Identification of Languages and Dialects of Italy Project Proposal

Sravya Sangaraju
George Mason University
ssangara@gmu.edu

Anisha Yidala
George Mason University
ayidala@gmu.edu

1 Introduction

1.1 Task

The idea for this project was taken from the Languages and Dialects of Italy (ITDI) task that was designed for the VarDial Evaluation Campaign 2022 . The task is to build a classification model that differentiates among 11 regional languages and varieties of dialects (Piedmontese, Venetian, Sicilian, Neapolitan, Emilian-Romagnol, Tarantino, Sardinian, Ligurian, Friulian, Ladin, Lombard) in Italy. We plan on using the same data-set that has been used for this task in the Vardial 2022 Campaign , which is a Wikipedia dump and hence preprocessing of this data will also be a major part of our task.

1.2 Motivation and Limitations of existing work

There is existing work of similar models built for classification and dialect identification on similar languages but not many researchers have worked specifically in the context of Italian language . The most common problem encountered while working on classification of language varieties included lack of reliable datasets with adequate representation of diversity , lack of availability of actual speaker centred data as dialect identification heavily relies on this rather than textual data. However, several researchers have achieved state of the art classification models through application of NLP techniques.

1.3 Proposed Approach

We aim to follow different approaches using the Deep Learning models with Transformer and CNN architectures for identification task of Italian dialects.

We plan to use a pre-trained BERT on Italian language and fine-tune it for our problem. In the domain of dialect identification, SOTA results were generated using the transformer models and especially fine-tuning a pre-trainined BERT gives good results for the dialect identification tasks.

Next, we also plan to implement a classifier along with a CNN that will be used for feature extraction on an embedding layer, since CNNs were previously used for dialect identification and successfully produced good results.

1.4 Likely challenges and mitigations

The main challenge we have to tackle in order to successfully develop this model is to carefully preprocess the raw wikipedia dumps. We have observed that this dataset is heavily imbalanced and there is inappropriate representation of wikipedia articles across dialects. Possible solutions we would like to explore to mitigate this issue are using techniques like weighted modelling and data sub sampling.

2 Related Work

Dialect classification , language variation identification or automatic dialect recognition has been an area of interest for researchers for a while now and we were able to go through a few papers that were published. Various methods were implemented for other languages and their dialects on classification problems.

One major issue is that no experiments on the task of language identification of Italian dialects are available, despite a lot of research is going on in the field of analysis of Italian dialects features"Andrea Zugarini and Maggini. (2020). Our goal is to find a solution to come up with good performance by studying the work done on different languages for dialect identification with the help of Transformer and CNN architectures.

Models that were built on the Transformer architecture "Ashish Vaswani and Polosukhin

(2017) were successful in achieving the SOTA outcomes as this architecture when introduced had transformed the way tasks in NLP were performed. Identification of dialects is one such NLP task which generated great results when implemented using the Transformers architecture. A fine-tuned BERT trained on Romanian corpora"George-Eduard Zaharia and Rebedea (2020) was used in the VarDial 2020 Evaluation Campaign for the task of Romanian vs Moldavian identification. This generated a weighted F1 score of 96.25 percent on Moroco dataset"Butnaru and Ionescu (2020).

Despite the fact that transformer-based models enabled various dialects identification tasks achieve SOTA, the usage of CNNs for solving such problems is higher. CNN models generated cutthroat results for identification of Romanian vs Moldavian in VarDial Evaluation Campaigns in both 2019"Tudoreanu. (2019) and 2020"Rebeja and Cristea. (2020).

3 Experiments

3.1 Datasets

The training dataset that we are using is in the form of raw Wikipedia dumps, that consists of Wikipedia articles from 1st March 2022 dumps. It is accessible from the git repository which was publicly made available for Findings of the VarDial Evaluation Campaign 2022"Aepli et al. (2022).

After a brief investigation of the data, a heavy imbalance among the representation of dialects was observed which needs to be carefully addressed.

3.2 Baselines

We have not found any papers that have been published for the same task on Italian language yet . We are yet to research and look for a Baseline SOTA implementation that we can reproduce for our checkpoint 2. However we have an approach to design our model using NLP models like CNNs and transformers which have been proven to produce great results for language classification tasks on other similar languages.

3.3 Timeline

The timeline for completion of this project is tentatively defined as follows :
1) Week 1 (October 11): Analysing the training data set and preprocessing the datasets.
2) Week 2 (October 18): Generating a Reliable test and dev data set from the available wiki repositories and Start Implementing the SOTA model which we plan to reproduce
3) Week 3 (October 25): Implementation of SOTA model and Error Analysis
4) Week 4 (November 1): Writing a PDF report for checkpoint 2 with the results we achieved and its comparison with our Baseline model.
We plan on working on every task as a team initially up to weak 2 . In Week 3 we would like to split and implement one part each to achieve efficiency.

References

Noëmi Aepli, Antonios Anastasopoulos, Adrian Chifu, William Domingues, Fahim Faisal, Mihaela Găman, Radu Tudor Ionescu, and Yves Scherrer. 2022. Findings of the VarDial evaluation campaign. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*.

Matteo Tiezzi Andrea Zugarini and Marco Maggini. 2020. Vulgaris:vulgaris: Analysis of a corpus for middle-age varieties of italian language.

Niki Parmar Jakob Uszkoreit Llion Jones Aidan N Gomez L ukasz Kaiser Ashish Vaswani, Noam Shazeer and Illia Polosukhin. 2017. Attention is all you need. In *In Advances in Neural Information Processing Systems, volume 30*.

Andrei M. Butnaru and Radu Tudor Ionescu. 2020. Moroco: The moldavian and romanian dialectal corpus.

Dumitru-Clementin Cercel George-Eduard Zaharia, Andrei-Marius Avram and Traian Rebedea. 2020. Exploring the power of romanian bert for dialect identification. In *In Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects, pages 232–241, Barcelona, Spain (Online)*.

Petru Rebeja and Dan Cristea. 2020. A dual-encoding system for dialect classification.

Diana Tudoreanu. 2019. VarDial 2019: Ensemble based on skip-gram and triplet loss neural networks for moldavian vs. romanian cross-dialect topic identification. In *In Proceedings of the Sixth Work- shop on NLP for Similar Languages, Varieties and Dialects*, page 202–208.