Italian Dialects: NLP For Local Linguistics

Maria Grazia Miccoli[1]

[1]*University of Bari - Aldo Moro*

Abstract

Italy's diverse dialects, integral to its cultural heritage, inspired the GeoLingIt Shared Task to classify dialects by region. This report investigates the use of a Support Vector Machine (SVM) for this task, utilizing geotagged Twitter posts that exhibit non-standard Italian. The SVM's proficiency in handling high-dimensional spaces and its versatility with different kernel functions make it well-suited for dialect classification. To improve the dataset, we combined the train and dev datasets and employed the LLaMAntino-3-ANITA-8B-Inst-DPO-ITA model for oversampling. Our approach achieved a 70% accuracy in dialect classification, with Sardinian dialects being identified most accurately. Additionally, we extended the task to include dialect translation into Italian, achieving a BLEU score of 0.17. These results underscore the potential of our approach while also highlighting its limitations, suggesting a need for larger datasets and manual translation to enhance accuracy.

Keywords

Italian dialect, NLP, Local Linguistics, Classification, Translation

1. Introduction and Motivations

In Italy, there is a wide variety of dialects, so numerous that their exact number remains undefined. Some of these dialects are still widely used, even alongside the official language, Italian, while others are less common but still in use. As Mazzaggio and Neri point out in their article [5], the identity of a region is also reflected in the languages that local people use, not only for speaking but also for expressing their values and beliefs through artistic expressions such as music, dialectal plays, literary works, and much more. Let us think ,for example, how neomelodic songs immediately evoke the region of Campania; how the Tarantella brings to mind Puglia and how a simple dialectal expression, such as 'Daje', makes us even identify a precise place (in this case Rome). Let's also consider the wonder a non-local person feels when discovering the traditions, typical foods, and much more through the dialect of the area, often incomprehensible to outsiders. Even locals, perhaps from nearby towns, can be surprised to find such a variety of expressions just a few kilometers away.

So, recognizing the significance of dialects, the GeoLingIt Shared Task site has introduced a challenge to classify Italian dialects by region. This initiative is part of a broader effort to document and preserve these linguistic varieties, ensuring they remain an integral part of Italy's cultural heritage. Additionally, to promote the dissemination and appreciation of dialects, the task has been extended to include a translation component. By translating dialects into Italian, this project aims to foster greater understanding and marvel at the linguistic diversity that characterizes Italy.

2. Related Work

In this section, we examine the existing state of the art and previous studies that are relevant to our project. The related work is organized into two main areas: classification of the dialect text by region of affiliation and translation. By examining these areas, we can contextualize our goal and build on solid foundations to develop our project.

NLP Course 2023-2024

m.miccoli45@studenti.uniba.it (M. G. Miccoli)

https://https://github.com/MariaGraziaMiccoli/ (M. G. Miccoli)

2.1. Classification of the dialect text

My project draws its foundational idea from the challenge proposed by GeoLingIT. GeoLingIT is the first shared task on the geolocation of linguistic variation in Italy, focusing on social media posts that exhibit non-standard Italian language. Part of the EVALITA 2023 evaluation campaign, this challenge aims to advance natural language processing (NLP) in handling non-standard Italian and to provide sociolinguistic insights from large-scale, quantitative analysis to enrich linguistic atlases. The motivation behind this challenge stems from the increasing amount of informal, user-generated text on social media, which reveals significant linguistic variation across Italy, a country known for its high linguistic diversity.

To tackle this challenge, I applied a machine learning algorithm called Support Vector Machine (SVM) [4] for classification. SVM is a supervised learning model that analyzes data and recognizes patterns, used for classification and regression analysis. The SVM classification works by finding the hyperplane that best separates different classes in the feature space. It is particularly effective in high-dimensional spaces and is versatile due to its different kernel functions. This approach allows for the effective classification of social media posts based on their linguistic content, thereby contributing to the objectives of GeoLingIT.

2.2. Translation of the dialect text

For the extension of the task to include dialect translation, we took inspiration from the article by Abdelaziz et al.[1], which describes a project focused on translating from Dialectal Arabic to Modern Standard Arabic (MSA) using neural networks. Given that this task is similar but in a different linguistic domain, our project attempts to adapt the techniques used by Abdelaziz et al. for Italian dialect translation.

Specifically, they started with a dataset composed of both original and synthetic Dialectal Arabic sentences and used a large language model (ChatGPT 3.5 [2]) to generate corresponding translations in MSA. They then developed a transformer-based model for automatic translation and facilitated its training using two neural machine translation frameworks, Marian NMT [6] and JoeyNMT[7], which are known for their efficiency in creating high-quality translation models. Finally, they evaluated their model on different versions of the dataset using the BLEU-score [8].

3. Proposed Approach

The various phases of the project will be described in detail below.

3.1. Dataset

My project is based on the dataset provided by GeoLingIT, which includes social media posts geotagged by Twitter (current X) showing a non-standard use of the Italian language. They provided two sets of data: one called ***train*** for training and the other called ***dev*** for testing. Language variations in GeoLingIT data can appear as single words or phrases (elements in a local language, dialect or regional synonyms such as "Guaglione", "Toso", "Picciotto" for "young man"), through code-switching (alternation of standard Italian and local language, dialect or regional variant), or as whole posts written in a specific local language or dialect. The data set is in a tab-separated format, with each example on a separate line and the first line serving as the header. Each example is described by three characteristics:

1. **id**: tweet identifier (anonymised to preserve user anonymity).
2. **text**: the text of the tweet (with anonymous user mentions, e-mail addresses, URLs and location mentions).
3. **region**: tweet region in string format.

Given the presence of lexical elements that were not important for the task, it was decided to apply to both pieces of dataset a pre-processing phase in which tags, emoticons and hashtags were deleted from each example.

3.1.1. Oversampling of the data

Analyzing the training dataset in detail, two main issues emerged. The first was the insufficient amount of total examples in the dataset, which did not allow to reach a classification accuracy higher than 50%. The second problem concerned some regions, such as Valle d'Aosta, Trentino-Alto Adige, Basilicata and Molise, whose dialects were represented by very few examples compared to other regions that had at least a hundred examples. As a result, these regions had minority classes. In contrast, Lazio and Campania, with respectively 5587 and 3016 examples, represented the majority classes.

Two steps were followed to address these issues. The first step was to combine the two datasets, train and dev, to increase the total number of examples. The second step was to identify an appropriate generative model that, using existing examples for each region, could generate new examples for regions with less representation (except Lazio and Campania). The oversampling phase proved to be the most difficult, because it was necessary to find a model capable of creating sentences not only of complete sense, but also respectful of the specific syntax of each dialect. Moreover, in this experiment, for the qualitative evaluation of the generated text, the Apulian dialect was taken into consideration, as we could have a direct comparison with the local population. The model that gave the best results for the Apulian dialect, was then used for the dialects of other regions. The models created from scratch are the follows:

1. **N-GRAM Model** [3]: the first model tested is the N-GRAM model. A language model is a probabilistic model that is used to assign a probability to a sequence of words. For example, if we have a group of words and we take the first word, the model can predict the next word, which is the one with the greatest probability of standing next to the first. In our case, the model considered 3-gram to predict the next word but, after various experiments, it created sentences that were simply a sequence of meaningless words.
2. **LSTM**[13]: A Long Short-Term Memory (LSTM) is a type of recurring neural network (RNN) designed to model long-term data sequences. It is particularly useful for natural language processing (NLP) applications such as text generation. LSTM overcomes the fading gradient problem of traditional RNN due to its special architecture that includes memory cells and port mechanisms (input, output and forget) that control the flow of information. In our case the neural network created based on it consists of 4 layers: embedding layer, for the internal representation of sentences, LSTM layer, dropout layer, to improve the efficiency of the network, and a dense layer that acts as an output layer. This model, however, does not satisfy the generation task as it simply gives us a piece of the same sentence given in input.
3. **VAE**[9] A Variational Autoencoder (VAE) is a type of neural network used to learn latent representations of data, useful for NLP text generation and modeling. The VAE combines autoencoder techniques with probabilistic generative models, allowing new data samples similar to training ones to be generated. Their structure consists of an encoder that maps the input data into a probabilistic latent space and a decoder that reconstructs the original data from the points in the latent space. But, in our case, not even this type of generative model can give us good results, most likely due to the limited number of examples.

At this point, the following pre-trained models were tested:

1. **Gemini-pro**[12]: through the API provided by Gemini, it was possible to experiment with the transformer based model gemini-pro. Already from an initial experiment we have seen that the model does not work so badly and generates sentences very close to the original dialect.
2. **GPT-2** [11]: GPT-2 is another transformer based model, but published by OpenAi. It gives good results, but occasionally gives results in English.

3. **LLaMAntino-3-ANITA-8B-Inst-DPO-ITA**[10]: The selected model is LLaMAntino-3-ANITA-8B-Inst-DPO-ITA which is a large language model developed for advanced natural language processing (NLP) applications in Italian, such as text generation, machine translation, completion of sentences and answers to questions. Thanks to its 8 billion parameters, it can handle complex tasks and provide more accurate and contextually relevant answers. This was the model that best met our expectations and was used to complete the over-sampling phase.

After completing the over-sampling phase using the latter model, it was possible to work with a dataset containing 34479 examples, where each region has at least 1000 examples each.

3.2. Main task: classification of examples by region.

To address this challenge, a machine learning algorithm known as Support Vector Machine (SVM) was used for classification purposes. The central mechanism of SVM classification is to identify the optimal hyperplane that effectively separates different classes within the space of functionality. This method is particularly powerful in large spaces, where it can manage data complexity with ease. In addition, SVM's versatility is enhanced by its various kernel functions, which allow it to adapt to a wide range of data types and distributions. By leveraging SVM for the classification of tweets, based on their linguistic characteristics, this approach contributes significantly to the achievement of the objectives set by the GeoLingIT project. To use this approach, the model in the scikit-learn library was used.

3.3. Extra task: translation of dialect phrases

The LLaMAntino-3-ANITA-8B-Inst-DPO-ITA model was used to conduct a first experiment in this field, which, with the following prompt, translated 100 Apulian dialect sentences into Italian:

```
prompt=f"""Testo: "{sentence[0]}".
        Il testo è scritto nel dialetto della regione italiana {sentence[1]},
        traducilo in lingua italiana senza cambiare
        nè la sua semantica nè la sua sintassi.
        Termina la traduzione con un "\n".
        Non devono essere date in output altre informazioni
        oltre la traduzione.
        Non aggiungere parentesi o altri commenti.
        Non aggiungere parole inglesi o italiane che
        non siano già presenti nel testo.
        Lascia invariati i termini che sono già all'interno
        del testo in lingua italiana o inglese.
        Rispondi solo con la traduzione letterale.
        Non saltare nessuna parte del testo.
        Neanche quelle che originariamente sono tra perentesi nel testo."""
```

The experiment was conducted only with the first 100 sentences from Puglia because the task required a manual translation for the evaluation phase of the sentences that took a little more time.

4. Evaluation

4.1. Evaluation for the main task

We can see the results obtained within Figure 1. We can see that:

1. Accuracy: The model achieved an overall accuracy of 70%, indicating that 70% of the phrases were correctly classified.
2. Recall medium: The average recall is 65%, indicating that the model is moderately effective in capturing all dialect phrases for each region.

```
==========================================================================================================
Accuracy: 0.7005137770278733
P=0.7101501776285007, R=0.6471214558008442, F1 Macro=0.671757058855882, F1 Micro=0.671757058855882
==========================================================================================================
                      precision    recall  f1-score   support

              abruzzo       0.64      0.51      0.57      1300
           basilicata       0.68      0.58      0.63      1304
             calabria       0.70      0.57      0.63      1508
             campania       0.78      0.85      0.81      3046
       emilia romagna       0.61      0.50      0.55      1379
friuli-venezia giulia       0.72      0.66      0.69      1387
                lazio       0.63      0.96      0.76      5614
              liguria       0.77      0.61      0.68      1319
            lombardia       0.70      0.60      0.64      1802
               marche       0.66      0.51      0.57      1321
               molise       0.66      0.57      0.61      1226
             piemonte       0.68      0.61      0.64      1413
               puglia       0.69      0.60      0.64      1429
             sardegna       0.91      0.85      0.88      1535
              sicilia       0.76      0.79      0.77      1848
              toscana       0.71      0.65      0.68      1649
  trentino-alto adige       0.71      0.60      0.65      1226
               umbria       0.65      0.51      0.57      1246
       valle d'aosta       0.77      0.65      0.71       948
              veneto       0.73      0.78      0.75      1979

            accuracy                           0.70     34479
           macro avg       0.71      0.65      0.67     34479
        weighted avg       0.70      0.70      0.69     34479
```

Figure 1: Result of the main task

3. F1-average score: The F1 average score is 67%, which represents a good balance between accuracy and recall.

In detail we have that:

1. The dialect of Sardinia is classified more precisely.
2. The dialect of the Emilia-Romagna region is classified with less precision.

4.2. Evaluation for the extra task

As did Abdelaziz et al., also in this case the BLUE score was used as a method of evaluation, going to obtain a score of 0.17. In general, a BLEU score of 0.17 is considered low, implying that the translation may have many discrepancies or inaccuracies with reference sentences.

5. Conclusions and Limitations

As for the first task, the classification task, we have achieved better results than we expected. The factor that limited accuracy to only 70% was, despite oversampling, the reduced number of examples. Moreover, the blue score so low about the translation task, is also due to the presence of synthetic sentences that do not capture the actual essence of the dialect. Consequently, further improvements that could follow are:

1. Repeat the experiment using a more accurate dataset with a much larger number of examples.
2. Associate each sentence in this dataset with a manual translation by local people.
3. Fine-tuning the LLaMAntino-3-ANITA-8B-Inst-DPO-ITA model for better machine translation.

References

[1] AhmedElmogtaba Abdelmoniem Ali Abdelaziz, Ashraf Hatim Elneima, and Kareem Darwish. "LLM-based MT Data Creation: Dialectal to MSA Translation Shared Task". In: *Proceedings of the 6th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT) with Shared Tasks on Arabic LLMs Hallucination and Dialect to MSA Machine Translation @ LREC-COLING 2024*. Ed. by Hend Al-Khalifa et al. Torino, Italia: ELRA and ICCL, May 2024, pp. 112–116. URL: https://aclanthology.org/2024.osact-1.14.

[2] Aram Bahrini et al. "ChatGPT: Applications, Opportunities, and Threats". In: *2023 Systems and Information Engineering Design Symposium (SIEDS)*. 2023, pp. 274–279. DOI: 10.1109/SIEDS58326.2023.10137850.

[3] Peter F Brown et al. "Class-based n-gram models of natural language". In: *Computational linguistics* 18.4 (1992), pp. 467–480.

[4] Chih-Chung Chang and Chih-Jen Lin. "LIBSVM: A library for support vector machines". In: *ACM Transactions on Intelligent Systems and Technology* 2 (3 2011). Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm, 27:1–27:27.

[5] Neri Binazzi Greta Mazzaggio. "Valorizzare il patrimonio immateriale: un'esperienza di digitalizzazione del dialetto". In: *DILEF. Rivista digitale del Dipartimento di Lettere e Filosofia - 3(2024), pp. 224-242. 10.35948/DILEF/2024.4348* (2024).

[6] Marcin Junczys-Dowmunt et al. "Marian: Fast Neural Machine Translation in C++". In: *Proceedings of ACL 2018, System Demonstrations*. Ed. by Fei Liu and Thamar Solorio. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 116–121. DOI: 10.18653/v1/P18-4020. URL: https://aclanthology.org/P18-4020.

[7] Julia Kreutzer, Jasmijn Bastings, and Stefan Riezler. "Joey NMT: A Minimalist NMT Toolkit for Novices". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*. Ed. by Sebastian Padó and Ruihong Huang. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 109–114. DOI: 10.18653/v1/D19-3019. URL: https://aclanthology.org/D19-3019.

[8] Kishore Papineni et al. "BLEU: a method for automatic evaluation of machine translation". In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. ACL '02. Philadelphia, Pennsylvania: Association for Computational Linguistics, 2002, pp. 311–318. DOI: 10.3115/1073083.1073135. URL: https://doi.org/10.3115/1073083.1073135.

[9] Lucas Pinheiro Cinelli et al. "Variational autoencoder". In: *Variational Methods for Machine Learning with Applications to Deep Networks*. Springer, 2021, pp. 111–149.

[10] Marco Polignano, Pierpaolo Basile, and Giovanni Semeraro. "Advanced Natural-based interaction for the ITAlian language: LLaMAntino-3-ANITA". In: *arXiv preprint arXiv:2405.07101* (2024).

[11] Alec Radford et al. "Language models are unsupervised multitask learners". In: *OpenAI blog* 1.8 (2019), p. 9.

[12] Demis Hassabis Sundar Pichai. "Introducing Gemini: our largest and most capable AI model". In: Google, 2023. URL: https://blog.google/technology/ai/google-gemini-ai/#sundar-note.

[13] Yong Yu et al. "A review of recurrent neural networks: LSTM cells and network architectures". In: *Neural computation* 31.7 (2019), pp. 1235–1270.