Identification of Languages and Dialects of Italy Report

Sravya Sangaraju
George Mason University
ssangara@gmu.edu

Anisha Yidala
George Mason University
ayidala@gmu.edu

1 Introduction

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. The shared task organizers provide a dataset consisting of a large pool of Wikipedia articles written in one of these dialects, in the form of Wikipedia dump. We have used the same for our task. Dialect classification represents a key task in the improvement of many other downstream tasks such as opinion mining and machine translation, where the enrichment of text with geographical information can potentially result in improved performances for real-world applications As a result, the interest in the study of language variation has been steadily growing in the last few years, as highlighted by the increasing number of publications and events related to the topic. 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 centered 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. 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-trained BERT gives good results for the dialect identification tasks. We have focused on explaining the performance of the model . The main challenge we had to tackle in order to successfully develop this model is to carefully pre-process the raw Wikipedia dumps. We have observed that this dataset is heavily imbalanced and there is inappropriate representation of Wikipedia articles across dialects. We have also noticed that certain dialects were not represented in the dev set provided in the VarDial task. F1 score accuracy was used as a metric to evaluate our model The transformer model that we implemented is performing with an accuracy of 0.73 which was approximate to the previously implemented models for the same task. After implementing the baseline model, we have analyzed and explained our models performance and how certain issues like the class imbalance in the data have impacted the efficiency.

2 Approach

Generating a training data set from raw Wikipedia dumps by extracting them and preprocessing them is the initial step. The raw Wikipedia dumps used

Language/Dialect	ST tag	Dump name with date	.bz2 size
Emiliano-Romagnolo	EML	emlwiki-20220301	9.3 MB
Friulian	FUR	furwiki-2022031	2.5 MB
Ladin	LLD	lldwiki-20220301	2.8 MB
Ligurian	LIJ	lijwiki-20220301	6.6 MB
Lombard	LMO	lmowiki-20220301	25 MB
Neapolitan	NAP	napwiki-20220301	5.4 MB
Piemontese	PMS	pmswiki-20220301	14 MB
Sardinian	SC	scwiki-20220301	7.2 MB
Sicilian	SCN	scnwiki-20220301	12 MB
Tarantino	ROA_TARA	roa_tarawiki-20220301	6.4 MB
Venetian	VEC	vecwiki-20220301	27 MB

Figure 1: WikiDumps

to generate the train data set are explained in Figure 1. For choosing an approach to implement a baseline model for the defined task, we have explored several linear classifiers machine learning methods and transformer methods. The use of transformer-based models yielding state of the art results has revolutionized many tasks in NLP and the task of dialect identification isn't any exception. In particular, the fine-tuning of pre-trained

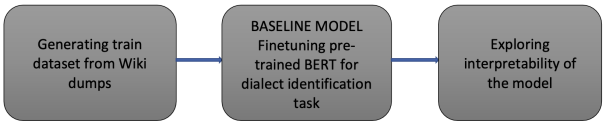


Figure 2: Flow of the task

BERT models obtained good results in this field. Following this line, our approach to implement a baseline model by adopting a version of BERT pre-trained with the Italian language and to fine-tune it on our task was coined. We have explored the interpretability of the model we implemented by using several techniques.

3 Experiments

3.1 Dataset

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

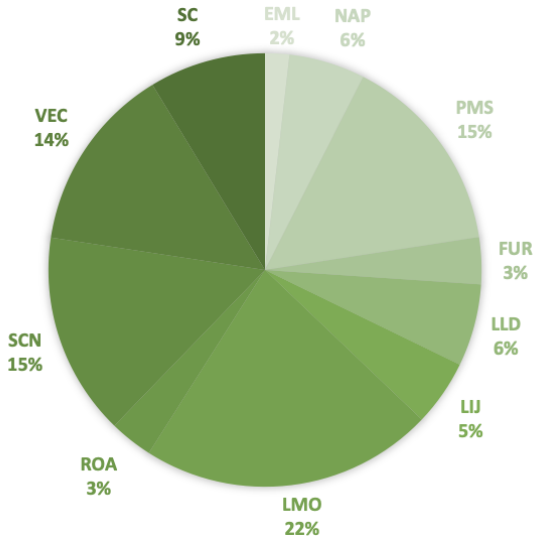


Figure 3: Distribution of 11 Dialects across Train data

was publicly made available for Findings of the VarDial Evaluation Campaign 2022"Aepli et al. (2022).

The development set consists of 6800 annotated sentences that cover only 7 out of the 11 dialects evaluated in the shared tasks (there are no development samples for Emilian, Neapolitan, Ladin and Tarantino dialects). Since the data is not a documented dataset, a preliminary exploration has been initially conducted to gain useful insight about it. This investigation highlighted a huge imbalance between classes as shown in Figure 3, since the 3 most represented dialect (Venetian, Piedmontese and Lombard) account for almost three quarters of the entire articles in the training data.

3.2 Generating and preprocessing the train dataset

As the training data is not a documented dataset we have generated the training dataset from the downloaded data after some preliminary processed data extraction using WikiExtractor a Python script that extracts and cleans text from Wikipedia database backup dumps. However, we noticed that this data contained information (like time stamps, contributions etc.) which might not help training the model. So, we have removed all

Pre-processing steps	No. Of documents
Original documents	265016
Remove length <50	245193
Remove duplicates	219178
Sentence split	562495
Sentence cleaning	388781

Figure 4: No. Of training documents after each preprocessing step

the HTML tags (e.g. br tags, amp; , etc.) and Wikipedia meta information (e.g. contributors, timestamps and comments) Then, we observe that most of the documents whose length is less than 50 are noisy observations, that come from documents for which WikiExtractor failed to extract any text at all or pages that contain simple and repetitive name entity definitions (e.g. small towns or years articles). The training set also contains duplicate documents (e.g., Web domains pages in Venetian Wikipedia) hence we remove all the duplicates in the remaining dataset. We split all the documents into sentences using an Italian spaCy tokenizer. The code for the pre-processing and generation of train.csv file that we used to train our model is available in generation.py.

3.3 Baseline Model and Interpretability of Model

We have fine-tuned a BERT model that was pre-trained on Italian language corpora for the classification task (ITDI). We used BERT base model pre-trained on Italian corpora from the hugging face. The choice of hyperparameters we made were according to the suggestions from the paper BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (1). The BERT model that we finetuned is performing diversly on various dialects. It was observed that it obtains worse results on certain dialects like Venetian but performs decently well on the other dialects. Our model has an accuracy of 0.73. The implementation of this is available on Italian-Dialects-Classification.ipynb. We have added a

Dialect	Precision	Recall	F1	Support
EML	0.99	0.71	0.83	825
NAP	0.85	0.77	0.81	2026
FUR	0.98	0.95	0.96	1323
LLD	0.98	0.56	0.71	2200
LIJ	0.87	0.89	0.88	2282
LMO	0.39	0.98	0.56	689
ROA_TARA	0.57	0.09	0.15	603
VEC	0.53	0.59	0.56	1139

Figure 5: Baseline model evaluation

predicted class label to the annotated dev dataset and after analyzing the results predicted by the baseline model against the actual labels we have noticed that approximately 3000 articles are misclassified out of the 8000 annotated articles in the dev set. We have generated confusion matrices to

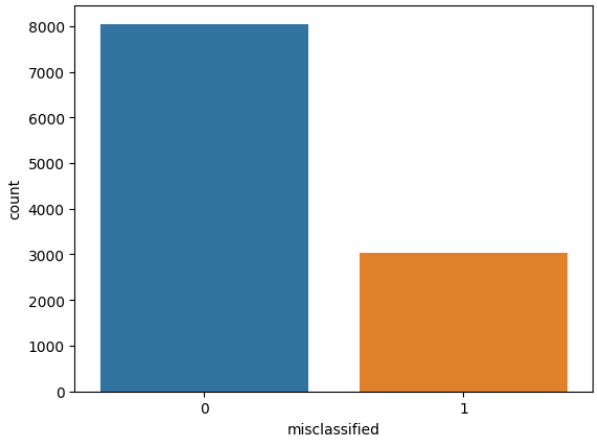
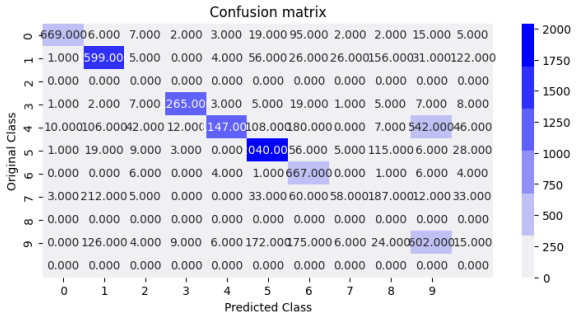
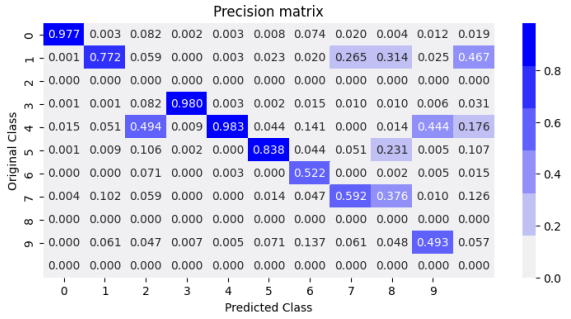


Figure 6: Misclassification of articles

understand which dialects were misclassified the most as other dialects. From the depicted confusion matrices we have identified the two dialects





that have been most misclassified as each other and tried to understand the reason for this. After using several techniques we have come across the Word cloud method , by drawing the wordclouds as shown in figures 7 and 8 for these dialects we have concluded that the dialects that have similarly spelled words and words that share common root words in both languages are at the highly likely to be misclassified by our model.

4 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

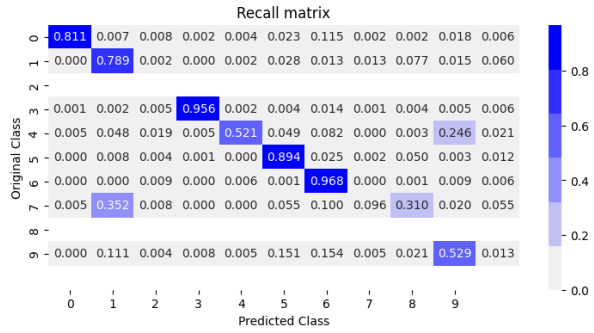


Figure 7: Wordcloud for EML

Figure 8: Wordcloud for NAP

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).

5 Conclusions and future work

Our approach to solve this task is designed by relying heavily on the high performance of transformer models for several Natural Language Processing tasks however the accuracy we have achieved by finetuning a Pretrained BERT model for the dialect identification tasks 0.73 is quite similar to that which was achieved on Linear models like SVM by other members who have worked on this task earlier. We have also tried to analize the results generated by our baseline model and identify the most common reasons for misclassification of dialects by our model which can be found in the Analysis.ipynb file. We have concluded that the dialects that have similar words and words that share common root words in both dialect are at the highly likely to be misclassified by our model.

References

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.