ML Olympiad

Toxic Language Detection (PTBR)

The problem

Main Goal

Classify Tweets as "Toxic" or "Non-Toxic". Toxicity is more related to semantics than syntactics, so we need to use a more robust model for this classification.

Linguistic Family

This solution is based on finetune. Therefore, we need to find a model capable of "understanding" comment data. Nothing better than using a model trained with Portuguese language data as a base.

Noises

Tweet comments tend to have a lot of noise. By noise we can consider: links, emoticons, hashtags and usernames.

Proposed Solution

Three-step solution

Step 1

Noise Remover

A function that removes any special characters found. The main ones were: usernames, emoticons, hashtags, links, retweet tags, characters from non-Latin alphabets.

Step 2

Foundation Model

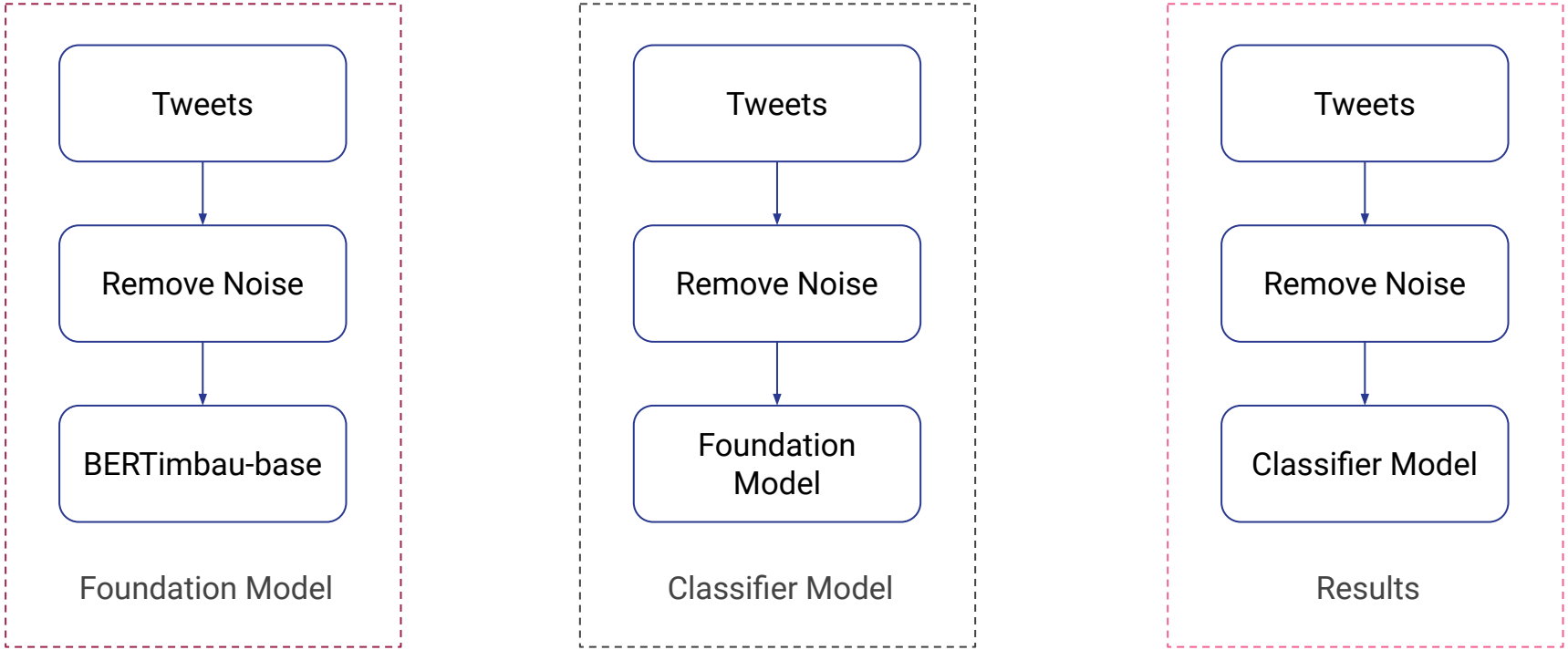
A model trained with Portuguese language data. Here we fine-tune the model with our data for the "Fill-Mask" task so that the model can better adapt to our data and understand the context of our information.

Step 3

Final Model

It uses the **Foundation Model** as a base, which in turn is fine-tuned with our data for the classification task.

Solution Diagram



Foundation Model

Model: BERTimbau-base
Parameters: 110M
Language: Portuguese
Type: Transformers
Tasks: Fill-Mask and Feature Extraction

Training

Epochs: 20
Learning Rate: 6e-5
Weight Decay: 0.01
Loss: Cross-Entropy
Training Loss: 1.76

Classification Model

Model: Foundation Model
Parameters: 110M
Language: Portuguese (+ Told BR)
Type: Transformers
Tasks: Fill-Mask and Feature Extraction

Training

Split: 0.9x Train – 0.1x Validation
Task: Classification
Max Steps: 1000
Learning Rate: 1e-5
Weight Decay: 0.01
Loss Function: Binary Cross-Entropy
Loss:

- Train: 0.516
- Validation: 0.480

Metrics:

- Validation Accuracy: 0.781

Results

Scores

- **Private:** 0.77380
- **Public:** 0.77666

Obs: I obtained the same results using the same approach for two different models.

- BERTimbau-base
- BERTweetBR

Scripts

Scripts

- **Github:** https://github.com/lrdsouza/told-br-classifier/