AI4D Malawi News Classification Challenge

We would like to reiterate our thanks to the ZINDI and the AI4D teams for organizing this competition. This document aims to bring more clarification on our final approach and to give an overview of the various experiments performed by our team.

Contents

1. Architecture of the solution

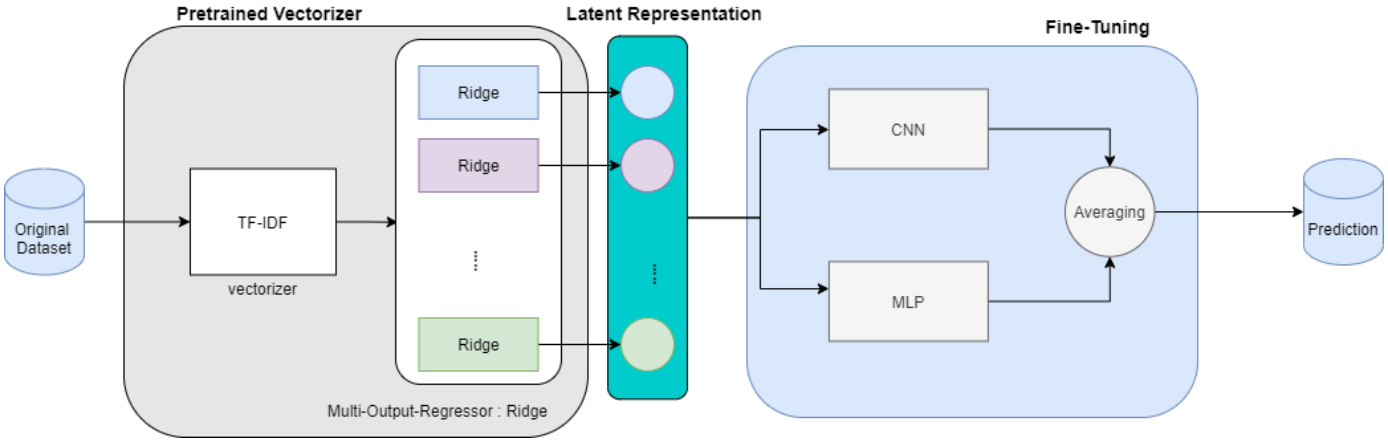

Figure 1: Final architecture of the proposed solution

The final solution involves a pre-trained Vectorizer upstream of the chain which consists of:

- a TF-IDF vectorizer,
- and a Multi-Output- Ridge Regressor.

The outputs produced by this vectorizer will be ingested by two models:

- A two-headed model composed of a 1D convolutional network and FFN
- MLP Classifier

The final prediction is computed by a weighted average of the output of these two models

2. Benchmarking

The following tables summarize a benchmark of the various approaches and experiments we have done:

Approach	**CV**	**LB**	**Inputs source**	**Model Size**	**Notebook Name**
Custom vectorizer + CNN + MLP	69.63	71.29	Chichewa	200Mo	1st_place_solution.ipynb
Custom vectorizer + MLP + SVC	69.22	71.29	Chichewa	13Mo**	MLP_SVC.ipynb
ResNet	66.15	70	Chichewa	200Mo	ResNet.ipynb

Table 1: Performance Benchmarking

3. Pre-training

3.1. Pseudo-Labeling

The choice of a pseudo-labeling approach is mainly motivated by the following reason: typically, when we are confronted with relatively long textual documents from various sources, it turns out that whatever method we use to label these documents, the **assigned classes will NOT mutually be exclusive**.
To illustrate our point, an article that talks about "WILDLIFE/ENVIRONMENT" is not totally different from an article that talks about "FLOODING", or an article that talks about "POLITICS" is not really different from an article about "ECONOMY", especially when the article talks about the implementation of an economic policy for example. From this observation, we conclude that using a "**Hard Labeling" approach may not be the best way to come up with**. This prompted us to adopt the "**Soft Labeling**" approach. Obviously manual soft labeling is excluded, we used an ensemble of several models to obtain "Pseudo soft labels" of the data, which will then be learnt by our vectorizer as latent representation. (see figure 2).

Teacher	Notebook	Execution time	Accelerator
TF-IDF + MLPClassifier	mlp_baseline.ipynb	10Min	CPU
TF-IDF + Truncated Label+ MLPClassifier	mlp_truncated_baseline.ipynb	10Min	CPU
DNN model	DNN_baseline.ipynb	50 Min	CPU
DNN +Truncated Label	DNN_truncated_baseline.ipynb	15 Min	CPU

Table 2: Teacher Model

3.2. Distillation

Once we have this latent representation, a smaller model will learn to map our original Chichewa texts to it. We found out that a **combination of Tf-idf combined with a ridge multi-regression** produces a much lighter model that allows to keep the performances of the teacher and which is much easier to handle and fine tune.

Student	Notebook	Execution time	Accelerator
TF-IDF + Multi Ridge	Distilation.ipynb	12Min	CPU

Table 3: Student Model

The diagram below summarizes the pre-training process.

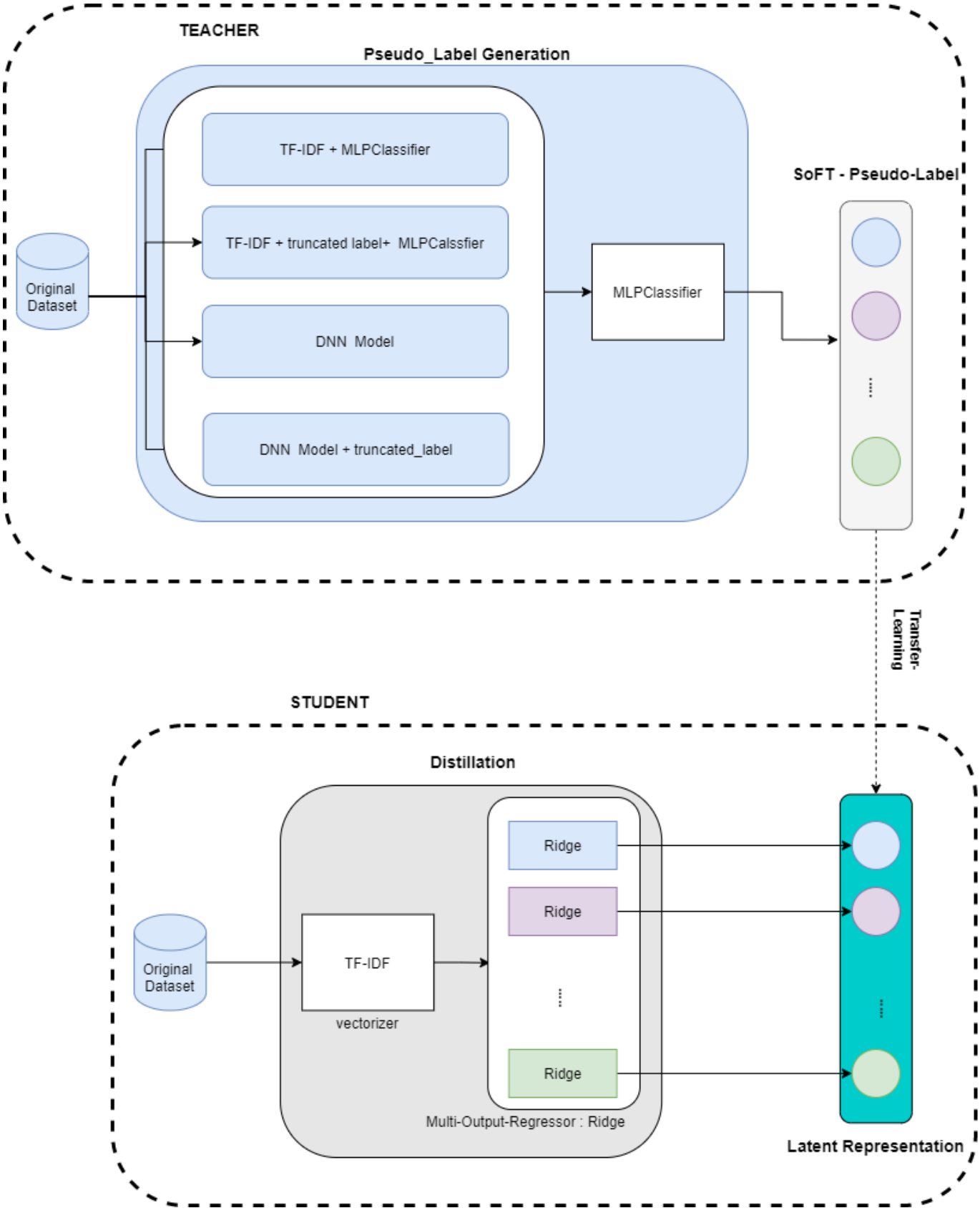

Figure 2: Pre-training of the vectorizer

4. What else can we do

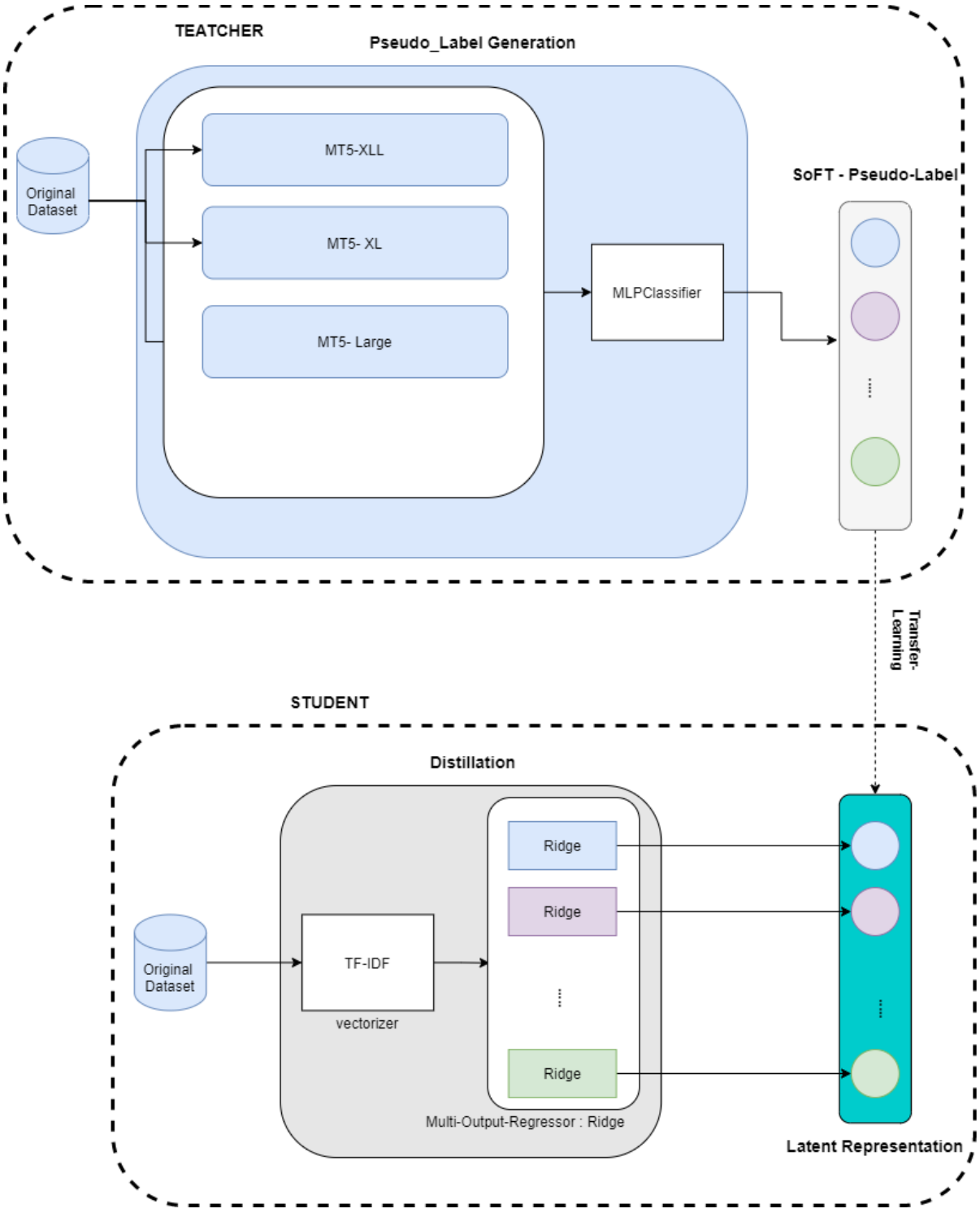

Figure 3: Pre-training of the vectorizer MT5

What we could have done, but couldn't. Actually, we have noticed that using a TF-idf combined with a Ridge model allows us to keep the performance. It would be interesting to train a MT5 model following the same strategy and get another one with a performance more or less equal to MT5 but is much lighter.