DANTE at GeoLingIt: Dialect-Aware Multi-Granularity Pre-training for Locating Tweets within Italy

Giuseppe Gallipoli[1,†], Moreno La Quatra[2,†], Daniele Rege Cambrin[1,†], Salvatore Greco[1,†] and Luca Cagliero[1]

[1]*Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Turin TO*
[2]*Kore University of Enna, Piazza dell'Università, 94100 Enna EN*

Abstract

This paper presents an NLP research system designed to geolocate tweets within Italy, a country renowned for its diverse linguistic landscape. Our methodology consists of a two-step process involving pre-training and fine-tuning phases. In the pre-training step, we take a semi-supervised approach and introduce two additional tasks. The primary objective of these tasks is to provide the language model with comprehensive knowledge of language varieties, focusing on both the sentence and token levels. Subsequently, during the fine-tuning phase, the model is adapted explicitly for two subtasks: coarse- and fine-grained variety geolocation. To evaluate the effectiveness of our methodology, we participate in the GeoLingIt 2023 shared task and assess our model's performance using standard metrics. Ablation studies demonstrate the crucial role of the pre-training step in enhancing the model's performance on both tasks.

Keywords

Linguistic Varieties, Region Localization, Text Classification and Regression, Italian NLP

1. Introduction

Italy is widely recognized for its linguistic diversity, with 20 distinct regions, each characterized by various unique and shared dialects [1]. These dialects exhibit further variations within each region, often associated with specific cities or provinces, and sometimes extend beyond regional boundaries. The intricate nature of Italy's linguistic landscape poses a significant challenge in accurately identifying the origin of a given text within the country.

This research is conducted in the context of the GeoLingIt shared task [2] at EVALITA 2023 [3]. It focuses on the geolocation of social media data, specifically Twitter posts. The task comprises two subtasks: *Coarse-grained variety geolocation* (Subtask A), whose aim is to determine the region from which a tweet originates within the 20 Italian regions, and *Fine-grained variety geolocation* (Subtask B), which focuses on predicting the latitude and longitude coordinates corresponding to the origin of a tweet within Italy. Linguistic variations within and across regions make it difficult to accurately associate a piece of text with its specific geographic origin. The challenge becomes even more significant due to the similarities each language variety may share with other languages, even outside Italy.

This paper presents the DANTE (Dialect ANalysis TEam)[1] submission for the GeoLingIt 2023 shared task, characterized by a two-step methodology involving pre-training and fine-tuning phases. By leveraging Italian or multilingual models, we propose a semi-supervised pre-training approach that combines standard and novel pre-training tasks to capture regional dialect information at multiple levels of granularity (i.e., sentence and token levels). Following the pre-training phase, the model undergoes a standard fine-tuning process tailored to the two subtasks proposed by the shared task. Through extensive experiments, we demonstrate the effectiveness of our methodology.

2. Background

Text classification is a fundamental task in NLP whose objective is to assign one (or more) predefined classes to a piece of text. It has many applications ranging from sentiment analysis to topic classification. In this work, we apply it to the prediction of the geographic region associated with the linguistic variety expressed in a tweet.

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Sep 7 – 8, Parma, IT

[†] Corresponding authors. These authors contributed equally.
giuseppe.gallipoli@polito.it (G. Gallipoli); moreno.laquatra@unikore.it (M. L. Quatra); daniele.regecambrin@polito.it (D. R. Cambrin); salvatore_greco@polito.it (S. Greco); luca.cagliero@polito.it (L. Cagliero)
https://mlaquatra.me (M. L. Quatra); https://darthreca.github.io/ (D. R. Cambrin); https://grecosalvatore.github.io/ (S. Greco)
0009-0003-1744-6674 (G. Gallipoli); 0000-0001-8838-064X (M. L. Quatra); 0000-0002-5067-2118 (D. R. Cambrin); 0000-0001-7239-9602 (S. Greco); 0000-0002-7185-5247 (L. Cagliero)

[1] The name "DANTE" is inspired by the Italian poet Dante Alighieri, widely regarded as one of the founding fathers of the Italian language.

The introduction of the Transformer [4] architecture for machine translation has represented a significant breakthrough in NLP, achieving superior performance also in other tasks, including text classification. Transformer-based classification models implement an encoder-only architecture whose objective is to extract a continuous representation from the input text. To do this, models are generally pre-trained on large corpora of unlabeled text using specific pre-training objectives.

The pre-training stage allows the model to learn language representations that enable it to capture the structure and semantics of the text more effectively. Our work follows the same approach by further pre-training several Transformer-based models as discussed in Section 3. After pre-training, the model is fine-tuned on labeled data tailored to the desired task. Specifically, the architecture is enriched by additional classification layers (i.e., classification head) trained in a supervised fashion to output the final probability for each class. Similarly, by introducing one or multiple linear layers, (multi-)regression tasks can also be performed.

Some of the most widely adopted Transformer-based classification models include: BERT and its multilingual version mBERT [5], DistilBERT [6], which is a distilled version of BERT, RoBERTa [7], and its multilingual version XLM [8], which are two variations of BERT including dynamic masking.

Computational linguistics research in Italian faces challenges due to the scarcity of large-scale datasets specifically designed for the language, as highlighted recently [9]. Also, the computational effort required to pre-train language models has resulted in only a few available architectures in Italian. Specifically, some of them are BERT-Italian and ELECTRA-Italian [10]. Furthermore, although they are not encoder-only architectures, the following are some of the other models available in Italian: GePpeTto [11], which is based on GPT-2 [12], IT5 [13], which is the Italian version of T5 [14], and the recently released BART-IT [15], which is the Italian version of BART [16].

3. Description of the system

The DANTE methodology for the GeoLingIt shared task aims to both identify the region of origin and predict the geographic coordinates of tweets within Italy.

3.1. Pre-training

The initial phase of our methodology involves pre-training the model to improve its ability to analyze different linguistic varieties. We initialize a Transformer-based encoder model using Italian or multilingual pre-trained weights and further pre-train it using both standard and novel pre-training tasks.

Masked Language Modeling (MLM) & Next Sentence Prediction (NSP). The MLM and NSP tasks are standard pre-training tasks used to train Transformer-based models. Both tasks contribute to language processing by helping the model learn the contextual information of words and their relationships.

Region Classification (RC). By leveraging region-specific data, we integrate into pre-training the supervised task of predicting the geographic region associated with the linguistic variety expressed in a given sentence.

Token-level Region Classification (TRC). We also include an additional (supervised) token classification task. It aims at predicting the geographic region associated with each token in a given sentence. To create training examples, we randomly combine multiple sentences belonging to text snippets labeled with different regions. This task aims at enabling the model to capture regional linguistic information with higher granularity.

Using a multi-task learning approach, the model is trained on multiple tasks simultaneously, allowing it to learn a shared representation useful for all tasks. We define a separate linear layer for each task (i.e., task-specific head) that operates on the shared representation and is trained using the corresponding labeled data. We experimented with two different multi-task learning setups: (1) *task-specific training (TST)*, where the model is trained on a single task at a time, with each batch randomly selecting one task from the set of all available tasks, and (2) *joint training (JT)*, where the model is trained on all tasks simultaneously, and the loss is computed as the average of the losses of all tasks. These two multi-task learning setups were inspired by recent findings in the literature [17].

3.2. Subtask A: Coarse-grained variety geolocation

The *Subtask A* within GeoLingIt 2023 shared task involves identifying the region of origin of a given tweet within Italy. It can be formulated as a classification task, where the model is trained to classify each tweet into its corresponding geographic region (i.e., one of the 20 Italian regions). To this end, we follow a standard fine-tuning approach, where the pre-trained model is adapted for the downstream task using the labeled training data. The representation of a special `[CLS]` token is used as the input to a linear layer trained to predict the region of origin of the tweet. The model is trained to minimize

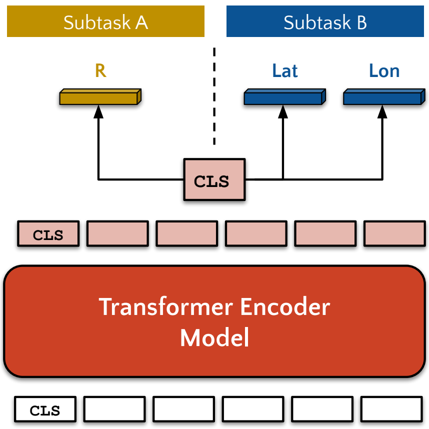

Figure 1: Fine-tuning architecture. It includes two branches: one for *Subtask A* which predicts the region class (represented as “R”), and another for *Subtask B* which predicts the latitude (represented as “Lat”) and longitude (represented as “Lon”).

the cross-entropy loss between the predicted and the ground-truth labels. For a visual representation of the fine-tuning process, please refer to the top-left part of Figure 1.

3.3. Subtask B: Fine-grained variety geolocation

The *Subtask B* aims to localize a given tweet’s origin within Italy by predicting its latitude and longitude coordinates. The task can be formulated as a regression problem, where the model is trained to jointly predict two values (i.e., the latitude and longitude coordinates of the tweet’s origin) using two separate linear layers. Similar to the previous case, the tweet representation is obtained by feeding the `[CLS]` token representation to both linear layers. The fine-tuning architecture for this subtask is illustrated in the upper right corner of Figure 1, showcasing the linear layers positioned on top of the `[CLS]` token.

The overall loss function for the regression task is defined as:

$$\mathcal{L}_{localization} = \frac{1}{2}\left(\mathcal{L}_{lat} + \mathcal{L}_{lon}\right)$$

Where $\mathcal{L}_{lat}$ and $\mathcal{L}_{lon}$ represent the mean squared error (MSE) loss for the latitude and longitude predictions, respectively.

It is worth noting that the model is separately fine-tuned for each task (i.e., coarse- and fine-grained variety geolocation). Thus, we do not use multi-task learning at this stage. Jointly optimizing the two tasks during fine-tuning could help the model to ensure consistency between the two tasks and improve the model’s performance. This is one of the possible future directions we plan to explore.

4. Dataset

4.1. Pre-training Dataset

To the best of our knowledge, there are no existing large-scale data collections specifically focusing on Italian language varieties. Therefore, we exploited web scraping to construct our pre-training dataset. From a web search, we identified the following two sources: (1) Dialettando[2]: a website that contains several proverbs, sayings, poems, rhymes, and stories from different regions; (2) Wikipedia: which comprises specific versions for some regional languages (e.g., nap[3] for Neapolitan). They were both accessed in January 2023. For the data collected from Wikipedia, we associated each language-specific Wikipedia portal with the region primarily representing the respective language. For example, data collected from the nap Wikipedia portal would be associated with the Campania region, which predominantly represents the Neapolitan language. After the data collection, we ended up with a corpus of 273,011 documents containing linguistic varieties of Italian from different regions. Out of these, 12,692 documents were collected from Dialettando, while the majority, 260,319, were obtained from Wikipedia. From Dialettando, we also collected the 12,692 Italian translations of the same documents in the corpus. This was done because the DiatopIt corpus utilized in the task contains instances that encompass regional Italian variations. Therefore, the final pre-training dataset is composed of 285,703 documents. Notice that a document can be a Wikipedia article or any text from Dialettando (e.g., proverb, saying, or story) without any difference, even if they can have different lengths. Indeed, we found that the mean number of tokens is 48 for Dialettando and 147 for Wikipedia[4]. However, both sources of texts can be helpful during the pre-training phase.

Figure 2 shows the distribution of the collected documents for each Italian region. Table 1 details the number of documents for each region and data source. As can be noticed, regions of the north of Italy, such as Piemonte, Lombardia, and Veneto, are predominant in the dataset, with approximately 60k texts (corresponding to approximately 20% of the entire collection) each. They are followed by some regions of the south, such as Sicilia, Campania, and Puglia, with around 25k, 14k, and 11k texts, respectively. Finally, regions such as Valle D’Aosta, Toscana, Umbria, Marche, Lazio, Molise, Abruzzo, and

[2] https://www.dialettando.com
[3] https://nap.wikipedia.org/
[4] Computed with the `bert-base-multilingual-cased` tokenizer of the HuggingFace library https://huggingface.co/bert-base-multilingual-cased.

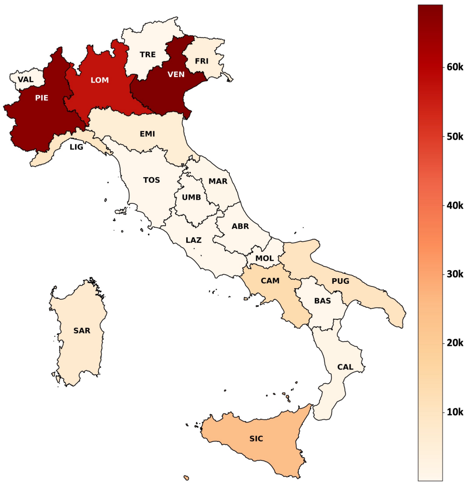

Figure 2: Distribution of the collected pre-training texts per region.

Region	Abbr.	# Documents	
		Dial.	Wiki
Abruzzo	ABR	325	–
Basilicata	BAS	260	–
Calabria	CAL	1,258	–
Campania	CAM	1,340	12,709
Emilia Romagna	EMI	492	5,194
Friuli Venezia Giulia	FRI	157	3,833
Lazio	LAZ	356	–
Liguria	LIG	1,667	7,838
Lombardia	LOM	1,332	55,743
Marche	MAR	690	–
Molise	MOL	153	–
Piemonte	PIE	854	66,407
Puglia	PUG	1,504	9,063
Sardegna	SAR	296	7,262
Sicilia	SIC	843	23,732
Toscana	TOS	242	–
Trentino Alto Adige	TRE	127	–
Umbria	UMB	161	–
Valle D'Aosta	VAL	102	–
Veneto	VEN	533	68,538
Total		*12,692*	*260,319*

Table 1
Number of collected documents for each Italian region. For each Italian region is reported: i) the abbreviation code; ii) the number of documents collected from Dialettando and Wikipedia. The total number of documents containing Italian language varieties from all regions is is 273,011.

Basilicata are lowly represented, with less than 1k texts each.

Notably, the data utilized for pre-training is sourced from the Dialettando website and Wikipedia, adhering to the guidelines of the GeoLingIt shared task by excluding social network data to avoid overlap with the fine-tuning data.

4.2. GeoLingIt Dataset

The GeoLingIt 2023 dataset for the *Subtasks A* and *B* is DiatopIt [18], a corpus of diatopic variations of language in Italy. It is composed of geotagged social media posts from Twitter. Each tweet also comprises the associated latitude, longitude, and the Italian region of origin. The dataset contains 13,669 examples for training, 552 for validation, and 818 for testing. This dataset is exploited in the fine-tuning phase to specialize the models to coarse- and fine-grained variety geolocation. The authors of the competition have already anonymized data. Specifically, user mentions, email addresses, and URLs have been replaced with specific placeholders for privacy reasons. However, the content of tweets is unfiltered and can exhibit non-standard language use (e.g., insults, bad words).

5. Results

Baseline models We compare our models with baseline models pre-trained on Italian or multilingual data, which undergo the same fine-tuning process described in Sections 3.2 and 3.3. The following baseline models are considered: (1) multilingual BERT model (mBERT)[5], (2) BERT model pre-trained on 13GB of Italian text (BERT-IT)[6], (3) BERT model pre-trained on 81GB of Italian text (BERT-IT-XXL)[7], (4) XLM-RoBERTa (XLM-R)[8], and (5) multilingual DistilBERT (dBERT)[9]. All models are used in their cased versions. By comparing the results of these baseline models with our approach, we can assess the benefits of the proposed pre-training phase.

Experimental settings The pre-training phase lasts five epochs and utilizes the Adam optimizer with a linear learning rate scheduler. The scheduler includes a warmup period (10% of the total training steps) followed by a linear decay of the learning rate until the end of training. The fine-tuning phase lasts for ten epochs and utilizes the same settings for the optimizer and scheduler as the pre-training phase.

[5] https://huggingface.co/bert-base-multilingual-cased
[6] https://huggingface.co/dbmdz/bert-base-italian-cased
[7] https://huggingface.co/dbmdz/bert-base-italian-xxl-cased
[8] https://huggingface.co/xlm-roberta-base
[9] https://huggingface.co/distilbert-base-multilingual-cased

5.1. Coarse-grained variety geolocation

We report the single models and an ensemble, which includes all models evaluated on the development set. The ensemble prediction is obtained through majority voting on the individual models' predictions. In case of a tie, a random selection is employed. The organizers provide Logistic Regression (LogReg) and Most Frequent Baseline (MFB) as baselines. According to the competition rules, we consider macro F1-score, precision, and recall as evaluation metrics.

In Table 2, we reported development and test sets results. There is a noticeable performance improvement when comparing deep learning techniques to classical machine learning methods. One common aspect shared by both approaches is the observed degradation in performance on the test set compared to the development set. This pattern can be attributed to the fact that the test set contains additional regions that are not present in the development set. Joint-Training (JT) consistently yields the best results in terms of F1-score, achieving significant improvements ranging from +2% to +7% compared to the absence of pre-training. This boost primarily manifests as enhanced precision.

Following GeoLingIt guidelines, we can only submit three models for test set evaluation. We have selected the top-3 models based on their performance on the development set: Jointly-Trained BERT-IT-XXL, Task-Specific-Trained BERT-IT-XXL, and the models' Ensemble. We show in Table 2b the performance of these models on the test set. The results show that the Ensemble method achieved the highest performance, followed by the Task-Specific-Trained (TST) BERT-IT-XXL model and the Jointly-Trained (JT) BERT-IT-XXL model. Surprisingly, the TST pre-training approach outperformed the others, exhibiting a significant +2% improvement in F1-score compared to the corresponding model pre-trained using JT.

Model	PT	Precision	Recall	F1-score
LogReg	✗	0.7686	0.5389	0.5872
MFB	✗	0.0160	0.0769	0.0265
	✗	0.7498	0.7914	0.7260
BERT-IT-XXL	JT	0.8401	0.7638	0.7923
	TST	0.8101	0.7374	0.7621
	✗	0.7276	0.7795	0.7005
BERT-IT	JT	0.7757	**0.8161**	0.7496
	TST	0.7591	0.8075	0.7310
	✗	0.7032	0.7555	0.6747
dBERT	JT	0.7196	0.7651	0.6962
	TST	0.7202	0.7830	0.6941
	✗	0.7096	0.7476	0.6874
mBERT	JT	0.7358	0.7872	0.7222
	TST	0.7408	0.7999	0.7170
	✗	0.6856	0.7257	0.6723
XLM-R	JT	0.7642	0.8112	0.7438
	TST	0.7325	0.7709	0.7136
Ensemble	E	**0.8609**	0.7878	**0.8121**

(a) Development Set

Model	PT	Precision	Recall	F1-score
LogReg	✗	0.6219	0.4243	0.4611
MFB	✗	0.0447	0.2115	0.0738
BERT-IT-XXL	JT	0.6518	0.6009	0.6172
BERT-IT-XXL	TST	0.6698	0.6265	0.6393
Ensemble	E	**0.7946**	**0.6375**	**0.6630**

(b) Test Set

Table 2
Subtask A results. The *PT* column indicates no-pre-training with ✗, Task-Specific Training with *TST*, Joint Training with *JT*, and models ensemble with *E*.

5.2. Fine-grained variety geolocation

Task organizers provide K-nearest-neighbor (KNN) and centroid baseline (CB) models as baselines. Following shared task guidelines, the models' performance is assessed using haversine distance. In this case, the lower, the better. We report the results of single models and an ensemble of the top-2 evaluated models. The ensemble prediction is obtained using the mean point between the two individual models' predictions.

In Table 3, we reported results for the evaluated model on the development and test sets. Similar to Subtask A, deep learning models outperform classical approaches. Notably, the test set's performance in this task shows higher scores than the development set.

The results on the development set confirm the effectiveness of the pre-training with one exception: Task-Specific-Training on BERT-IT. The differences span from -19 km to -161 km with respect to the models without a specific pre-training. In most cases, Joint-Training consistently yields the best results, except for XLM-R.

We submit the top-3 most promising solutions according to the development set results for evaluation on the test set: Jointly-Trained dBERT, Task-Specific-Trained dBERT, and Ensemble. We report their performance on the test set in Table 3b. The Ensemble model achieved the best performance, with the TST pre-training demonstrating a 1.5 km average distance improvement compared to the JT counterpart.

Model	PT	Avg Dist (km)
KNN	✗	281.03
CB	✗	301.65
BERT-IT-XXL	✗	161.88
	JT	128.18
	TST	136.00
BERT-IT	✗	187.47
	JT	153.26
	TST	215.35
dBERT	✗	143.61
	JT	118.57
	TST	124.15
mBERT	✗	192.20
	JT	135.04
	TST	164.99
XLM-R	✗	316.90
	JT	156.86
	TST	155.83
Ensemble	E	**117.64**

(a) Development Set

Model	PT	Avg dist (km)
KNN	✗	263.35
CB	✗	281.04
dBERT	JT	114.00
dBERT	TST	112.58
Ensemble	E	**110.35**

(b) Test Set

Table 3
Subtask B results. The *PT* column indicates no-pre-training with ✗, Task-Specific Training with *TST*, Joint Training with *JT*, and models ensemble with *E*.

6. Discussion

Our analysis assessed the effectiveness of widely used deep language models in the context of both coarse- and fine-grained variety geolocation tasks. We also offer an interactive demo[10] to showcase our best-performing models and release both the code and pre-trained models[11]. It is worth noting that social media data composing the fine-tuning dataset may contain profanities, slurs, hateful content, and stereotypes. Although pre-training data is collected using controlled sources, a similar statement may apply to them. A community partially manages both the Dialettando website and Wikipedia portals. Therefore their content may not be carefully curated. As a result, the models may exhibit label correlations based on the presence of such offensive language, potentially influencing their region identification capabilities. The proposed methodologies are not intended to offend anyone; since they may be inaccurate in some cases, it is possible to get improprieties.

7. Conclusion and Future Works

This paper presents an effective solution for modeling language varieties within Italy, achieving excellent results and ranking 1st and 2nd among other teams for Subtask A and Subtask B, respectively. However, there are still promising avenues for future research. Utilizing multi-task learning during the fine-tuning phase can improve consistency and performance by training on multiple related tasks using the same backbone model. Regarding model architecture, we aim to investigate the development of a specific model focused on identifying portions of the text belonging to specific language varieties. This model will be designed to identify the distinctive linguistic features within tweets accurately. By successfully identifying these features, the model would have the potential to concentrate on the relevant parts of the text, which may lead to improved localization capabilities. Finally, preliminary experiments show that incorporating curriculum learning techniques during pre-training can optimize the learning process and enhance the overall model's performance.

Acknowledgments

This study was partially carried out within the MICS (Made in Italy – Circular and Sustainable) Extended Partnership and received funding from Next-GenerationEU (Italian PNRR – M4 C2, Invest 1.3 – D.D. 1551.11-10-2022, PE00000004). This study was also partially carried out within the FAIR (Future Artificial Intelligence Research) and received funding from Next-GenerationEU (Italian PNRR – M4 C2, Invest 1.3 – D.D. 1555.11-10-2022, PE00000013). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

References

[1] A. Ramponi, Nlp for language varieties of italy: Challenges and the path forward, arXiv preprint arXiv:2209.09757 (2022).

[2] A. Ramponi, C. Casula, GeoLingIt at EVALITA 2023: Overview of the geolocation of linguistic variation in Italy task, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and

[10] https://huggingface.co/spaces/DGMS/DANTE-GeoLingIT2023

[11] https://github.com/MorenoLaQuatra/DANTE-GeoLingIT2023

Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[3] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, 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, 2023.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[5] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[6] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, ArXiv abs/1910.01108 (2019).

[7] 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). URL: http://arxiv.org/abs/1907.11692.

[8] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 8440–8451. URL: https://aclanthology.org/2020.acl-main.747. doi:10.18653/v1/2020.acl-main.747.

[9] A. Koudounas, M. La Quatra, L. Vaiani, L. Colomba, G. Attanasio, E. Pastor, L. Cagliero, E. Baralis, Italic: An italian intent classification dataset, arXiv preprint arXiv:2306.08502 (2023).

[10] S. Schweter, Italian bert and electra models, 2020. URL: https://doi.org/10.5281/zenodo.4263142. doi:10.5281/zenodo.4263142.

[11] L. D. Mattei, M. Cafagna, F. Dell'Orletta, M. Nissim, M. Guerini, Geppetto carves italian into a language model, 2020. arXiv:2004.14253.

[12] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners, 2019.

[13] G. Sarti, M. Nissim, IT5: Large-scale text-to-text pretraining for italian language understanding and generation, ArXiv preprint 2203.03759 (2022). URL: https://arxiv.org/abs/2203.03759.

[14] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, J. Mach. Learn. Res. 21 (2020).

[15] M. La Quatra, L. Cagliero, BART-IT: An Efficient Sequence-to-Sequence Model for Italian Text Summarization, Future Internet 15 (2022) 15.

[16] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 7871–7880. URL: https://aclanthology.org/2020.acl-main.703. doi:10.18653/v1/2020.acl-main.703.

[17] C. Fifty, E. Amid, Z. Zhao, T. Yu, R. Anil, C. Finn, Efficiently identifying task groupings for multi-task learning, Advances in Neural Information Processing Systems 34 (2021) 27503–27516.

[18] A. Ramponi, C. Casula, DiatopIt: A corpus of social media posts for the study of diatopic language variation in Italy, in: Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023), 2023, pp. 187–199. URL: https://aclanthology.org/2023.vardial-1.19.