# Galliz at GeoLingIt: Enhancing BERT with Vocabulary Knowledge for Predicting the Region of Language Varieties of Italy

Tiziano Labruna[1,2], Simone Gallo[3,4]

[1]*Fondazione Bruno Kessler, Trento, Italy*
[2]*Free University of Bozen-Bolzano, Italy*
[3]*HIIS Laboratory, CNR - ISTI, Pisa, Italy*
[4]*University of Pisa, Italy*

Abstract

The linguistic diversity of the Italian peninsula and its islands, characterized by several language varieties, represents a linguistic condition and a cultural treasure unique in Europe. However, the oral nature of these varieties poses a challenge to their preservation in the written form. While significant research efforts have been dedicated to standard Italian language processing, less attention has been given to the language varieties of Italy and the development of supporting resources. This paper aims to study the peculiarities of language varieties of Italy and identify the region of origin of tweets written in *non-[Standard Italian]* varieties. To achieve this goal, we utilized two main techniques: fine-tuning a language model (BERT) and implementing an algorithm that utilizes dictionaries of regional varieties and word frequency. Our results show that integrating lexical analysis with BERT could be a promising approach for this particular task. We present an overview of the data, methodology, and evaluation results, then discuss the implications of our findings.

**Keywords**

Natural Language Processing, Language varieties, Tweets classification

## 1. Introduction and Motivations

The Italian peninsula and its islands present considerable linguistic variation among the different regions that compose it, as well as within the regions themselves. The presence of many different language varieties makes this linguistic situation special and unique in Europe [1], as well as a treasure of cultural diversity, interpretation, and expression of the reality to which they belong. However, these linguistic diversities are in danger of being lost, as most of them are passed on only orally, leaving less room for written usage [1]. Despite significant research efforts being devoted to processing techniques for standard Italian (e.g., [2, 3, 4]), less effort has been devoted to supporting language varieties, both from a technological point of view and in terms of curated resources [5]. In this paper, our main goal is to study varieties of Italy in order to develop effective methods for classifying the region of origin of Twitter posts (tweets) written from Italy. We address two different tasks of GeoLingIt 2023 [6] from EVALITA 2023 [7]: one for classifying non-standard Italian tweets according to their region of origin at the country level (*"standard track"*), and another for classifying tweets according to a subset of regions (*"special track"*, in this case, Lazio and Toscana). To tackle this problem, we rely on the combination of two different techniques: the first one is based on the fine-tuning of a language model (i.e., BERT), while the second is based on an algorithm that utilizes regional varieties dictionaries and the frequency of words present in the tweets. The classification results obtained from both techniques are normalized and combined to derive the final result. In the following sections, we provide an overview of the data and resources used for the tasks. We then describe the methodology applied, including data augmentation, prediction using the two different techniques, and global prediction. Next, we present the results obtained during the evaluation phase. Finally, we discuss the findings and draw some conclusions.

## 2. Data and Resources

The dataset used for the tasks was collected by retrieving geotagged tweets classified as *IT* by Twitter. The curators only kept posts that exhibit *non-standard* language, along with the region information that falls within the Italy territory [6].

### 2.1. Provided Data

The provided data consists of training and development splits for both the *standard track* and the *special track*

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

tlabruna@fbk.eu (T. Labruna); simone.gallo@isti.cnr.it (S. Gallo)
0000-0001-7713-7679 (T. Labruna); 0000-0002-5162-0475 (S. Gallo)

tasks. The data is provided in a tab-separated format, with each column defining three properties:

- "id": an integer that uniquely identifies the tweet;
- "text": a string representing the text of the tweet in a *non-standard* language variety of Italy. This variety may be present as a single word or phrase (there are many cases of code-switching), or the entire tweet can be written in that variety. Any sensitive information has been replaced with placeholders by the curators (e.g., "@tagged_user" is replaced with "[USER]");
- "region": a string representing the tweets' region of provenance (e.g., Lazio, Sicilia, Toscana).

The training set contains 13,669 tweets, covering all the administrative regions of Italy. The development set consists of 552 tweets from 13 selected regions, namely Calabria, Campania, Friuli-Venezia Giulia, Emilia Romagna, Lazio, Liguria, Lombardia, Piemonte, Puglia, Sardegna, Sicilia, Toscana, and Veneto. The training set exhibits a strong imbalance, with highly represented regions like Lazio (5549 items) and Campania (2971 items), while regions such as Valle d'Aosta and Molise have only 14 and 35 items, respectively. The overall class distribution is shown in Fig. 1.

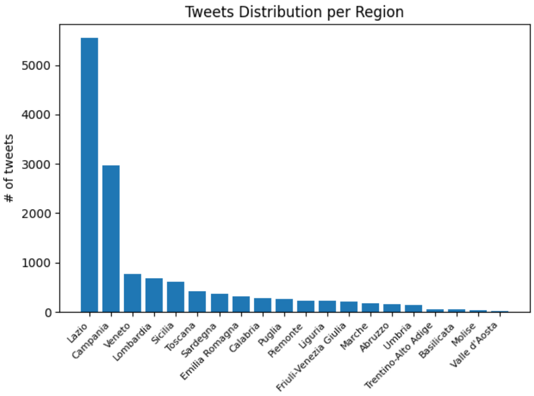

**Figure 1:** Distribution of tweets for each region in the training set.

## 2.2. Language Model

We employed BERT (Bidirectional Encoder Representations from Transformers) [8] as our language model, which was introduced by Google in 2018 and has gained significant prominence in various Natural Language Processing (NLP) tasks. Unlike traditional language models that utilize left-to-right or right-to-left approaches, BERT utilizes bidirectional pre-training, allowing all tokens in the input to contribute to the prediction process. Its effectiveness is evident from its state-of-the-art performance on multiple NLP benchmarks, including GLUE, SQuAD and RACE.
Although BERT has been in existence for five years, we believe that it remains highly suitable for our specific task. In light of its robust performance across diverse NLP applications, leveraging BERT finely aligns with the requirements of our work.
During the model selection, we also took into account Italian variants of the classic BERT model (e.g., "bert-base-italian"[1]), but the preliminary results reported, in some case, worst performance compared to the "standard" English BERT (i.e., "bert-base-uncased"[2])

## 2.3. Vocabularies

To build the vocabularies needed for our purpose, we utilized various online resources as well as the text from the provided tweets. Some of these vocabularies were used for the *standard track*, while other ones for the *special track*, as more precisely described in the rest of this Section. All vocabularies are publicly available[3].

**Global vocabulary** We obtained a "global" vocabulary, containing words from language varieties spoken in every Italian region, by performing web scraping on the dictionary available at "Dialettando.com"[4]. This resulted in a JSON file containing all the available words for each region. This vocabulary was used for the *standard track* only.

**Unique words vocabulary** This vocabulary was generated starting from the provided training set and considering the occurrences of the words for every tweet from each region. Specifically, it contains all the unique words present in the tweets, along with their corresponding frequencies, grouped by region. This vocabulary was used for the *standard track*, and a subset of this vocabulary with only the regions Toscana and Lazio was used for the *special track*.

**Distinctive words vocabulary** Similarly to the previous one, this vocabulary was generated from the provided training set. This time, we keep only the distinctive words from each region's tweets (i.e., only words that are exclusive to a specific region and do not appear in other regions are considered), along with the corresponding frequencies. This vocabulary was used for the *standard*

[1] https://huggingface.co/dbmdz/bert-base-italian-cased
[2] https://huggingface.co/bert-base-uncased
[3] https://github.com/simog-dev/Galliz_geolingit
[4] https://www.dialettando.com/dizionario/dizionario.lasso, *retrieved on May 9th, 2023*

*track*, and a subset of this vocabulary with only the regions Toscana and Lazio was used for the *special track*.

**Toscana vocabulary** We obtained this vocabulary by performing a web scraping of the terms present in the website of "Vocabolario del Fiorentino Contemporaneo"[5] and thus converting the content of the website into a JSON file. This vocabulary was used for the *special track* only.

**Lazio vocabulary** We obtained this vocabulary by performing a web scraping of the terms present in the website of "The Roman Post" website[6] and thus converting the content of the website into a JSON file. This vocabulary was used for the *special track* only.

## 3. Methodology

In order to predict the most likely region of origin of a given sentence, we decided to make use of both LLMs (fine-tuned on an augmented training set) and lexical information of regional varieties, taken from the vocabularies presented above. As shown in Fig. 2, we first consider the predictions of the two strategies individually, and then we merge both contributions for coming to one final prediction of the region enclosing the particular variety for the sentence.

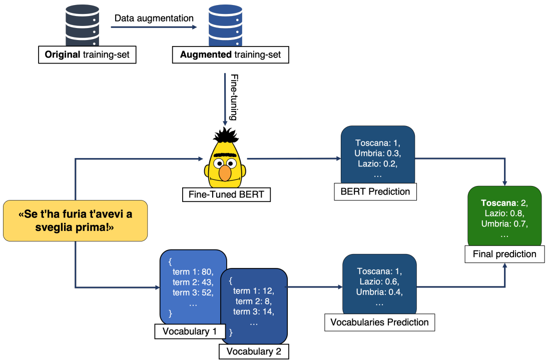

**Figure 2:** System Architecture

### 3.1. Data Augmentation

The first step was to increase the size of the training set, both for producing better training of the classifier, and for equalizing the distribution of the samples for each region, which was initially extremely unbalanced (cfr. § 2.1).

Our approach for implementing data augmentation was to generate new sentences that are equal to the original one, but with one random word that is substituted by a different word, semantically similar to the original one. We utilized established approaches from literature [9, 10] to implement word substitution in the text, by changing the value of the portion of the text and maintaining all the rest unchanged. As an example, the original sentence "Fa ancora na sfaccim e per andare in #moto sulle mie montagne" was transformed into "Fa ancora na sfaccim e per tornare in #moto sulle mie montagne", since the verbs "andare" and "tornare" are semantically similar and do not change the global meaning of the sentence.

In order to find similar words, we used a Word Embedding model for Italian[7] (due to the similarity between Italian and the majority of linguistic varieties spoken in Italy) fine-tuned on our training set, and then we selected one among the vectors that are closer to the vector of the word we want to substitute, using the library "Word2Vec.most_similar"[8].

The augmented dataset ensures an equal distribution of sentences (5549 sentences each) for every region. This quantity corresponds to the initial number of sentences for Lazio, the region with the highest number of entries. In each region, except for Lazio, the number of newly generated sentences equalled the difference between the initial number of sentences in that region and the initial number of sentences in Lazio. Therefore, if a region had a lower initial sentence count, the same sentence was used more frequently for augmentation (thus the number of times a single sentence is used for augmentation is 5549 divided by the number of sentences). As an example, let's consider Sicily which had 608 initial sentences. In this case, each sentence has been used ≈ 9 times to create new data, resulting in 4941 newly generated sentences.

### 3.2. Prediction Through Language Model

The process for classifying the sentences using BERT was the following: (i) we fine-tune BERT on our augmented training set (cfr. § 3.1); (ii) for every sentence in the test set, we use the model to get a prediction on the regional variety of the sentence; (iii) a confidence score for each region is returned.

[5] https://www.vocabolariofiorentino.it/ricerca/lemmi, *retrieved on May 10th, 2023*

[6] https://www.theromanpost.com/2016/06/dizionario-dialetto-romanesco, *retrieved on May 10th, 2023*

[7] https://github.com/MartinoMensio/it_vectors_wiki_spacy

[8] https://tedboy.github.io/nlps/generated/generated/gensim.models.Word2Vec.most_similar.html

| Run | Precision | Recall | Macro F1 |
|---|---|---|---|
| Standard-Track_Run-1 | **0.83** | **0.52** | **0.56** |
| Standard-Track_Run-2 | 0.69 | 0.45 | 0.48 |
| Standard-Track_Run-3 | 0.75 | 0.50 | 0.52 |
| Logistic_regression baseline | 0.62 | 0.42 | 0.46 |
| Most_frequent baseline | 0.05 | 0.21 | 0.07 |
| Special-Track_Run-1 | 0.72 | 0.80 | 0.73 |
| Special-Track_Run-2 | 0.72 | 0.80 | 0.73 |
| Special-Track_Run-3 | 0.81 | **0.83** | **0.82** |
| Logistic_regression baseline | **0.92** | 0.67 | 0.71 |
| Most_frequent baseline | 0.39 | 0.50 | 0.44 |

**Table 1**

Results of the classification of Italian regions. The upper group shows the performances of the 3 runs for the standard track, compared to the "logistic regression" and "most frequent" baselines. The lower group shows the performances of the 3 runs for the special track, compared with the correspondent baselines.

### 3.3. Prediction Through Vocabularies

Among the vocabularies presented in § 2.3, we selected the ones relevant to the specific task (it will be discussed in § 4). The first step was to normalize the format of the different vocabularies, associating each regional word with its frequency value. For the vocabularies retrieved from the web, we first performed web-scraping, disregarding the standard Italian translation of the terms, and then assigned 1 as the frequency for each one of the vocabulary entries. Since the vocabularies generated from the training set came with different frequencies, we had to normalize those values, reassigning a value between 0 and 1, by maintaining the same proportion of the original frequencies. A process of normalization was also performed on the vocabulary words: accented characters were converted into unaccented equivalents, IPA representations presented in the terms were deleted, combinations of words (sometimes multiple entries were considered as one in the web vocabulary) were divided into individual entries (e.g., the entry "c(o/u)mpà" becomes two different entries "compa" and "cumpa").

Once we obtained all the normalized vocabularies, we then merged them into one single vocabulary, by summing all frequencies for the same term in the same variety (e.g. if "compà" has frequency 1 in Vocabulary_A for Sicilian and frequency 0.9 in Vocabulary_B, it will have frequency 1.9 in the global vocabulary, assuming there are only 2 vocabularies).

Finally, to predict the regional variety of a sentence, we sum the frequencies of each word in the sentence for each regional variety, using the frequencies present in the global vocabulary. The region with the highest scoring is the predicted regional variety of the sentence.

### 3.4. Global Prediction

Our final prediction on the region of origin a given sentence leverages both the predictions from BERT (cfr. § 3.2) and from the vocabularies (cfr. § 3.3). We introduce a variable K which regulates the proportions of the contributions given by each one of the 2 predictions. We define the sum of the 2 predictions $S$ for every regional variety $r$ as follows:

$$S(r) = B[r] + K * V[r] \tag{1}$$

where $B[r]$ is the confidence of BERT for the region $r$ and $V[r]$ is the confidence of the vocabularies algorithm for the region $r$.

The final prediction for the believed variety of the sentence is defined by the following expression:

$$\text{global\_pred} = \arg\max_{r \in R} S(r) \tag{2}$$

where $R$ is the set of all regional varieties.

## 4. Experimental Setup

Our experiments were divided into 2 parts: the first one aims at classifying a regional variety among all the 20 Italian regions and was targeted to the Standard Track of GeoLingIt (cfr. § 4.1), while the second one aims at classifying only Toscana and Lazio varieties and was targeted to the Special Track of GeoLingIt (cfr. § 4.2).

### 4.1. Standard Track

For classifying one variety among the 20 Italian regions, we used the 3 vocabularies presented in § 2.3 and the BERT classifier, as described in § 3. During the fine-tuning process of BERT, we experimented with different numbers of training epochs, learning rates, and values of K (cfr. § 3.4), and used the 3 configurations that gave better results on the validation set as the 3 runs of the task:

- **Run-1**: 2 epochs, learning rate $6 * e^{-5}$, K=1;
- **Run-2**: 2 epochs, learning rate $2 * e^{-5}$, K=1;
- **Run-3**: 2 epochs, learning rate $6 * e^{-5}$, K=0.5.

Although we tried with a lower and greater number of epochs, we observed that 2 was the best value for the relatively small training set that we used for fine-tuning. The same approach was used to set the Adam optimiser's learning rate. We started with a learning rate of $2 * e^{-5}$ as suggested by TensorFlow documentation, and gradually decrease or increase it.
Finally, we noted that relying too much on the vocabularies rather than the LLM (using a K greater than 1), did not bring high results, thus we focused on values of K between 0 and 1.

### 4.2. Special Track

We followed the same process also for the special track, using BERT fine-tuned on a corpus of only Toscana and Lazio samples (filtered out from the augmented dataset, described in § 3.1), and the 4 vocabularies presented in § 2.3, 2 generated from the original training-set, one for the Toscana lexicon, and one for the Lazio lexicon. We tried different configurations and used the best 3 as the runs for the special track:

- **Run-1**: 2 epochs, learning rate $4 * e^{-5}$, K=0.5;
- **Run-2**: 2 epochs, learning rate $4 * e^{-5}$, K=0.1;
- **Run-3**: 2 epochs, learning rate $2 * e^{-5}$, K=0.1.

Again, we found that 2 epochs were the optimal value on the validation set and that the best learning rate values were around the suggested value from the literature. In this case, we observed that the contribution of the vocabularies did not bring many advantages, and therefore we kept a low value of K.

## 5. Results

Table 1 shows the results for our 3 runs for the standard track (classification on the 20 Italian regions) and 3 runs for the special track (classification on the Toscana and Lazio varieties), as described in section 4. In addition, we reported also the results for the 2 baselines provided by the organizers of the tasks, one based on a logistic regression model and the other one which simply predicts the most frequent label in the training set for every inference, as described by Ramponi and Casula [11]. These baselines are reported in the table once per group and are relative to the task of the correspondent group.
According to the task's indications, we employ macro-averaged precision, recall and f1-score as evaluation metrics.

## 6. Discussion

For the standard track group, we can observe that our best run obtained a value for the f1-score 8 times higher than the Most_frequent baseline and 10 points higher than the Logistic_regression baseline. Comparing the 3 runs, besides the differences in the hyper-parameters choice (a learning rate of $6 * e^{-5}$ appears to give better results), it is interesting to note that the use of the vocabulary has a positive effect on the performances, since we observe an improvement in all evaluation metrics when we pass from $K = 0.5$ to $K = 1$ (from Run-3 to Run-1, which are identical for all other parameters).
For the special track group, we improved the Most_frequent and Logistic_regression baselines of 38 and 11 points respectively, for what concerns the f1-score. Here, in contrast, the use of the vocabulary seems to be less influential than the learning rate, which results to work better with a value of $6 * e^{-5}$.
This difference between the two tasks can be partially explained by the relative lexical similarity between Toscana and Lazio varieties: in both regions, the regional spoken language does not differ too much from standard Italian and therefore the strategy of distinguishing a sentence on the basis of the vocabulary does not seem to be the right approach. On the other side, when it comes to classifying a sentence between all the 20 Italian regions varieties, the lexical terms differ significantly and therefore our prediction based on the vocabularies proves to be a great improvement to the classification made through LLMs only.

## 7. Conclusion

In this paper, we addressed the classification of *non-standard* Italian Twitter posts according to their regional variety, by combining the prediction obtained using a language model (BERT) with an algorithm utilizing regional varieties dictionaries and word frequency. We contribute to two tasks: a classification at the country level, and a classification according to a subset of regions (Lazio and Toscana).

After briefly introducing the Italian regional varieties, the tasks addressed, and the methodology applied, a description of the data provided and other additional resources retrieved (e.g., online vocabularies) follows. We then explain in detail the methodology used, starting from the augmentation of the training set data, and going through the different techniques used for obtaining the intermediate and final predictions for both tasks. Finally, the evaluation results are shown: the first task achieved a macro F1 score of 0.56, outperforming both the logistic regression baseline (0.46) and the most frequent baseline

(0.07); while in the second task, even if the macro F1 score and the recall show substantial improvement, the overall precision is 11 points lower with respect to the logistic regression baseline (0.81 vs 0.92).

Overall, the knowledge captured from regional dictionaries and word frequencies seems to be effective in capturing the nuances and characteristics of regional varieties. Furthermore, by leveraging BERT's bidirectional pre-training, the system can consider the entire context of a sentence, thereby contributing to accurate predictions.

The availability of curated resources, such as regional dictionaries, played an important role in enhancing the system's performance. However, we acknowledge the limitations of the vocabularies used in our experiments, and further efforts should be made to expand and refine these resources. Finally, we think there are still areas for improvement. Future research could explore more sophisticated and tailored methods for data augmentation and investigate alternative techniques for integrating vocabulary lexical analysis with BERT. Moreover, the inclusion of additional linguistic features and the exploration of ensemble methods could potentially lead to further performance improvements.

## References

[1] C. Moseley, Atlas of the World's Languages in Danger, Memory of peoples Series, UNESCO Publishing, 2010. URL: https://books.google.it/books?id=kFVthqmDs_kC.

[2] C. Bosco, F. Dell'Orletta, S. Montemagni, M. Sanguinetti, M. Simi, The evalita 2014 dependency parsing task, The Evalita 2014 Dependency Parsing Task (2014) 1–8.

[3] F. Dell'Orletta, Ensemble system for part-of-speech tagging, Proceedings of EVALITA 9 (2009) 1–8.

[4] 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.

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

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

[7] 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.

[8] J. Devlin, M.-W. Chang, K. Lee, Google, kt, language, ai: Bert: pre-training of deep bidirectional transformers for language understanding, in: Proceedings of NAACL-HLT, 2019, pp. 4171–4186.

[9] W. Wang, Z. Zhang, J. Guo, Y. Dai, B. Chen, W. Luo, Task-oriented dialogue system as natural language generation, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2022, pp. 2698–2703.

[10] T. Labruna, B. Magnini, Fine-tuning bert for generative dialogue domain adaptation, in: Text, Speech, and Dialogue: 25th International Conference, TSD 2022, Brno, Czech Republic, September 6–9, 2022, Proceedings, Springer, 2022, pp. 513–524.

[11] 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.