NYU Center for Data Science

META-LEARNING FOR CROSS-LINGUAL COVID-19 FAKE NEWS DETECTION IN LOW-RESOURCE LANGUAGES

Group Members: Giacomo Bugli, Luigi Noto, Anirudh Nistala

Advisor: Will Merrill

INTRODUCTION

As we went through the COVID-19 pandemic, serious questions arose surrounding the credibility of news about medical and political information, posing the need for frameworks able to detect and flag unreliable articles before their spreading. Deep Learning models for text classification are notorious for being data- and computation-intensive, making the great majority of languages in the world under-resourced for the successful application of such methods.

To overcome the limitations imposed by these approaches, a solution to the problem of automatic fake-news detection in low-resource languages could be resorting to meta-learning. We thus decided to replicate the paper by van der Heijden et al. (2021), proposing a meta-learning framework for few-shot cross-lingual adaptation and multilingual joint-learning for document classification tasks in different domains.

Motivated by the substantial improvements and state-of-the-art achieved through the experiments on document classification in limited-resource settings, we propose to apply this approach for few-shot cross-lingual COVID-19 fake news detection in low-resource languages.

META-LEARNING

Meta-learning refers to the idea of using previous knowledge experiences to guide efficient new tasks learning (learning to learn) by optimizing performance on the task distribution.

Each learning task (e.g. each language) is associated with a dataset $\mathcal{D}_i$ containing both features and labels, and is split into a support set (used for fast adaptation) and a query set (used to evaluate performance after adaptation).

In the meta-training part, the goal is to optimize for the best performance on the task distribution. So given training tasks $\mathcal{T}_i$ with labelled datasets $\mathcal{D}_i$, find a common meta-parameter (often model parameter initialization) that plays a role in the task distribution

$$\omega^* = \arg\min_{\omega} \sum_i \mathcal{L}_\omega(\mathcal{D}_i)$$

In the meta-testing part, given an unseen target task $\mathcal{T}_j$, the goal is to use the learned meta-knowledge ω^* to obtain optimal task parameters with few support samples

$$\theta_j^* = \arg\min_{\theta} \mathcal{L}_\theta(\mathcal{D}_j \mid \omega^*)$$

The meta-learning approaches considered in this project are MAML (Finn et al., 2017) and ProtoMAML (Triantafillou et al., 2020).

DATASETS AND METRICS

Amazon Sentiment Analysis Dataset used in the paper replication, consisting of a collection of product reviews in three different categories in English, French, German and Japanese. For each language we concatenate the product categories and distinguish between positive (rating > 3) and negative (rating < 3) reviews obtaining 12 000 samples per language. We augmented it by adding 22 000 Chinese product reviews from JD.com

MM-COVID Dataset used in the extension, consisting of a collection of fake news content in English, Spanish, Portuguese, Hindi, French, and Italian. Please refer to Table 1 for the stats about MM-COVID.

Label	en	es	pt	fr	hi	it
Fake	1847	564	293	147	260	81
Real	4749	1830	637	246	1205	937

Table 1. Statistics of COVID-MM after pre-processing

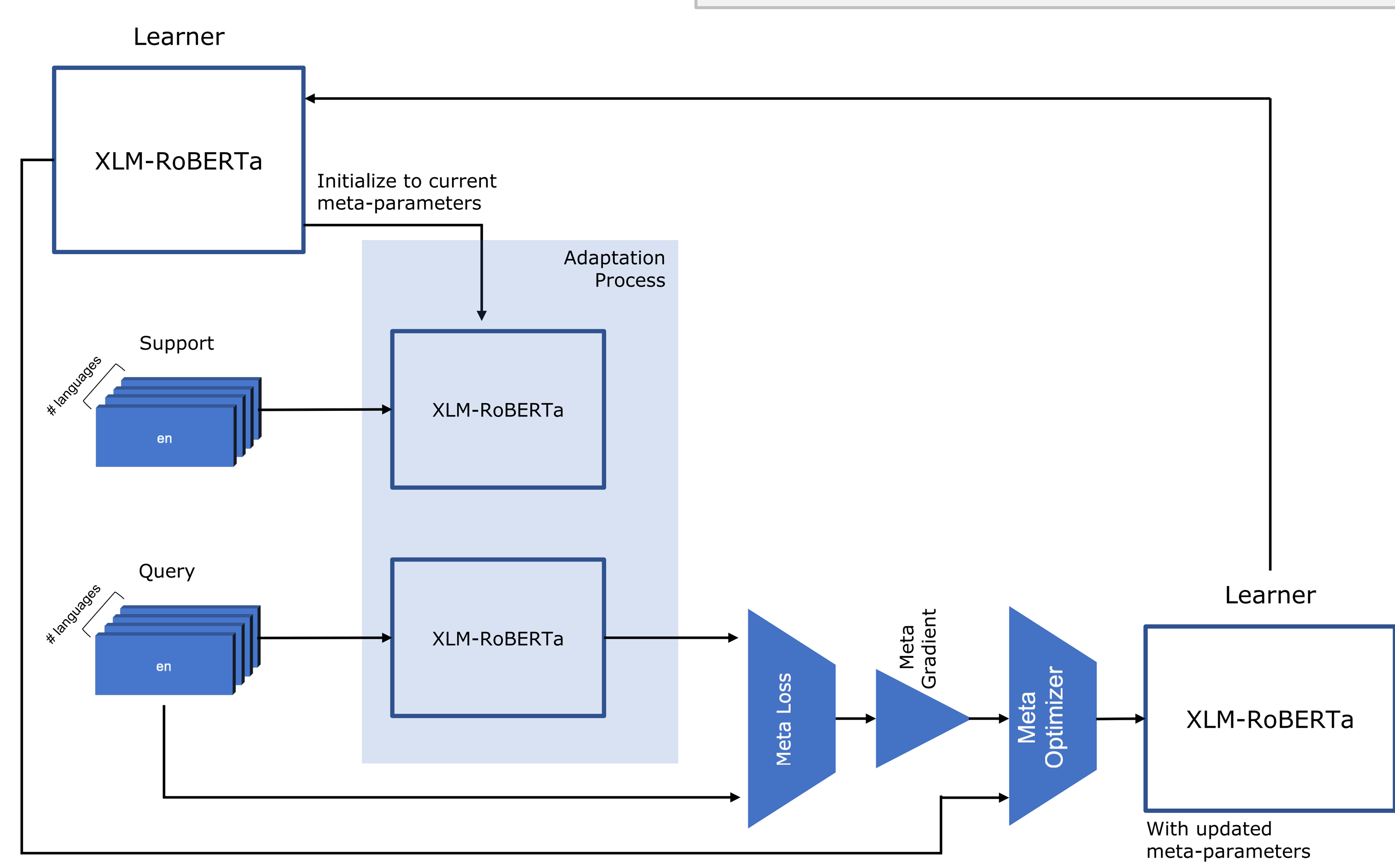

Figure 1. Illustration of the meta-training process on different languages (tasks) with XLM-RoBERTa as base-learner. Adapted from Cloudera Fast Forward Blog.

Method	Replication					Original Paper				
	de	fr	ja	zh	Δ	de	fr	ja	zh	Δ
Zero-Shot	87.4	87.3	83.7	79.8	84.6	91.2	90.7	87.0	84.6	88.4
Non-Episodic	89.1	86.5	84.1	80.3	85.0	91.6	91.0	85.5	87.9	89.0
foMAML	89.6	90.4	84.9	86.0	87.7	91.4	92.5	88.0	90.4	90.6
foProtoMAMLn	90.2	90.7	86.8	87.2	88.7	92.0	93.1	88.6	89.8	90.9

Table 2. Average accuracy of 5 different seeds on the unseen target languages for Amazon. Δ corresponds to the average accuracy across test languages.

METHODS AND EXPERIMENTS

Two meta-learning approaches have been considered:

- First-order approximation of MAML (foMAML)
- First-order approximation of ProtoMAML with L_2-normalized prototypes (foProtoMAMLn)

In all cases the base-learner is XLM-RoBERTa.

Paper Replication:
We replicate experiments in cross-lingual document classification (i.e. target tasks kept unseen until meta-testing) and multi-lingual document classification (target tasks are seen in meta-training) settings on the Amazon Sentiment Analysis dataset.
We compare the performances of foProtoMAML against three baselines (see Table 2).

Extension:
Using the same framework developed for the replication, we compare the performances of foProtoMAMLn against a baseline without meta-learning proposed by Li et al. (2020), dEFEND\C (see Table 3).

Method	Metric	pt	fr	hi	it	Δ
dEFEND\C	Accuracy	75.0	84.0	79.0	85.0	77.2
	F1-score	75.0	84.0	78.0	85.0	76.6
foProtoMAMLn	Accuracy	92.7	92.9	82.9	90.6	89.8
	F1-score	94.8	94.3	88.5	94.6	93.1

Table 3. Average accuracy and F1-score of 5 different seeds on the target languages for MM-COVID. Note that Li et al. (2020) trained dEFEND\C in a multilingual joint-training setting (with additional language dataset).

RESULTS

From Table 2 and Table 3 we can see that in both cases the meta-learning approaches outperform the baselines without meta-learning.

As regards the COVID-19 fake news detection application, foProtoMAMLn gains 12 percentage points in accuracy on average on the unseen target languages. In particular, we obtained a good performance increase on a low-resource language such as Hindi.

The improvements obtained suggest that meta-learning is effective in improving the model generalization capabilities to unseen tasks.

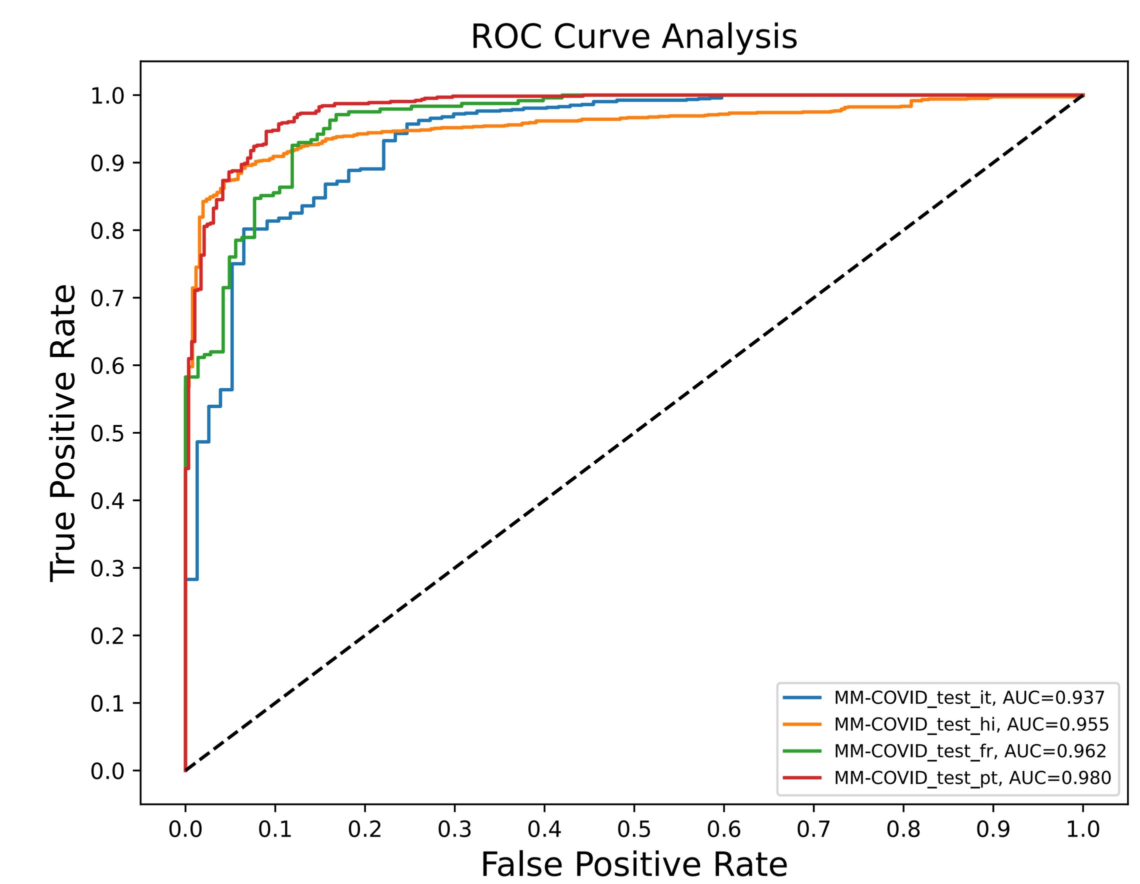

Figure 2. ROC curve in meta-testing for each unseen target language in the cross-lingual setting for MM-COVID. The predictions used to produce the curve are the ones obtained from the last of the 5 different seeds used to compute the test metrics.

REFERENCES

Y. Bengio, S. Bengio, and J. Cloutier. Learning a synaptic learning rule. In IJCNN-91-Seattle International Joint Conference on Neural Networks, volume ii, pages 969 vol.2–, 1991.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. CoRR, abs/1703.03400, 2017.

Yichuan Li, Bohan Jiang, Kai Shu, and Huan Liu. MM-COVID: A multilingual and multimodal data repository for combating COVID-19 disinformation. CoRR, abs/2011.04088, 2020.

Peter Prettenhofer and Benno Stein. Cross-language text classification using structural correspondence learning. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 1118–1127, Uppsala, Sweden, July 2010.

Eleni Triantafillou, Tyler Zhu, Vincent Dumoulin, Pascal Lamblin, Utku Evci, Kelvin Xu, Ross Goroshin, Carles Gelada, Kevin Swersky, Pierre-Antoine Manzagol, and Hugo Larochelle. Meta-dataset: A dataset of datasets for learning to learn from few examples. In International Conference on Learning Representations, 2020.

Niels van der Heijden, Helen Yannakoudakis, Pushkar Mishra, and Ekaterina Shutova. Mutilingual and cross-lingual document classification: A meta-learning approach. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 1966–1976, Online, April 2021.