

The curious case of Logistic Regression for Italian Dialects Identification

Giacomo Camposampiero 🌱, Francesco Di Stefano 🌱, Quynh Anh Nguyen 🌱🚀
 🌱ETH Zürich, 🚀University of Milan

Introduction

Automatic Language Identification represents an important task for improving many real-world applications. We propose an extensive evaluation of different approaches for the identification of **Italian dialects and languages**, spanning from classical machine learning models to more complex neural architectures. This work was developed in the context of the Identification of Languages and Dialects of Italy task organized at **VarDial 2022**.

Method Overview

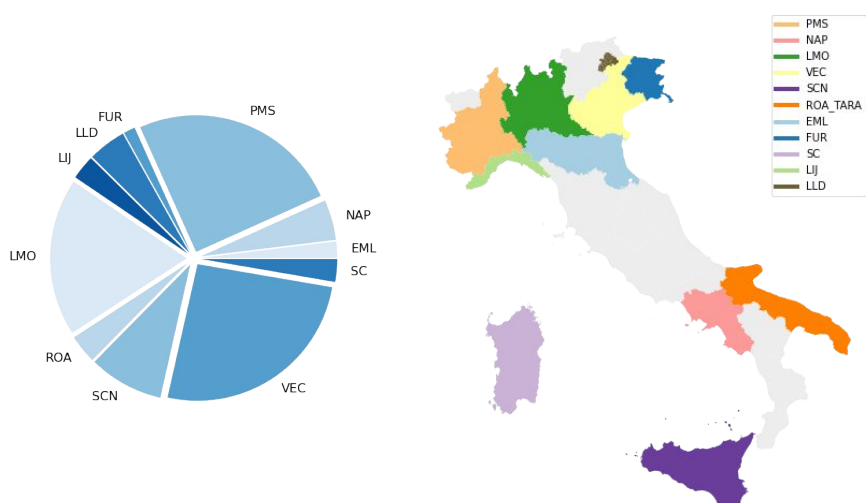
We tackled the identification problem exploiting three different architecture:

- ❑ **Linear models:** We experimented with three different models, namely Linear Support Vector Machines, Naïve Bayes classifiers and Logistic Regression. The models are trained on scaled word-level TF-IDF feature vectors.
- ❑ **CNN:** We implemented both word-based and character-based networks. All networks are 3 layers deep, with 2 convolutional layers and 1 fully- connected layer.
- ❑ **Transformers:** We fine-tuned 6 different HuggingFace BERT models pre-trained on Italian corpora for two epochs.

Predictions are evaluated with f1-micro at sentence level.

Data Exploration

- ❑ The dataset was provided by the organizers and consists of Wikipedia dumps including samples for all the eleven dialects.
- ❑ **Challenges** observed with the given dataset:
 - ❑ Imbalanced classes.
 - ❑ Missing dialects in the validation set, difficult to fine-tune the models for those.
 - ❑ Validation and test sources different from Wikipedia, possible poor out-domain model performances.



Results and Discussion

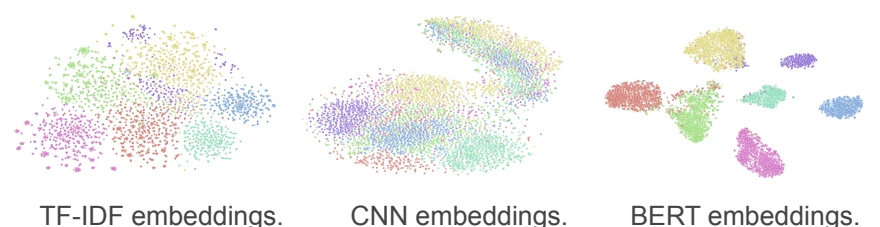
The evaluation results on the validation set for the best-scoring models of each categories are shown in the table.

Model	f1-micro
Logistic Regression	0.9445
Character-level CNN	0.7987
BERT _{LARGE}	0.8907

- ❑ Logistic Regression achieved the best performance.
- ❑ Key speculations on the model
 - ↑ It leverages the consistent linguistic variety between the evaluated Italian dialects and languages.
 - ↑ It is simple and explainable. The number of parameters learned by the model is relatively small (~5 million), compared to other investigated models (BERT has 110 million parameters).
 - ↓ Impossibility of handling OOV problem and possibility of model overfitting to the validation set.

Analysis

- ❑ **Error analysis:** confusion matrix and explained logistic regression predictions
- ❑ Training and inference **time comparison** between models
- ❑ Feature space **visualization**



Conclusions

- ❑ Logistic Regression model achieved the best results, outperforming the other two models and ranking within the top 5 submissions the Vardial 2022 ITDI shared tasks.
- ❑ No notable difference in the performance of character-based and word- based CNN, of which the vast vocabulary size is more costly in terms of training time.
- ❑ BERT models performed weakly in this cross-domain language identification task, generalising less than linear models.



[Github](#)



[Demo](#)