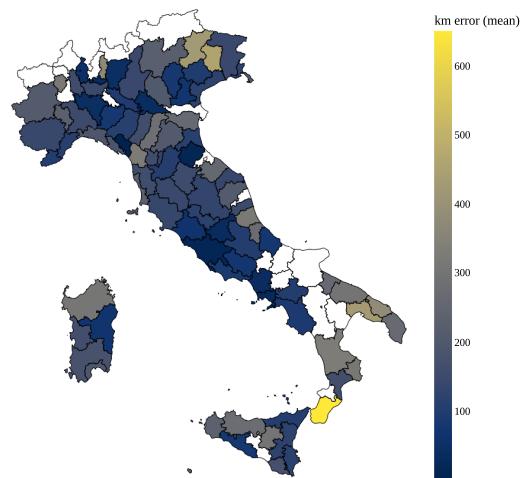
(c) Heatmap of the **sum** of the regression error (in km) over Italian provinces for ANITA.



(d) Heatmap of the **mean** of the regression error (in km) over Italian provinces for ANITA.



(e) Heatmap of the **sum** of the regression error (in km) over Italian provinces for Minerva.



(f) Heatmap of the **mean** of the regression error (in km) over Italian provinces for Minerva.

Figure 4: Heatmaps of the the regression error (in km) over Italian provinces for all the tested models. The figures in the left column (4a, 4c and 4e) show the sum of distance error over the area of a province. Instead, figures in the right column (4b, 4d and 4f) show the average distance error over the same areas.