Proposal for Exist @ IberLEF 2022

march, 2022

At Exist@IberLEF 2022, the idea to evaluate is the comparison of two strategies for prediction assembling. The first one is to combine the prediction made by state of the art and finetuned language models (i) in different languages (i.e., es, en, fr, de, pt, it). The second strategy is to make prediction augmentation for each example by making back-translation and then combine the prediction over these paraphrases (ii).

For both strategies we employ RoBERTa-based pretrained language models (PLM) and introduce the data form MAMI, which is almost aligned with the current task, HAHA 2021, by analyzing which of the tweets related to targets like woman, men and their combination with other elements such as profession, lgbt or age, are or not sexist. Finally we include the data from HAHACKATHON, by just taking positive examples among the offensive ones and which contain wildcards such as femeini*, bitch, woman.

In (i) we evaluate how the ensemble of models pretrained in different languages and tuned on translated data, ranks with respect to a more data-centric apporach which consist in augmenting the data with examples provided in MAMI, HAHACKATHON and HAHA making also translation in such cases when it is needed in order to finetune two simple Language Models, one for english language and another for spanish.
This will allows us to conclude if for such tasks where detecting sexist messages is needed, incorporating more criteria given by the prediction on different languages could be more effective than just augmenting the data in order to give to the PLM a more general point of view of this issue.
On the other hand, these criteria taken into account for predicting on the ensemble, will be compared with (ii), where the criteria is expanded by making backtransaltion at testing time.

The prediction assembling in (i) will be evaluated through a simple majority vote and by taking prediction probability from each prediction source to employing a metaclasifier.

Regarding preliminary results obtained by the evaluation on the test set proposed for EXIST 2021 the strategy with best performance was (i) for both tasks 1 (accuracy: 0.7885, f1: 0.7868) and 2 (accuracy: 0.6532, f1: 0.5719), combining the prediction of LM in (es, en, fr, it, pt), whereas the criterion augmentation at testing time of (ii) obtained (accuracy: 0.7612, f1: 0.7609) for task 1 and (accuracy: 0.6163, f1: 0.5197) for 2. Finally the alternative for (i) to combine the predictions into a vector and separating their representation by a SVM achieved (accuracy: 0.7756, f1: 0.7752) for task 1 and (accuracy: 0.5911, f1: 0.5252) for 2.

Considering these results and in (i) we simply take the input sequence into different domains where its corresponding model is trained on, and in (ii) we paraphrase the input sequence by making the semantic information flowing towards a pivot language and afterwards back to the original language we may hypothesize:

a. For (ii) some phrases can be contrastive among their different translations in terms of the sexism perception of the evaluated Language Model. This possibly caused when the text flowed from the pivot language back to the original language the performance of the prediction decreased, i.e., went down from (accuracy: 0.7672, f1: 0.7672) for task 1 and (accuracy: 0.6220, f1: 0.5197) for 2.

b. Taking the texts into different domains and employing specific LM for each domain helped smoothing the noise produced for an individual model in this sexism detection task.

With respect to the best ranked team in the last year competition our approach introduces new data from sexism-related task and extends the use of a wider range of languages which allow us to outperform their result of 0.780 of acc and makes our model functional not just for english or spanish languages, but also with (fr, de, pt, it). We evaluated denoising the predictions not just by ensembling multiple language models but introducing back-translation on testing-time, which allow us to conclude the existence of vanishing in te sexism perception by the language model when the phrases are taken into another domain where the LM are not trained in.