IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received December 15, 2021, accepted January 19, 2022, date of publication January 28, 2022, date of current version February 10, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3147588

Cross-Lingual Few-Shot Hate Speech and Offensive Language Detection Using Meta Learning

MARZIEH MOZAFARI, (Student Member, IEEE), REZA FARAHBAKHSH, (Member, IEEE), AND NOEL CRESPI, (Member, IEEE)
CNRS Lab UMR5157, Institut Polytechnique de Paris, 91764 Palaiseau, France

Corresponding author: Marzieh Mozafari (marzieh.mozafari@telecom-sudparis.eu)

ABSTRACT Automatic detection of abusive online content such as hate speech, offensive language, threats, etc. has become prevalent in social media, with multiple efforts dedicated to detecting this phenomenon in English. However, detecting hatred and abuse in low-resource languages is a non-trivial challenge. The lack of sufficient labeled data in low-resource languages and inconsistent generalization ability of transformer-based multilingual pre-trained language models for typologically diverse languages make these models inefficient in some cases. We propose a meta learning-based approach to study the problem of few-shot hate speech and offensive language detection in low-resource languages that will allow hateful or offensive content to be predicted by only observing a few labeled data items in a specific target language. We investigate the feasibility of applying a meta learning approach in cross-lingual few-shot hate speech detection by leveraging two meta learning models based on optimization-based and metric-based (MAML and Proto-MAML) methods. To the best of our knowledge, this is the first effort of this kind. To evaluate the performance of our approach, we consider hate speech and offensive language detection as two separate tasks and make two diverse collections of different publicly available datasets comprising 15 datasets across 8 languages for hate speech and 6 datasets across 6 languages for offensive language. Our experiments show that meta learning-based models outperform transfer learning-based models in a majority of cases, and that Proto-MAML is the best performing model, as it can quickly generalize and adapt to new languages with only a few labeled data points (generally, 16 samples per class yields an effective performance) to identify hateful or offensive content.

INDEX TERMS Hate speech, offensive language, few-shot learning, meta learning, transfer learning, XLM-RoBERTa, cross-lingual classification.

I. INTRODUCTION

The proliferation of social media platforms (e.g., Twitter, Facebook, and Instagram) changes the way people communicate with each other. According to the statistics in the DataReportal report,[1] there are 4.33 billion social media users around the world at the start of 2021. The great promise of these platforms is to provide a safe place for users to communicate their opinions and share information. However, concerns are growing that they enable abusive behaviors, e.g., threatening or harassing other users, cyberbullying, spreading hate speech, racial and sexual discrimination, as well.[2]

Given the high progression of online hate speech and its severe negative effects, institutions, social media platforms, and researchers have been trying to react as quickly as possible. The recent advancements in Natural Language Processing (NLP), Machine Learning (ML), and Deep Learning (DL) have enabled research communities to develop a variety of automatic hate speech detection methods [1]–[6], where, in general, hate speech is defined as any type of communication that is abusive, insulting, intimidating, harassing, and/or inciting violence or discrimination, disparaging a person or a vulnerable group based on some protected characteristics, e.g., gender, sexual orientation, religion, ethnicity, race, etc. Recently, the introduction of transformer-based models, most notably BERT (Bidirectional Encoder Representations from Transformers) [7], RoBERTa [8] and XLM-RoBERTa (XLM-R) [9], has led to the considerable use of these models in hate speech and offensive language

The associate editor coordinating the review of this manuscript and approving it for publication was Yiqi Liu.

[1] https://datareportal.com/social-media-users

[2] https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/

identification tasks with promising results; the proof of this is achieving competitive scores and topping the leaderboards in recent shared tasks HASOC [10] and OffensEval [11] by these models.

Although a major research has been dedicated to the automatic identification of hate speech and offensive language in English [1], [2], [4], [12], creating annotated corpora and analyzing hateful and offensive content on other languages such as Italian [13], Spanish [14], Danish [15], Turkish [16], Arabic [17], etc. have raised many concerns recently. Since multilingual social media foster their users to interact in their primary languages, it is essential to develop automatic technologies including hate speech and offensive language detection tools for low-resource languages to reduce the vulnerability among users with different languages, other than English. However, the lack of sufficient annotated data containing hatred, offense, and abuse for low-resource languages has made this problem far from being solved at scale. Indeed collecting and labeling data is a labor- and time-consuming work, and the complex, subjective, and implicit nature of hate speech makes the reliability of annotation process more difficult.

Regarding the aforementioned challenges, we investigate the problem of the limited availability of labeled training datasets in low-resource languages for hate speech detection by proposing a few-shot cross-lingual approach based on meta learning. Meta learning is an effective solution proposed for few-shot learning problems, in which we have a few labeled data for a target task, and it has shown a great performance in different computer vision tasks, such as classifying new image classes with a few available instances of that class [18], [19]. Recently, meta learning has raised attentions regarding few-shot learning problems in NLP tasks as well, where a diverse tasks with different numbers of labels across tasks were studied [20], [21]. However, to the best of our knowledge, this is the first attempt to investigate the feasibility of meta learning in cross-lingual hate speech detection in order to tackle the problem of low availability of labeled data. Here, we study two popular tasks, namely hate speech and offensive language detection, separately and try to transfer knowledge from resource-rich languages to a low-resource target language by leveraging a meta learning approach derived from two optimization-based and metric-based methods; Model-Agnostic Meta-Learning (MAML) [18] and Proto-MAML [22].

The primary contributions of this study are:

- It evaluates the feasibility of a meta learning approach in few-shot cross-lingual hate speech detection and demonstrates its effectiveness on different languages with a low-resource setting. Simple but effective modifications are applied on two existing meta learning methods (MAML and Proto-MAML) to accomplish this goal.
- The first large-scale analysis of few-shot cross-lingual hate speech and offensive language detection is realized by assessing the performance of meta learning-based models over transfer learning models (e.g., XLM-R) on two diverse collections of different publicly available corpora comprising 15 datasets across 8 languages for hate speech and 6 datasets across 6 languages for offensive language.
- An evaluation using a few-shot setting in which only k samples per class are available from a target language is performed. The experiments demonstrate the superiority of the meta learning approach to generalize quickly to a new language in our few-shot classification tasks in comparison to the transfer learning-based baselines.

II. RELATED WORK

Previous work in this area has focused on different aspects of hate speech, including, but not limited to: (1) its definition [1], [2] and typology [23]; (2) the data collection and annotation process [24]–[26]; (3) investigation of automatic machine learning and deep learning [3], [27], [28] classification models and their generalizations [29]; (4) investigation of the most effective features of hate speech classification [1], [30]; (5) the unintended bias(es) in datasets or classification models [29], [31]–[33]; and (6) some of the relevant ethical principles [34]. Here, we present a concise overview of monolingual, multilingual, and cross-lingual hate speech and offensive language detection models along with the few-shot learning problem in this domain.

A. HATE SPEECH AND OFFENSIVE LANGUAGE DETECTION

Abusive language emerging in a variety of forms, including hateful and offensive expressions, toxicity, misogyny, racism, sexism, cyberbullying, etc. is widely poisoning the online social media environment. A wide range of studies has therefore been dedicated to developing automatic methods to detect these types of content in social media. Our main focus here is on two of the most popular tasks, hate speech and offensive language detection in monolingual and multilingual settings.

The vast majority of studies have investigated the development of computational models for hate speech and offensive language identification tasks predominantly in a single language, English. These efforts have relied on simple feature engineering methods, including Bag-of-Words (BoW), Term Frequency–Inverse Document Frequency (TF-IDF), word-level and character-level n-grams [35] accompanying different traditional supervised classifiers such as Multinomial Naïve Bayes, SVM, and Random Forest [1], [2], [36]. Furthermore, the impacts of additional features such as syntactic and linguistic features, distributional semantics (word-embeddings), and user-based and platform-base metadata have been investigated [30], [37], [38].

Thanks to the advances in neural network models and the volume of available labeled data in this domain, mainly just in English, various neural network-based approaches such as Recurrent Neural Network (RNN) [4]), Long Short-Term Memory (LSTM) [27], Convolutional Neural Network (CNN) [39], bidirectional LSTM (BiLSTM) [40], Gated Recurrent Unit (GRU) [3] have been employed in hate speech content identification. Most of these approaches outperform traditional machine learning models. In addition, the recent advancements in language representation models such as ELMO [41], BERT [7], and XLM-R [9] have led to the considerable use of transformer-based pre-trained language models in hate speech and offensive language detection with competitive and promising results [42]–[44].

B. MULTILINGUAL AND CROSS-LINGUAL HATE SPEECH DETECTION

The multilingual nature of social media has underscored the importance of hate speech detection in multilingual settings. Several studies have investigated the multilingual classification of hate speech and offensive language using multilingual, cross-lingual, or joint-learning approaches. We summarize the works in multilingual (where the robustness of the proposed models is evaluated across multiple languages without experimenting in a cross-lingual setting) and cross lingual (where the proposed models are evaluated in a zero- or few-shot setting), below.

1) MULTILINGUAL

There has been a great interest in providing monolingual datasets of various languages other than English, such as Arabic [17], Danish [15], Turkish [16], Greek [45], Italian [13], French [46], Spanish [14], Dutch and German [47], Portuguese [48], Indonesian [49] and more. In addition, various competitions [10], [11], [14], [50], [51] and workshops [52], [53] have been dedicated to introducing new computational methods for the identification of hate speech, offensive language, cyberbullying, etc. in different languages.

Building multilingual classifiers to automatically detect hate speech is a very recent topic that would be a notable step forward in this area. Ousidhoum *et al.* [46] presented the first multilingual multi-aspect hate speech analysis dataset in English, French, and Arabic tweets and evaluated several multilingual multi-task learning approaches for the identification of hate in a multilingual setting. Ibrohim and Budi [54] investigated the effect of the machine translation approach in multilingual hate speech detection in Hindi, English, and Indonesian, by comparing classifiers trained with/without translating samples. Ranasinghe and Zampieri [44] employed a cross-lingual contextual word embeddings model, XLM-R, to transfer knowledge from a resource-rich language, English, to a low-resource language (i.e., Bengali, Hindi, or Spanish) to predict offensive content in less-resourced languages. First, they used XLM-R to train a classification model on English. Then, they used training data from Bengali, Hindi, and Spanish to fine-tune the model on target language. Corazza *et al.* [55] proposed a robust recurrent neural architecture to identify hate speech in different languages (i.e., English, German, and Italian), and also evaluated the effect of different type of embeddings, additional features (word-level, tweet-level, or emotion-based), and hashtag and emoji normalization in the architecture’ s performance. Vashistha and Zubiaga [56] proposed a hierarchical deep neural network for the identification of hate speech in English, Hindi, and Hindi code-mix language to investigate the effect of a combination of CNN filters or pre-trained BERT embedding into a BiLSTM model.

Multilingual Offensive Language Identification in Social Media (OffensEval-2020) [11] is a pioneering effort to analyze multilingual offensive language in social media by providing multilingual datasets in five languages: Arabic, Danish, English, Greek, and Turkish. Using the English dataset annotated with a three-level annotation scheme to identify offense content, the target audience and the type of offense, participants contributed in this task by a variety of traditional machine learning and deep neural network models. For the languages other than English, data was annotated in one level as either offensive or non-offensive content. More than half of the contributions associated pre-trained transformer-based models: BERT [7], mBERT [7], RoBERTa [8], XLM-R [9], ALBERT [57], etc. with fine-tuning and data-augmentation strategies to tackle the problem of offensive language detection. Wang *et al.* [58] proposed a multi-lingual method leveraging the transformer-based pre-trained model XLM-R and ERNIE to predict offensive language and its target and type. Wiedemann *et al.* [59] performed an exhaustive experimental evaluation using different transformer models such as BERT-base and BERT-large, RoBERTa-base and RoBERTa-large, XLM-R, and different version of the ALBERT model to fine-tune the models on English offensive language data and found that using an ensemble combining different ALBERT models outperforms other models.

2) CROSS-LINGUAL

The cross-lingual setting in which there are few or non-existent training data sets in the target language is a relatively new concept in the hate speech detection domain. Some recent works have discussed the use of cross-lingual models, along with few- or zero-shot learning methods for addressing the problem of hate speech identification across different languages. Stappen *et al.* [60] proposed an architecture for uni-lingual and cross-lingual zero- and few-shot hate speech detection from English to Spanish and vice versa. Their system used a frozen transformer language model, BERT or XLM, to extract the contextual representation of input samples without fine-tuning the models. Their next step utilized an attention-based classification block, Attention-Maximum-Average Pooling (AXEL), as a trainable layer to condense hate speech specific representations from general text representations of BERT or XLM. Aluru *et al.* [5] analyzed hate speech in a multilingual setting by considering 9 languages from 16 publicly-available hate-speech datasets on Facebook and Twitter. In a few-shot setting, they considered datasets of $n-1$ languages as training and an nth language as the unseen target language (test) to train models based on multilingual embedding models LASER and BERT, using an incremental approach to include target language samples in the training process. Pamungkas *et al.* [6]

employed a machine translation mechanism and proposed two joint-learning architectures based on a multilingual pre-trained model called MUSE (Multilingual Unsupervised and Supervised Embeddings) with an LSTM network and a multilingual pre-trained BERT model to identify hate content among 11 publicly available datasets in 7 different languages. To configure a zero-shot setting, these researchers considered English as the training set and other languages as the test sets. Although this model has yielded a cross-domain robust system, there is a limitation attributed to potential excessive data noise which is produced during the translation and is propagated to downward learning modules. Therefore, we do not use translation mechanism in our proposed few-shot learning model.

C. FEW-SHOT META-LEARNING IN NLP

Establishing ways to classify inputs based on only a limited number of samples, known as few-shot learning, has attracted much attention in the research community. One of the most popular solutions for few-shot learning is meta learning, or learning-to-learn, mainly used in the computer vision area [18], [19]. Meta learning has also become popular recently for few-shot learning problems in NLP. Gu *et al.* [20] introduced a MAML-based meta learning method for low-resource neural machine translation by exploiting large samples of high-resource languages pairs to learn how to adapt to target languages. They considered 18 high-resource language translation tasks as training tasks, and five low-resource ones as testing tasks.

Regarding multiple-tasks and monolingual settings, Dou *et al.* [61] explored multiple MAML-based approaches for low-resource Natural Language Understanding (NLU) tasks on the General Language Understanding Evaluation (GLUE) benchmark, but only for English. They used the four high-resource tasks SST-2, QQP, MNLI, and QNLI as the training tasks and the low-resource tasks CoLA, MRPC, STS-B, and RTE as the testing tasks. Their results indicated the superiority of meta learning approaches to the fine-tuned BERT and multi-task learning approaches. Bansal *et al.* [21] introduced a new MAML-based meta learning model to perform few-shot learning across 17 NLP tasks with different numbers of classes. Their results for k-shot learning (k = 4, 8, 16) indicated the superiority of meta-learning to the BERT-based transfer learning models. To learn the interactions between tasks and languages in a meta-learning setting, Nooralahzadeh *et al.* [62] studied a cross-lingual meta-learning method based on MAML for few- and zero-shot learning in Natural Language Inference and Question Answering tasks by pre-training on a high-resource language, English, meta-learning using low-resource languages, auxiliary languages, and zero-shot or few-shot learning on the target languages. Meanwhile, Tarunesh *et al.* [63] proposed a meta-learning model to more effectively share parameters across multi tasks and languages by experimenting on five different tasks and six different languages from the XTREME multilingual benchmark dataset. Sui *et al.* [64] proposed a few-shot meta learning approach for sentiment classification task on the Amazon reviews dataset for 23 types of products for which three different binary classification tasks exist. They used a 5-shot meta learning model to classify the sentiment of 12 tasks from four domains (Books, DVDs, Electronics, and Kitchen).

Based on our literature review, the potential of applying meta learning algorithms to address the problem of few-shot learning in cross-lingual hate speech and offensive language detection tasks has not yet been thoroughly explored. Furthermore, no study has been devoted to investigating cross-lingual hate speech and offensive language detection as two separate tasks with large-scale datasets. Our approach could thus be the first step towards creating a benchmark dataset in hate speech detection similar to other NLP tasks (e.g., GLUE).

III. METHODOLOGY

In this section, we introduce the terminology and definitions related to few-shot learning and meta learning and describe the adaptation of few-shot learning concepts and meta learning approaches to our cross-lingual problem.

A. FEW-SHOT LEARNING IN CROSS-LINGUAL HATE SPEECH

Deep neural networks' requirement of large amounts of training data to achieve promising results makes these models inefficient when there is a lack of training data. Meanwhile, hate speech and offensive language are a common phenomenon in social media that does not respect language barriers, so the lack of sufficient labeled data in some languages, mainly low-resource ones, renders automatic detection algorithms impractical. Meta learning is thus a potential answer to this training data lacunae.

In this setting, we have a dataset including labeled samples in different languages. We formulate our cross-lingual problem for each target language as an N-way K-shot classification, given:

1) A support set composed of K labeled samples per each N classes for a target language; and
2) A query set composed of Q unlabeled samples of a target language.

where we aim to classify the Q unlabeled samples of target language into the N classes given the $N \times K$ labeled samples in the support set during the training. Given the insufficient training data that we have in each target language (K samples per each of N classes), we characterize our few-shot learning problem as a meta learning problem in which training on other languages helps to achieve better results in a target language. Since we consider our problem a binary classification problem (hate/non-hate or offensive/non-offensive), the number of classes N is fixed at two. Our assumption here is that all the K samples in each target language are new.

B. META LEARNING

Meta learning, or learning to learn, is a general paradigm for few-shot learning that learns to quickly adapt to new tasks. Given a classification problem, classical learning algorithms learn how to classify from the training data, and evaluate the performance of a task using test data. However, a meta learning algorithm learns to learn on a diverse set of training tasks and then evaluate new tasks at test time [65].

We consider a model f parameterized by θ to map each training sample with input vector x to output label y; f is often referred to as the *base-learner*. In a meta learning scenario, the model is trained to learn to adapt to a large number of tasks. Therefore, we assume a set of M related tasks in our formulation as $\mathcal{T} = \{\tau_1, \tau_2, \ldots, \tau_M\}$ with a distribution over tasks $\tau_i \sim p(\mathcal{T})$, where each task potentially has a large amount of training data $\mathcal{D}_i \in \mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_M\}$, containing feature vectors and ground truth labels $\mathcal{D}_i = \{(\mathbf{x}_i, y_i)\}$. Each $\mathcal{D}_i$ is divided into a training set $\mathcal{D}_i^{train}$ (or support set) to adjust the model parameters to the specific task and a test set $\mathcal{D}_i^{test}$ (or query set) to evaluate the performance, denoted as $\mathcal{D}_i = \langle \mathcal{D}_i^{train}, \mathcal{D}_i^{test} \rangle$. In each meta-training step (i.e., an *episode*) a task τ_i is sampled from $p(\mathcal{T})$. Then, considering task τ_i as an N-way K-shot task, the model f is trained with K samples (per N classes) from $\mathcal{D}_i^{train}$ using feedback from the corresponding loss function $\mathcal{L}_i$ from τ_i and evaluated on $\mathcal{D}_i^{test}$ to compute a loss with respect to the model' s parameter initialization. The loss on $\mathcal{D}_i^{test}$ is used to adjust the model parameters. The validation error of the sampled tasks τ_i serves as the training error of the meta learning process in which updating the parameters of the base-learner (f_θ) continues by performing the described episodic training process until some stop criteria is reached. Finally, to generalize the model on a new task $\tau_{M+1} \sim p(\mathcal{T})$, the model uses its learning procedure to adapt to the task τ_{M+1} with only K samples per class of its train set.

An overview of our cross-lingual meta learning-based framework is depicted in Figure 1. Using a few-shot learning fashion, we use a diverse set of tasks in different languages to train a model in the meta-training step, and then in the meta-testing step the model is further fine-tuned in only k labeled samples from the unseen target language. Each training task corresponds to a language with labeled instances respecting Support and Query sets. During the training, the base-learner (a cross-lingual pre-trained language model) uses Support sets from different languages to train its parameters. Then, Query sets are used to calculate loss and update the parameters. During the test, the base-learner uses the Support set of an unseen target language to adapt to the new language and evaluates the performance on unlabeled Query set of it.

There are three different approaches for performing meta learning: metric-based [66], [67], model-based [68], and optimization-based [18], [69]. Metric-based methods learn similarities between feature representations of input samples from different training sets given a similarity metric such as cosine similarity or Euclidean distance, to calculate how close two samples are in the space. Model-based methods learn to update their parameters and incorporate new information rapidly with a few training steps by leveraging an internal storage buffer (e.g., memory networks) or another meta-learner model. Optimization-based methods try to find a good point of parameter initialization across tasks and adapt to new tasks with a few steps of gradient descent. In this study, we propose to use an optimization-based meta learning algorithm, Model-Agnostic Meta-Learning (MAML), for our few-shot classification task due to its superior performance at several computer vision [18] and NLP tasks [21], [64]. In addition, an adaptation of the MAML method, Proto-MAML [22] is also investigated. In the following sections, we introduce the respective characteristics of these algorithms.

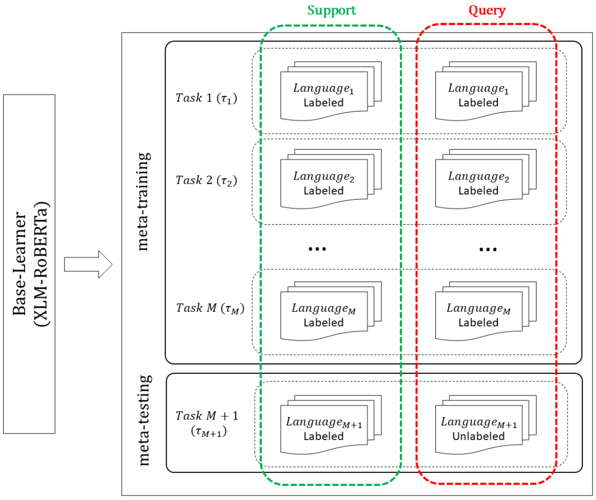


FIGURE 1. An overview of the cross-lingual meta learning-based framework for few-shot hate speech classification task.

1) MODEL-AGNOSTIC META-LEARNING

The idea of MAML is to perform meta learning by finding a good initialization of parameters through multiple tasks and then quickly adapting to new tasks with relatively few training samples [18]. Considering a model represented by a parametrized function f_θ with parameters θ, in general, for a single training dataset for one task, neural network parameters are randomly initialized and optimized via gradient descent; however, MAML extends the gradient descent by optimizing parameters θ to yield good performance on a set of related tasks $\mathcal{T} = \{\tau_1, \tau_2, \ldots, \tau_M\}$.

Given a sampled task τ_i from the distribution $p(\mathcal{T})$, the parameters θ of model f for τ_i are updated to θ_i' using one or a few gradient descent steps on the $\mathcal{D}_i^{train}$ of task τ_i, as follows:

$$\theta_i' = \theta - \alpha \nabla_\theta \mathcal{L}_{\tau_i}(f_\theta) \tag{1}$$

where α is the step size (learning rate), f_θ is the learned model, and $\mathcal{L}_{\tau_i}$ is the loss on the specific test set $\mathcal{D}_i^{test}$ of task τ_i. The model parameters θ are trained to optimize the performance of the base-learner $f_{\theta_i'}$ on the unseen test examples $\mathcal{D}_i^{test}$ in order to generalize on the specific task τ_i. This step is known as inner-loop optimization. Considering a distribution of tasks $p(\mathcal{T})$ the meta learning objective is:

$$min \sum_{\tau_i \sim p(\mathcal{T})} \mathcal{L}_{\tau_i}(f_{\theta_i'}) = \sum_{\tau_i \sim p(\mathcal{T})} \mathcal{L}_{\tau_i}(f_{\theta - \alpha \nabla_\theta \mathcal{L}_{\tau_i}(f_\theta)}) \tag{2}$$

Finally, MAML performs meta-optimization, known as outer-loop optimization, across tasks via a stochastic gradient descent (SGD) as follows [18]:

$$\theta \leftarrow \theta - \beta \nabla_\theta \sum_{\tau_i \sim p(\mathcal{T})} \mathcal{L}_{\tau_i}(f_{\theta_i'}) \tag{3}$$

where β is the meta step size (learning rate) and the α and β may be fixed as hyper-parameters or be meta-learned.

Although MAML is an elegant and very powerful method that has produced state-of-the-art-results in different settings for computer vision tasks [18], it suffers from some drawbacks such as instability during training, limitations on the model generalizability, high computational requirements in both training and inference times, and being costly in terms of hyperparameter tuning. To address these disadvantages, Antoniou *et al.* [70] proposed various modifications to MAML that stabilize the system as well as improve the generalization performance, convergence speed and computational efficiency. Following the work of *Antoniou et al.* [70], we adapt some modifications to our MAML-base few-shot learning model as follows:

Regarding Equation 3, optimizations through gradient update steps in MAML require computing second-order derivatives, which is very expensive. One possible solution is to compute only the first-order approximation of the gradient derivatives to speed up the process, however, this can have a negative impact on the final generalization error [70]. Therefore, we use a derivative-order annealing approach in which in the early steps of training the first-order gradients are computed to speed up the training process, and then in the later training steps the second-order gradients are computed to improve the generalization performance.

Regarding Equation 2, the learning rate α is shared across all update steps and all parameters, which results in a high computational cost for finding the correct learning rate for a specific task. Instead, we use an initial learning rate per layer and per step to be jointly learned during the meta-learning steps of MAML. Furthermore, MAML uses a fixed step size β to optimize its meta-objective in Equation 3 with an Adam optimizer, which results in both generalization performance and computational costs issues. We anneal this learning rate on the optimizer by utilizing a cosine annealing function proposed by Loshchilov and Hutter [71], to achieve higher generalization performance. Although we have made these modifications to the original MAML method, for simplicity we use MAML to refer to this model in this study.

2) PROTO-MAML

the Prototypical Network algorithm proposed by Snell *et al.* [67] is one of the more successful metric-based approaches in meta learning and has yielded substantial improvements in the few-shot learning problem. This approach hypothesizes the existence of an embedding (a prototypical representation) in which all the samples belonging to a specific class cluster around a single prototype representation for that class. Then, a new sample in few-shot learning is classified based on its distance with the prototypical representation of each class. Therefore, this metric-based algorithm requires an embedding function f_θ to extract the embeddings of all samples, and a distance function d to compute the distance between new samples and the prototypical representation of each class. Given an embedding function f_θ and a few-shot classification with support set S and query set Q, the embeddings of all samples in S are encoded by f_θ, and then a prototype $\mathbf{c}_k$ is computed for each class k in S by taking the mean embeddings of samples of the respective class as follows:

$$\mathbf{c}_k = \frac{1}{|S_k|} \sum_{(\mathbf{x}_i, y_i) \in S_k} f_\theta(\mathbf{x}_i) \tag{4}$$

where $\mathbf{c}_k$ is the prototype of class k. Given the prototypes of the classes in S, each unlabeled sample in Q is encoded by f_θ and is then classified by:

$$p(y = k|\mathbf{x}) = \frac{exp(-d(f_\theta(\mathbf{x}), \mathbf{c}_k))}{\sum_{k'} exp(-d(f_\theta(\mathbf{x}), \mathbf{c}_{k'}))} \tag{5}$$

where d is a distance function, mainly the squared Euclidean distance. To obtain the probability distribution over classes, a Softmax function with a negative log-likelihood loss function is applied on the distance vectors and then the sample is assigned to the class with the highest probability value.

Triantafillou *et al.* [22] utilized the simple inductive bias of Prototypical Network algorithms along with the simple and flexible adaptation mechanism of MAML to introduce a new meta learning algorithm, called Proto-MAML. Taking a Softmax on the squared Euclidean distance function between f_θ and the prototypes of the classes $\mathbf{c}_k$ in Prototypical Networks makes them a linear model [67], where the equivalent linear layer has weights $\mathbf{W}_k$ and biases b_k corresponding to class k which are computed as follows (regarding Equation 4):

$$\mathbf{W}_k := 2\mathbf{c}_k \quad \text{and} \quad b_k := -\left\|\mathbf{c}_k\right\|^2 \tag{6}$$

Therefore, Proto-MAML adapts MAML in such a way that these weights and biases can be employed in the task-specific linear layer of each episode in the MAML instead of using random initializations. This simple modification yields significant improvements in the optimization process of MAML [22]. We adapt this strategy in MAML and use Proto-MAML as a meta learning algorithm in our cross-lingual few-shot classification problem. We apply all the modifications which are used in MAML in Proto-MAML as well. Although we make some modifications to the original Proto-MAML method, we use the same name to refer to this model for simplicity.

TABLE 1. Dataset description for hate speech and offensive language detection tasks. Class 0 and Class 1 represent the number of normal and hate/offensive labels in the datasets, respectively. The last column indicates the total number of samples for each language.

Task	Language	Dataset	Class 0	Class 1	#Samples per language
Hate Speech	English	Davidson *et al.* [2]	4,163	1,430	79,348
		Basile *et al.* [14]	7,530	5,470	
		Founta *et al.* [26]	53,851	4,965	
		Ousidhoum *et al.* [46]	661	1,278	
	Arabic	Ousidhoum *et al.* [46] Mulki *et al.* [72]	915 3,650	755 468	5,788
	Spanish	Basile *et al.* [14]	3,861	2,739	12,600
		Pereira-Kohatsu *et al.* [73]	4,433	1,567	
	German	Roß *et al.* [25]	315	54	4,648
		Mandl *et al.* [10]	4,126	152	
	Indonesian	Ibrohim and Budi [49]	7,608	5,561	13,882
		Alfina *et al.* [74]	453	260	
	Italian	Bosco *et al.* [13]	2,704	1,296	4,000
	French	Ousidhoum *et al.* [46]	821	399	1,220
	Portuguese	Fortuna *et al.* [48]	3,882	1,788	5,670
Offensive Language	English	Zampieri *et al.* [12]	9,460	4,640	14,100
	Arabic	Mubarak *et al.* [17]	8,085	1,915	10,000
	Danish	Sigurbergsson and Derczynski [15]	3,159	441	3,600
	Turkish	Çöltekin [16]	28,464	6,847	35,284
	Greek	Pitenis *et al.* [45]	7,376	2,911	10,287
	Persian		4,376	1,624	6,000
Total			159,893	46,560	206,453

C. BASE-LEARNER MODEL

Since the optimization-based meta learning algorithms used in this study are model-agnostic, they are compatible with any base-learner model that learns through gradient descent. Here, we chose the cross-lingual pre-trained language model XLM-R [9] as the base-learner for hate speech and offensive language classification tasks. The base-learner extracts the last hidden-state layer of the first token of the sequence (the classification token) in size 768 and processes it by a linear layer and a *tanh* activation function to do the classification.

IV. DATASET DESCRIPTION

Given the varying definition of hate speech and offensive language content in publicly available datasets, we decided against the combination of datasets with hatred and offense samples. Hence, we consider two separate tasks, *hate speech detection* and *offensive language detection*, with different datasets in different languages. The hateful datasets consist of insults targeted toward a group based on some protected characteristics such as sexual orientation, religion, misogyny, nationality, gender, ethnicity, etc., whereas offensive datasets contain any form of non-acceptable language or a targeted offense including insults, threats, and posts containing profane language or swear words [11].

A. HATE SPEECH DATA

We use 15 publicly available sources in 8 languages provided by the research community. Most of the datasets are selected according to the *hatespeechdata*[3] web page that catalogs datasets annotated for hate speech, online abuse, and offensive language. Although different datasets have different classes, in this study, we only select samples including hateful or normal content. The details regarding all hateful datasets are included in appendix A.

B. OFFENSIVE LANGUAGE DATA

We use the original multilingual offensive language dataset provided in OffensEval-2020 [11], a shared task at SemEval-2020, which focused on multilingual offensive language identification in 5 languages Arabic, Danish, English, Greek, and Turkish. All of the languages followed the OLID annotation schema proposed by Zampieri *et al.* [12], and had only offensive and non-offensive classes in the first level of annotation. In addition, we use a Persian offensive language dataset collected and annotated by us following the annotation guidelines at [12]. The details regarding all offensive datasets are included in appendix B.

The statistics of these datasets are presented in Table 1 where the second column represents the datasets from different languages in Hate Speech and Offensive Language categories and the third and fourth columns represent the number of normal (non-hateful or non-offensive) and hateful/offensive content as Class 0 and Class 1, respectively. The total number of samples in each language is reported in the final column.

V. EXPERIMENT AND RESULT

This section presents the details of different training models including the meta learning models and different baselines used in this study. In addition, it describes the experimental

[3] https://hatespeechdata.com

setup and implementation details as well as the results of our experiments.

A. TRAINING MODELS

1) MAML AND PROTO-MAML

In our cross-lingual few-shot classification task, we have a set of training, validation, and test tasks which are including samples from different languages (mutually exclusive). To investigate the performance of the meta learning approach in cross-lingual hate speech detection, we divide each dataset (hate speech or offensive language) into three meta-datasets: 1) a training set L_{train} comprising of training languages to train MAML; 2) a validation set L_{val} consisting of validation languages to tune MAML hyper-parameters; and 3) a test set L_{test} consisting of test (or target) language to evaluate generalization ability of the model on an unseen target language. Therefore, using labeled data in L_{train} the model is trained. Then, by using samples from L_{val}, we tune the hyper-parameters and set early stopping condition. As we consider a few-shot setting, we do not rely on a large validation set and use a held-out validation set of a specific language in validation set L_{val}. For evaluating the method on L_{test}, at first, we fine-tune the model using a sample of k-shot training data (k samples per label in target language 's train set) and then test the model on the entire test set of target language. Therefore, the target language is unknown during both training and model selection. All the settings in MAML and Proto-MAML are the same.

2) BASELINES

We create two transfer learning baselines (based on XLM-R model) to evaluate the ability of these approaches as well as our proposed model for cross-lingual few-shot hate speech and offensive language detection tasks. The baselines are as follows:

- **XLM-R** Aluru *et al.* [5] have recently proposed a multilingual BERT-based model for multilingual hate speech detection in which all samples in different languages except a target language l_{tgt} (test language) are used as training data and then the model is further fine-tuned with a portion of training data of l_{tgt} and evaluated in a held-out test set of l_{tgt}. Inspiring this study, we create a baseline for our few-shot cross-lingual model where we use the pre-trained model XLM-R with a two-step fine-tuning method. During the fine-tuning, first, the model is trained on all languages except l_{tgt} and the best model is selected according to the held-out validation set of l_{tgt}. Then the selected model is fine-tuned with only k samples (per class) in l_{tgt}. At the end, the model is evaluated on the test set of l_{tgt}. Samples from different languages in training, validation, and test steps of this model are considered as L_{train}, L_{val}, and L_{test}. We note that according to Aluru *et al.* [5], this model uses target language for both model selection and test step. Therefore, the target language will be unknown only during training phase.
- **Non-episodic** To measure the exact impact of meta learning on the performance of model versus standard supervised learning, we use a non-episodic approach to train a model in which support and query sets of training languages in L_{train} are merged, and by using a mini-batch gradient descent with cross-entropy loss function the model is trained. In the test step, first, the trained model is fine-tuned on k-shot training data of L_{test}, and then is evaluated on test set of L_{test}. The target language will be unknown during both training and model selection.

B. TRAINING SETUP AND IMPLEMENTATION

1) TRAINING SETUP

We consider hate speech and offensive language detection as two separate tasks in which a binary classification is trained based on transfer learning or meta learning approaches. To create and initialize each model, we use the configuration, tokenizer, and pre-trained weights of the XLM-R (xlm-roberta-base) model from publicly available Transformers[4] library for Pytorch (pytorch-transformers). Then, each model will be fine-tuned on the downstream task by adding a classification head on top of the pre-trained XLM-R encoder. As hate speech and offensive language detection are binary classification tasks, we directly modify and fine-tune the classification class of the XLM-R model (*XLMRobertaForSequenceClassifcation*).

For MAML-based meta leaning models, we consider 50 epochs and sample 100 training episodes per epoch to perform meta training. The learning rate of inner loop α (adaptation stage) and learning rate of outer loop β are set initially to 3e-5 and 6e-5, respectively. We use Adam optimizer to update the parameters. The number of update steps in the inner-loop is set to 10. During the first 30 epochs, we calculate the first-order derivatives and in the rest of training process we calculate the second-order derivatives in MAML. We perform evaluation on the samples in L_{val} set with 5 different seeds after each epoch, and to avoid overfitting, we apply early stopping when the validation accuracy failed to decrease for 5 epochs. In the few-shot setting, we chose $k \in \{4, 8, 16\}$ to evaluate how models generalize to new target language with a limited labeled data k per class.

For the XLM-R baseline, the maximum sequence length of the input sentences is set to 256 and in case the input length is shorter or longer, it will be padded with zero values or truncated to the maximum length, respectively. The model is fine-tuned with a batch size of 16 for 5 epochs. An Adam optimizer with a learning rate of 2e-5 is used to minimize the Cross-Entropy loss function. For non-episodic baseline, the model is trained for 5 epochs on L_{train} and is evaluated after each epoch on L_{val} set.

[4] https://github.com/huggingface/transformers

TABLE 2. Results of k-shot classification on the unseen target languages of hate speech dataset in terms of macro F1-measure with standard deviation. The values in bold indicate the best performing model in each k-shot setting. The last column corresponds to the row-wise average F1-measure across all target languages.

Models	k-shot	Target Languages							
		ar	de	es	fr	id	it	pt	avg
XLM-R [5]	4	42.32 ±0.91	37.90 ±1.41	46.06 ±1.37	46.23 ±0.62	45.74 ±1.65	37.22 ±2.11	41.67 ±0.73	42.44
	8	38.78 ±2.31	**46.13** ±0.81	39.77 ±1.78	**46.44** ±1.78	46.98 ±2.78	39.02 ±0.58	43.50±1.08	42.94
	16	43.01 ±1.32	50.23 ±2.01	45.64 ±0.82	52.32 ±2.24	49.86 ±0.87	45.22 ±1.41	51.03 ±2.82	48.18
Non-episodic	4	41.07 ±2.91	36.29 ±0.36	45.33 ±2.41	45.77 ±1.61	44.41 ±2.47	37.05 ±0.55	40.31±0.89	41.46
	8	35.84 ±0.59	42.22 ±0.90	38.49 ±2.51	45.24 ±2.18	34.41 ±0.10	37.89 ±0.78	**45.30** ±1.83	39.91
	16	39.00 ±0.46	37.04 ±1.23	45.29 ±0.69	35.55 ±1.25	41.67 ±0.47	34.38 ±0.62	50.20 ±0.57	40.44
MAML	4	**45.62** ±1.90	40.06 ±0.84	**49.97** ±1.90	44.93 ±2.01	45.15 ±0.93	36.96 ±0.55	**47.97** ±1.24	44.38
	8	36.99 ±0.75	45.77 ±2.35	34.30 ±1.44	44.16 ±2.16	36.97 ±0.33	35.85 ±1.82	31.27 ±0.57	37.90
	16	**51.48** ±1.76	39.87 ±1.40	42.44 ±0.81	40.87 ±0.78	37.53 ±0.41	35.90 ±0.39	35.39 ±0.55	40.49
Proto-MAML	4	44.31 ±1.80	**45.23** ±2.75	45.47 ±3.10	**48.16** ±3.51	**60.99** ±2.41	**45.34** ±2.52	43.62 ±3.61	**47.58**
	8	**46.85** ±5.23	44.48 ±3.71	**44.92** ±2.21	42.93 ±2.40	**60.51** ±0.43	**49.93** ±1.26	44.43 ±3.08	**47.72**
	16	44.42 ± 2.91	**61.94** ±0.74	**61.24** ±2.91	**67.36** ±1.54	**64.41** ±1.14	**70.64** ±0.06	**68.80** ±1.08	**62.68**

2) IMPLEMENTATION

As the implementation and execution environment, we use Lab-IA[5] platform provided by The French National Center for Scientific Research[6] (CNRS) with a NVIDIA Tesla V100 GPU with 32 GiB of RAM (NVLink).

C. RESULTS AND DISCUSSIONS

In this section, we evaluate the training models on hate speech and offensive language detection tasks with different languages.

1) HATE SPEECH DETECTION

In this task, we combine all datasets in each language as reported in Table 1. Due to the lack of a held-out benchmark test set for each dataset, after combining all datasets in each language, we select 20% of samples in each language as test set by performing a stratified sampling. To have a variety of tasks during the meta-training step, we leverage different languages with different hateful content where all languages except two are selected as training set. For example, to evaluate meta-learning models on Arabic as a target language with k labeled samples per class, we consider one language for validation and the rest of languages for training, where $L_{train} = \{English, French, German, Indonesian, Spanish, Portuguese\}$, $L_{val} = \{Italian\}$, and $L_{test} = \{Arabic\}$. According to the literature in low-resource NLP classification tasks [75], it can be unreasonable to assume that we have a large validation set; thus we consider only one language in L_{val} set for all experiments. Performing initial experiments led us to choose Italian as validation language. Therefore, in all experiments we set $L_{val} = \{Italian\}$ except when Italian is used as a target language at which we set $L_{val} = \{Spanish\}$. The ratio of validation samples is set to 20% of the original dataset. As English has been frequently used in hate speech detection tasks with a large labeled data, we consider it as a high-resource language and fix it in L_{train} during all experiments.

2) OFFENSIVE LANGUAGE DETECTION

In this task, there exists one dataset per language that has a specific held-out test set, provided by OffensEval 2020, and we use this test set for evaluation. Only for Persian, which is provided by us, we sample a ratio of 20% of the data as test set. Similar to the hate speech dataset, in each experiment we consider all languages except two as training set, where English is always included. We set $L_{val} = \{Turkish\}$ except when Turkish is used as a target language at which we set $L_{val} = \{Arabic\}$. The ratio of validation samples is set to 10% original dataset.

Towards a faithful evaluation amongst all models, we keep the same train, validation, and test samples in all experiments. In our few-shot setting, we evaluate the models on $k \in \{4, 8, 16\}$ and due to the sensitivity of models to the k samples chosen from the target language in test set, we perform each experiment based on 10 testing episodes (for each k) and report the average performance in terms of macro F1-measure over 5 different random seeds. For the XLM-R and non-episodic baselines, we select 10 different random sets in size k and report the average performance.

Tables 2 and 3 present the results for k-shot hate speech and offensive language detection datasets, respectively. The performance of each model for each k-shot setting is displayed in terms of macro-averaged F1-measure along with the standard deviations. Each column corresponds to an unseen target language and the last column shows the average performance of each model on all target languages,

[5] https://doc.lab-ia.fr/

[6] http://www.cnrs.fr/

TABLE 3. Results of k-shot classification on the unseen target languages of offensive language dataset in terms of macro F1-measure with standard deviation. The values in bold indicate the best performing model in each k-shot setting. The last column corresponds to the row-wise average F1-measure across all target languages.

Models	k-shot	Target Languages					
		ar	da	fa	gr	tr	avg
XLM-R [5]	4	33.76 ±0.92	40.26 ±2.55	**43.47** ±1.16	32.17 ±0.56	38.76 ±2.32	37.68
	8	37.60 ±2.42	39.60 ±3.05	45.04 ±2.38	38.87 ±0.68	**46.62** ±1.30	41.54
	16	40.32 ±1.62	42.09 ±2.15	45.76 ±1.08	39.26 ±0.78	46.95 ±1.18	42.87
Non-episodic	4	30.67 ±0.93	35.27 ±1.84	30.08 ±1.34	31.36 ±1.12	36.29 ±0.83	32.73
	8	47.69 ±1.13	32.02 ±1.14	43.60 ±1.26	34.32 ±1.24	39.62 ±0.68	39.45
	16	48.67 ±1.02	40.83 ±2.12	44.36 ±0.72	31.72 ±3.05	49.16 ±1.21	42.94
MAML	4	**51.12** ±1.11	46.64 ±1.66	30.90 ±0.58	**41.14** ±1.75	**42.82** ±4.22	42.52
	8	54.04 ±0.90	45.51 ±1.51	**54.68** ±1.81	51.52 ±1.98	40.38 ±0.44	49.22
	16	48.89 ±1.12	47.81 ±2.08	46.35 ±0.49	41.21 ±2.75	56.55 ±0.59	48.16
Proto-MAML	4	41.05 ±1.32	**57.84** ±2.60	43.21 ±0.91	40.50 ±1.21	41.70 ±2.01	**44.86**
	8	**58.65** ±2.06	**57.73** ±3.12	45.98 ±2.24	**59.70** ±2.50	46.25 ±3.52	**53.66**
	16	**64.16** ±3.14	**59.80** ±2.71	**72.92** ±4.77	**60.55** ±2.25	**60.64** ±2.36	**63.61**

for the sake of comparison. The values in bold indicate the best performing model in each k-shot setting.

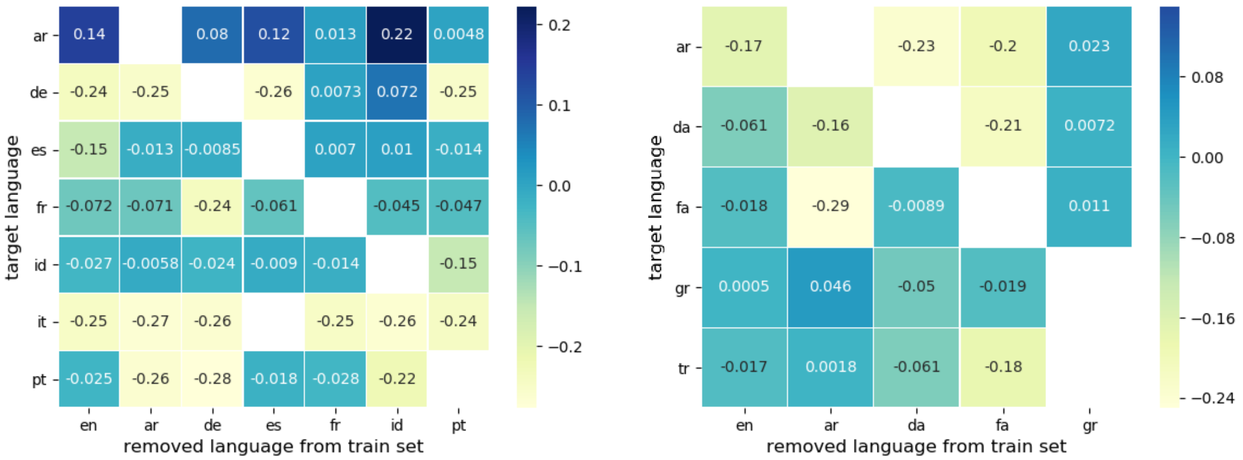
Generally, the results clearly demonstrate that meta learning-based models, MAML and Proto-MAML, outperform other models in most cases, and Proto-MAML achieves the best performance across two datasets in the majority of settings. Regarding the last columns in both tables, when comparing against MAML, Proto-MAML improves notably by 6.7%, 20%, and 35% on average in 4-,8-, and 16-shot classification for hate speech dataset and by 5.2%, 8.3%, and 24.3% on average in 4-,8-, and 16-shot classification for offensive language dataset, respectively. Therefore, this specifies the high ability of Proto-MAML in generalizing to the new language given a few samples.

Considering two baselines XLM-R and non-episodic, we observe that in most settings, XLM-R achieves better results. Since the non-episodic baseline trains in a non-episodic fashion and concatenates the samples of all training languages during training, it performs the training process the same as XLM-R baseline. However, the main difference between these two baselines is in the choice of validation language to select the best model; where XLM-R uses the target language for validation, whereas non-episodic uses two different languages in the validation and test steps. Although the results show that using the same language for the best model selection (validation step) and test step yields better performance, it is not aligned with our assumption in cross-lingual few-shot setting in which test language remains unseen during training and validation.

An interesting observation is that although all models are initialized and fine-tuned with pre-trained language model XLM-R, the cross-lingual knowledge in hate and offensive contexts is not transferred by baselines across languages well. Whereas, Proto-MAML leverages the cross-lingual class prototypes along with initial parameters performing equally well across languages in meta-training step to benefit from XLM-R embeddings.

Regarding Table 2, we perceive that hateful content in different languages are more transferable through meta learning-based models (MAML and Proto-MAML) in comparison with transfer learning-based model (XLM-R); where, Proto-MAML and MAML achieve the best performances with different k values, except when German, French, and Portuguese are target languages with $k = 8$. Results indicate that increasing the number of labeled data per class (k) does not necessarily lead to better performance incrementally, however $k = 16$ is a stable number for Proto-MAML to perform well on different languages; except for Arabic language with $k = 16$ where MAML outperforms Proto-MAML. An interesting observation is that although we have a heterogeneous set of languages in training, where Arabic and Indonesian are from different language families with low typological commonalities with other languages, meta learning-based models can generalize to these languages with better performance quickly; which is very practical in real applications.

Regarding Table 3, offensive content is well generalized across languages where Proto-MAML is the best-performed model in all target languages with $k = 16$. Hate speech and offensive language are subjective and contextual-based phenomenon and the substantial improvements for languages such as Arabic, Persian, and Turkish indicate that meta learning is most beneficial when we have tasks with heterogeneous languages. More precisely, in hate speech and offensive language we are facing with a domain drift problem in which some context cannot be captured across different languages easily, such as cultural differences. However, our results show that meta learning can alleviate this problem.



(a) Hate speech detection

(b) Offensive language detection

FIGURE 2. Differences in the performance of Proto-MAML after removing each training language from the train set, in terms of F1-measure. Rows correspond to target languages and columns correspond to the removed language from the original train set. Each cell reports performance differences between training on the original train set and the train set without a specific training language.

3) ABLATION STUDY

To analyze the contributions of different training languages on performance of the meta-training process in Proto-MAML model, as the best-performing model, we conduct an ablation study. To that end, we repeat the experiments with training Proto-MAML model with $k = 16$ while removing each language in training set one by one and calculating the performance differences compared with original results (which is reported in Tables 2 and 3 for Proto-MAML/$k = 16$), in terms of F1-measure. Figure 2 shows the relative change in performance when each training language is held out from original train set of hate speech and offensive language detection datasets; where rows indicate target languages and each column corresponds to an held-out training language. Positive values show an improvement in performance after removing a specific training language while the negative values indicate a reduction in the performance. It is noted that, in hate speech, choosing *es* as validation language when the target language is *it* causes an empty cell regarding the case in which the target language and the removed one are *it* and *es*, respectively. Furthermore, in offensive language, choosing *gr* as validation language when the target language is *tr* causes an empty cell regarding the case in which the target language and the removed one are *tr* and *gr*, respectively.

The results specify the effectiveness of each language in the generalization of the model where removing each of them results in a reduction of performance, in major cases (where the performance differences are negative). For the hate speech detection task, as shown in Figure2a, removing the training languages $\{en, ar, de, es, pt\}$ gives rise to a performance reduction except when the target language is *ar*, which indicates a positive contribution of each language in the model' s ability to generalize to the target language with a few labeled samples. A small improvement is observed in the performance of the model for target languages *de* and *es*, when we remove *fr* or *id* from the training set. In addition, surprisingly for target language *ar*, we observe that removing each language during meta-training leads to the performance improvement where removing *id* has the largest impact. We hypothesize that the large distance between *ar* and other languages is the cause of this observation; where *ar* belongs to Afro-Asiatic, *id* belongs to Austronesian, and the rest of languages belong to Indo-European language families. Therefore, in this situation, the choice of training languages has different implications for an unseen target language, and has a crucial impact in the ability of meta learning model to adapt to the new language. The relationship between the training languages and an unseen target language in terms of typology and distance must be investigated further in the future.

For offensive language detection task, as shown in Figure 2b, we observe that removing the training languages $\{en, ar, da, fa\}$ results in a performance reduction in all cases except when *gr* is a target language; at which there exists a small improvement by removing *ar* and *en*. On the other hand, removing *gr* from training set causes an improvement in the performance for all target languages. This indicates that *gr* language has a negative impact across different target languages in meta-training process. However, the diversity of other training languages has positive effect in the performance of the model.

VI. CONCLUSION

Although pre-trained transformer models have yielded promising results in hate speech detection, they require a large amount of labeled data in a specific language; which is not always feasible for low-resource languages. In this paper, we studied the problem of few-shot learning in cross-lingual hate speech and offensive language detection tasks by exploring the feasibility of meta learning approach as a potential solution for the first time, to our knowledge. To that end, we collected a diverse set of publicly available datasets containing hateful and offensive content from different languages to create two benchmark datasets for cross-lingual hate speech and offensive language classification tasks. We employed a meta learning approach based on optimization-based and metric-based methods (MAML and Proto-MAML) to train a model being able to generalize quickly to a new language with a few labeled data (k samples per classes). The experiments demonstrate that meta learning based models outperform transfer learning based models in a majority of cases, and Proto-MAML is the best performing model where it can quickly generalize and adapt to new languages with a few labeled data (mainly 16 sample per class yields an effective performance) to identify hateful or offensive content. In addition, MAML also performs strongly, however transfer learning-based baselines notably presents the lowest results. Our future work will extend this study to investigate different sampling strategies for training tasks and see how different languages in training set affect on the performance of meta learning models for an unseen target language. We will also perform a typological analysis to study the relationships between different language families and the performance of meta learning in cross-lingual hate speech detection task.

APPENDIX A
HATE SPEECH DATASETS

We use 15 publicly available sources in 8 languages provided by research community as follows:

Arabic This category consists of two hateful datasets in Arabic, explained in the following:

- Mulki *et al.* [72] introduced the first Levantine hate speech and abusive Twitter dataset in size of 5,846. Levantine is one of the Arabic dialects used on Twitter. The dataset was collected based on different strategies including: 1) querying for tweets containing the potential entities that are usually targeted by hate or abusive language, and 2) using user timelines belonging to certain politicians, social/political activists and TV anchors with high probability of receiving hate content regarding their tweets and tweets' replies. Three Levantine native speakers annotated the data as *hate*, *abusive*, or *normal*. Here, we only select the tweets labeled as hate or normal.

- Ousidhoum *et al.* [46] built a dataset containing 13,014 tweets in English (5,647), French (4,014), and Arabic (3,353) from Twitter. Here, we just select tweets that have hateful or normal sense in their annotation labels from Arabic samples (3,353).

English This category consists of four different hateful datasets in English, explained in the following:

- Basile *et al.* [14] introduced a multilingual hate speech dataset in English and Spanish for HatEval 2019, a shared task at SemEval 2019, which focuses on identification of multilingual hate speech against immigrants and women in Twitter. The dataset was collected by employing different approaches such as monitoring potential victims of hate accounts, using a set of keywords to filter tweets, and downloading the history of identified haters, and resulted in a composition of 19,600 tweets for English (13,000) and Spanish (6,600). Authors used the crowdsourcing platform Figure Eight (F8) to annotate the data in three categories including: 1) hate speech (*hate speech* or *not hate speech* towards immigrant or women, 2) target range (*generic* or *individual*), and 3) aggressiveness (*aggressive* or *not aggressive*). Here, we only select the first category of annotation for English, in which each tweet is labeled as hate speech or not hate speech.

- Davidson *et al.* [2] built a dataset by crawling and annotating 24,783 tweets in English with using the Twitter API. This dataset was collected using a hate speech lexicon containing words and phrases issued by Hatebase[7] dictionary, and was annotated using the crowdsourcing platform CrowdFlower.[8] Each tweet was labeled as *hate speech*, *offensive*, or *neither*. Here, we only select tweets that are labeled as hate speech or neither.

- Founta *et al.* [26] proposed a methodology for annotating a large-scale dataset that were randomly sampled from Twitter utilizing the Twitter Stream API. The randomly sampled data, in size of 32 million tweets, was boosted with tweets that are likely to belong into the minority classes (containing inappropriate speech) and resulted in 80K tweets. The dataset was annotated to four classes: *hate speech*, *abusive*, *spam*, and *normal* by using a crowdsourcing platform CrowdFlower. Here, we only select tweets marked as either hate or normal.

- Ousidhoum *et al.* [46] built a dataset containing 13,014 tweets in English (5,647), French (4,014), and Arabic (3,353) from Twitter. The authors proposed a multi-aspect annotation schema to annotate the dataset as *offensive*, *disrespectful*, *hateful*, *fearful*, *abusive*, or *normal* using a crowdsourcing mechanism with the Amazon Mechanical Turk[9] platform. They also considered directness and target of hatred and the sentiment of the annotator in their annotation process. Here, we only select tweets in English that have hateful or normal sense in their annotation label.

French We use the dataset introduced in [46], containing 13,014 tweets in English (5,647), French (4,014), and Arabic (3,353) from Twitter. Here, we just select tweets that have hateful or normal sense in their annotation label that results in 1,220 samples in French.

German This category consists of two different hateful datasets in German, explained in the following:

[7] https://hatebase.org

[8] Now the name of platform is changed to Appen: https://appen.com/

[9] https://www.mturk.com/

- Mandl *et al.* [10] created a corpus of size 17,657 in three languages English (7,005), Hindi (5,983), and German (4,669) from Twitter and Facebook, which was introduced in the first edition of HASOC track (Hate Speech and Offensive Content Identification in Indo-European Languages shared task in FIRE 2019). The dataset was collected using a set of hashtags and keywords containing offensive content and users' timelines with potential hateful content. The dataset was annotated in a three-layer annotation schema as: 1) identification of *Hate and Offensive* or *Non Hate-Offensive*, 2) identifying the type of hate as *Hate speech*, *Offensive*, *Profane*, and 3) identifying whether a post is containing *Targeted Insult* or *Untargeted*. Here, we only select samples from the first and second layers of annotation labeled as hate speech or not hate speech.

- Roβ *et al.* [25] introduced the first hate speech corpus, consisting of 469 tweets, for the refugee crisis in German language. The aim of the study was to measure the reliability of hate speech annotations. To collect the dataset, they used a list of hashtags with potential insulting or offensive meaning towards refugees. Two experts annotated the corpus as *hate speech* or*not hate speech*. In addition, the offensiveness of each tweet was rated from 1 (Not offensive at all) to 6 (Very offensive). Here, we select tweets according to a complete agreement between annotators, which results in 369 tweets.

Indonesian We use two following datasets proposed for hate speech detection in Indonesian [49], [74].

- Alfina *et al.* [74] introduced a dataset for hate speech detection in Indonesian containing 713 tweets and was collected from Twitter based on a set of hashtags related to the political events. The dataset was annotated as *hate speech* or *not hate speech* by a group of 30 college students as annotators.

- Ibrohim and Budi [49] built an Indonesian Twitter corpus in size of 13,169 to detect hate speech and abusive language along with the target, category, and level of hate. The dataset contains a combination of existing datasets and new dataset collected from Twitter using Twitter Search API for a duration of 7 months. The dataset was annotated by a large group of annotators using crowdsourcing mechanism and resulted in a multi-label hate speech and abusive language dataset. Here, we only select tweets that are labeled as *hate speech* or *not hate speech*.

Italian Bosco *et al.* [13] used two Italian corpus form Twitter and Facebook for the Hate Speech Detection (HaSpeeDe) task at EVALITA 2018. The first dataset is a collection of 4,000 Facebook posts provided by Vigna *et al.* [76], and the second dataset is a collection of 4,000 tweets from Twitter built by Sanguinetti *et al.* [77]. To keep platform consistency across different datasets, we only use the Twitter dataset here. The Twitter dataset was collected by considering three potential hate speech targets in the Italian context: immigrants, Muslims, and Roma and with employing a set of neutral keywords associated with these groups. Using a combination of experts and crowdsourcing annotators, the dataset was annotated as *hate speech*, *aggressiveness*, *offensiveness*, *irony*, *stereotype*, and *intensity*. Here, we only select tweets labeled as hate speech or not hate speech.

Portuguese This dataset composes of 5,668 tweets in Portuguese [48]. Tweets were collected using a set of hate-related keywords and hate-related profiles. The authors used two annotation schemas: 1) binary annotation (*hate* vs. *no-hate*) relying on non-expert annotators and 2) multi-label hate speech hierarchical annotation (including 81 hate categories) relying on an expert annotator (a researcher in hate speech domain who was trained in social psychology). Here, we only select tweets annotated with binary annotation schema as hate or no-hate.

Spanish This category consists of two hateful datasets in Spanish, explained in the following:

- Basile *et al.* [14] introduced a multilingual hate speech dataset in English and Spanish for HatEval 2019, a shared task at SemEval 2019, which focuses on identification of multilingual hate speech against immigrants and women in Twitter. The dataset was composed of 19,600 tweets for English (13,000) and Spanish (6,600). Here, we only select the first category of annotation for Spanish (6,600), in which each tweet is labeled as hate speech or not hate speech.

- Pereira-Kohatsu *et al.* [73] introduced a dataset on hate speech in Spanish consisting of 6,000 tweets filtered from a corpus of two million tweets, sampled from Twitter using the Twitter Rest API. The filtering process was based on different dictionaries containing absolute hate or relative hate with generic insults. Using expert annotators, the dataset was labeled as *hate speech* or *not hate speech*. Here, we use all samples in the dataset.

APPENDIX B
OFFENSIVE LANGUAGE DATASETS

We use the multilingual offensive language dataset provided in OffensEval-2020, a shared task at SemEval-2020, which focused on multilingual offensive language identification in 5 languages Arabic, Danish, English, Greek, and Turkish. In addition, we collected and annotated an offensive language corpus in Persian from Twitter as a low-resource language in this task.

Arabic This dataset contains 10,000 tweets in Arabic collected from Twitter and annotated by an experienced annotator who is a native Arabic speaker and familiar with several Arabic dialects [17]. The authors considered a specific pattern in tweets to increase the chance of having offensive content, so that an initial collection of 660K tweets having at least two vocative particles ("yA" in Arabic - meaning "O") were collected. The intuition was that the vocative particle ("yA") is mainly used in directing the speech to a specific person or group and this vocative is widely observed in all Arabic dialects containing offensive language. Then 10k out of the initial corpus was selected and annotated as offensive or clean. If a tweet is offensive, then annotator searched for any potential vulgar or hate speech content. Therefore, each tweet is given one or more labels: *offensive*, *vulgar*, *hate* *speech*, or *clean*. In this study, we consider tweets annotated as offensive or clean.

Danish This dataset contains 3,600 comments collected from different three popular social media platforms among Danish speakers: Twitter, Facebook, and Reddit. An initial platform-specific lexicon containing abusive terms in Danish collected through a crowd-sourcing mechanism in Reddit was used in data collection process [15]. The annotation process followed the three-layer annotation scheme proposed in [12], for English, to identify the type and the target of offense. Here, we just used the first level of annotation where each comment is annotated as offensive or non-offensive.

English The Offensive Language Identification Dataset (OLID) containing over 14,000 English tweets, is introduced at SemEval-2019 for identification of offensive language, the type of offensive content, and the target of offensive in English [12]. The OLID targeted different kinds of offensive content and was annotated using a fine-grained three-layer annotation scheme to identify the type and the target of offense as well. In the first level of annotation, tweets are annotated as *offensive* or *non-offensive*. In the second level, offensive tweets are annotated as *targeted insult* or *untargeted*, and in the third level, targeted offensive tweets are annotated as *individual*, *group*, or *other*. Here, we just used the first level of annotation where each tweet is annotated as offensive or non-offensive.

Greek The first version of this dataset, named Offensive Greek Tweet Dataset (OGTD), contains 4,779 posts from Twitter collected between May and June 2019 [45]. Different sampling strategies were used in collecting data including: 1) using popular and trending hashtags in Greek attributed to the television programs, reality and entertainment shows and political tweets, querying for tweets containing keywords usually found in offensive content such as curse words, expletives and their plural forms, and searching for tweets containing (eisai, ''you are'') as a keyword. Following the same annotation guidelines proposed in [12], the dataset was annotated as *offensive*, *not offensive* and *spam*, by a group of three volunteers annotators through the LightTag[10] platform. The spam tweets were filtered out from the dataset. To enrich the corpus for OffensEval 2020, the second version of the dataset, in size 5,508, was collected and annotated in November 2019 with the same approach used in the first version. The combination of two versions results in 10,287 tweet samples that we use in this study.

Turkish This dataset contains over 35,000 tweets extracted from Twitter using Twitter streaming API, from March 2018 to September 2019 [16]. Although a list of frequent words in Turkish tweets was used to filter Twitter streams, all the tweets were sampled uniformly without any strategy such as offensive keywords for extracting offensive content specifically. To annotate the corpus by volunteers, the annotation guidelines proposed in [12] with a small divergence was used; where at the top level, tweets were labeled as *offensive* or *non-offensive* and then offensive content were labeled as *targeted* or *profanity*. Similar to [12], the targeted offensive content were divided to *individual*, *group*, or *other*. Here, we just used the first level of annotation where each tweet is annotated as offensive or non-offensive.

Persian This dataset has been created by the authors of this paper in another research work (which is under review in a journal now) to investigate the problem of offensive language detection in Persian as a low-resource language. This dataset contains 6,000 tweets crawled from Twitter using Twitter streaming API in a two-month interval from June to August 2020. We used two main strategies for data collection: 1) random sampling and 2) lexicon-based sampling in which we used the HurtLex[11] as a multilingual computational lexicon of offensive, aggressive, and hateful words organized in 17 categories in over 50 languages including Persian. Employing random and lexicon-based sampling left us to 320K and 200K tweets, respectively. Finally, we selected 3,000 tweets randomly from each sampling sets (random and lexicon-based) for annotation step. Inspiring the annotation guidelines proposed in [12], we developed an hierarchical annotation protocol for our Persian corpus.

Three highly educated volunteers from the author's personal contacts, who were Persian native speakers, were enrolled for annotating the corpus. Two of annotators were supposed to annotate all the selected tweets at three levels offensive language detection, categorization, and target identification and in the case of agreement the final label was set. Otherwise, the third annotator was asked for labeling the tweet again and then we took a majority vote. Tweets were labeled as *offensive* or *non-offensive* and then offensive content were labeled as *targeted* or *profanity*. At the end, the targeted offensive content were divided to *individual*, *group*, or *other*. Annotation consensus for two annotators on three levels of annotation schema was approximately 73%, in which the agreements in the first level of annotation schema (offensive vs non-offensive) was very high as 86%, in the second level 75%, and in the third level 60%. Here, we just used the first level of annotation where each tweet is annotated as offensive or non-offensive.

REFERENCES

[1] Z. Waseem and D. Hovy, ''Hateful symbols or hateful people? Predictive features for hate speech detection on twitter,'' in *Proc. NAACL Student Res. Workshop*, San Diego, CA, USA, Jun. 2016, pp. 88–93. [Online]. Available: https://www.aclweb.org/anthology/N16-2013

[2] T. Davidson, D. Warmsley, M. Macy, and I. Weber, ''Automated hate speech detection and the problem of offensive language,'' in *Proc. Int. AAAI Conf. Web Social Media*, 2017, pp. 512–515. [Online]. Available: https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15665

[3] Z. Zhang, D. Robinson, and J. Tepper, ''Detecting hate speech on Twitter using a convolution-gru based deep neural network,'' in *The The Semantic Web*, A. Gangemi, R. Navigli, M.-E. Vidal, P. Hitzler, R. Troncy, L. Hollink, A. Tordai, and M. Alam, Eds. Cham, Switzerland: Springer, 2018, pp. 745–760.

[10] https://www.lighttag.io

[11] https://github.com/valeriobasile/hurtlex

[4] A. M. Founta, D. Chatzakou, N. Kourtellis, J. Blackburn, A. Vakali, and I. Leontiadis, "A unified deep learning architecture for abuse detection," in *Proc. 10th ACM Conf. Web Sci.*, New York, NY, USA, Jun. 2019, pp. 105–114, doi: 10.1145/3292522.3326028.

[5] S. S. Aluru, B. Mathew, P. Saha, and A. Mukherjee, "Deep learning models for multilingual hate speech detection," *CoRR*, vol. abs/2004.06465, 2020. [Online]. Available: https://arxiv.org/abs/2004.06465

[6] E. W. Pamungkas, V. Basile, and V. Patti, "A joint learning approach with knowledge injection for zero-shot cross-lingual hate speech detection," *Inf. Process. Manage.*, vol. 58, no. 4, Jul. 2021, Art. no. 102544. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306457321000510

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.* Minneapolis, MN, USA, vol. 1, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[8] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.

[9] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 8440–8451. [Online]. Available: https://www.aclweb.org/anthology/2020.acl-main.747

[10] T. Mandl, S. Modha, P. Majumder, D. Patel, M. Dave, C. Mandlia, and A. Patel, "Overview of the HASOC track at FIRE 2019: Hate speech and offensive content identification in Indo-European languages," in *Proc. 11th Forum Inf. Retr. Eval.*, New York, NY, USA, 2019, pp. 14–17, doi: 10.1145/3368567.3368584.

[11] M. Zampieri, P. Nakov, S. Rosenthal, P. Atanasova, G. Karadzhov, H. Mubarak, L. Derczynski, Z. Pitenis, and Ç. Çöltekin, "SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020)," in *Proc. 14th Workshop Semantic Eval.*, Barcelona, Spain, 2020, pp. 1425–1447. [Online]. Available: https://aclanthology.org/2020.semeval-1.188

[12] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar, "Predicting the type and target of offensive posts in social media," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.* Minneapolis, MN, USA, vol. 1, Jun. 2019, pp. 1415–1420. [Online]. Available: https://www.aclweb.org/anthology/N19-1144

[13] C. Bosco, F. Dell'Orletta, F. Poletto, M. Sanguinetti, and M. Tesconi, "Overview of the EVALITA 2018 hate speech detection task," in *Proc. 6th Eval. Campaign Natural Lang. Process. Speech Tools Italian Final Workshop (EVALITA) Co-Located 5th Italian Conf. Comput. Linguistics (CLiC)* (CEUR Workshop Proceedings), vol. 2263. T. Caselli, N. Novielli, V. Patti, and P. Rosso, Eds. Turin, Italy, Dec. 2018. [Online]. Available: http://ceur-ws.org/Vol-2263/paper010.pdf

[14] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. Rangel Pardo, P. Rosso, and M. Sanguinetti, "SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter," in *Proc. 13th Int. Workshop Semantic Eval.*, Minneapolis, MN, USA, 2019, pp. 54–63. [Online]. Available: https://www.aclweb.org/anthology/S19-2007

[15] G. Sigurbergsson and L. Derczynski, "Offensive language and hate speech detection for Danish," in *Proc. Int. Conf. Lang. Resour. Eval.*, May 2020, pp. 3498–3508.

[16] C. Çöltekin, "A corpus of Turkish offensive language on social media," in *Proc. 12th Lang. Resour. Eval. Conf.*, Marseille, France, 2020, pp. 6174–6184. [Online]. Available: https://www.aclweb.org/anthology/2020.lrec-1.758

[17] H. Mubarak, A. Rashed, K. Darwish, Y. Samih, and A. Abdelali, "Arabic offensive language on Twitter: Analysis and experiments," 2020, *arXiv:2004.02192*.

[18] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. 34th Int. Conf. Mach. Learn.* (Proceedings of Machine Learning Research), vol. 70, D. Precup and Y. W. Teh, Eds. Sydney, NSW, Australia, Aug. 2017, pp. 1126–1135. [Online]. Available: http://proceedings.mlr.press/v70/finn17a.html

[19] L. Zintgraf, K. Shiarli, V. Kurin, K. Hofmann, and S. Whiteson, "Fast context adaptation via meta-learning," in *Proc. 36th Int. Conf. Mach. Learn.* (Proceedings of Machine Learning Research), vol. 97. K. Chaudhuri and R. Salakhutdinov, Eds. 09–15, Jun. 2019, pp. 7693–7702. [Online]. Available: http://proceedings.mlr.press/v97/zintgraf19a.html

[20] J. Gu, Y. Wang, Y. Chen, V. O. K. Li, and K. Cho, "Meta-learning for low-resource neural machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Brussels, Belgium, 2018, pp. 3622–3631. [Online]. Available: https://www.aclweb.org/anthology/D18-1398

[21] T. Bansal, R. Jha, and A. McCallum, "Learning to few-shot learn across diverse natural language classification tasks," in *Proc. 28th Int. Conf. Comput. Linguistics*, Barcelona, Spain, 2020, pp. 5108–5123. [Online]. Available: https://aclanthology.org/2020.coling-main.448

[22] E. Triantafillou, T. Zhu, V. Dumoulin, P. Lamblin, U. Evci, K. Xu, R. Goroshin, C. Gelada, K. Swersky, P. Manzagol, and H. Larochelle, "Meta-dataset: A dataset of datasets for learning to learn from few examples," in *Proc. 8th Int. Conf. Learn. Represent. (ICLR)*, Apr. 2020, pp. 1–24. [Online]. Available: https://openreview.net/forum?id=rkgAGAVKPr

[23] Z. Waseem, T. Davidson, D. Warmsley, and I. Weber, "Understanding abuse: A typology of abusive language detection subtasks," in *Proc. 1st Workshop Abusive Lang. Online*, Vancouver, BC, Canada, 2017, pp. 78–84. [Online]. Available: https://www.aclweb.org/anthology/W17-3012

[24] Z. Waseem, "Are you a racist or am i seeing things? Annotator influence on hate speech detection on Twitter," in *Proc. 1st Workshop NLP Comput. Social Sci.*, Austin, TX, USA, 2016, pp. 138–142. [Online]. Available: https://aclanthology.org/W16-5618

[25] B. Roß, M. Rist, G. Carbonell, B. Cabrera, N. Kurowsky, and M. Wojatzki, "Measuring the reliability of hate speech annotations: The case of the European refugee crisis," Sep. 2016, *arXiv:1701.08118*.

[26] A.-M. Founta, C. Djouvas, D. Chatzakou, I. Leontiadis, J. Blackburn, G. Stringhini, A. Vakali, M. Sirivianos, and N. Kourtellis, "Large scale crowdsourcing and characterization of Twitter abusive behavior," in *Proc. 11th Int. Conf. Web Social Media (ICWSM)*, 2018, pp. 491–500.

[27] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep learning for hate speech detection in tweets," in *Proc. 26th Int. Conf. World Wide Web Companion (WWW Companion)*, Geneva, Switzerland, 2017, pp. 759–760, doi: 10.1145/3041021.3054223.

[28] P. Fortuna, J. Soler-Company, and L. Wanner, "How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets?" *Inf. Process. Manage.*, vol. 58, no. 3, May 2021, Art. no. 102524. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306457321000339

[29] A. Arango, J. Pérez, and B. Poblete, "Hate speech detection is not as easy as you may think: A closer look at model validation," in *Proc. 42nd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, New York, NY, USA, Jul. 2019, pp. 45–54, doi: 10.1145/3331184.3331262.

[30] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang, "Abusive language detection in online user content," in *Proc. 25th Int. Conf. World Wide Web*, Geneva, Switzerland, Apr. 2016, pp. 145–153, doi: 10.1145/2872427.2883062.

[31] J. H. Park, J. Shin, and P. Fung, "Reducing gender bias in abusive language detection," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Brussels, Belgium, 2018, pp. 2799–2804. [Online]. Available: https://www.aclweb.org/anthology/D18-1302

[32] M. Mozafari, R. Farahbakhsh, and N. Crespi, "Hate speech detection and racial bias mitigation in social media based on BERT model," *PLoS ONE*, vol. 15, no. 8, pp. 1–26, Aug. 2020. [Online]. Available: https://doi.org/10.1371/journal.pone.0237861

[33] M. Wich, J. Bauer, and G. Groh, "Impact of politically biased data on hate speech classification," in *Proc. 4th Workshop Online Abuse Harms*, 2020, pp. 54–64. [Online]. Available: https://www.aclweb.org/anthology/2020.alw-1.7

[34] S. Kiritchenko, I. Nejadgholi, and K. C. Fraser, "Confronting abusive language online: A survey from the ethical and human rights perspective," *J. Artif. Intell. Res.*, vol. 71, pp. 431–478, Jul. 2021.

[35] Y. Mehdad and J. Tetreault, "Do characters abuse more than words?" in *Proc. 17th Annu. Meeting Special Interest Group Discourse Dialogue*, Los Angeles, CA, USA, 2016, pp. 299–303. [Online]. Available: https://www.aclweb.org/anthology/W16-3638

[36] S. Malmasi and M. Zampieri, "Challenges in discriminating profanity from hate speech," *J. Exp. Theor. Artif. Intell.*, vol. 30, no. 2, pp. 187–202, Mar. 2018, doi: 10.1080/0952813X.2017.1409284.

[37] D. Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini, and A. Vakali, "Measuring #GamerGate: A tale of hate, sexism, and bullying," in *Proc. 26th Int. Conf. World Wide Web Companion*, Geneva, Switzerland, 2017, pp. 1285–1290, doi: 10.1145/3041021.3053890.

[38] E. F. Unsvåg and B. Gambäck, "The effects of user features on Twitter hate speech detection," in *Proc. 2nd Workshop Abusive Lang. Online (ALW2)*, Brussels, Belgium, 2018, pp. 75–85. [Online]. Available: https://www.aclweb.org/anthology/W18-5110

[39] J. H. Park and P. Fung, "One-step and two-step classification for abusive language detection on Twitter," in *Proc. 1st Workshop Abusive Lang. Online*. Vancouver, BC, Canada, Aug. 2017, pp. 41–45. [Online]. Available: https://www.aclweb.org/anthology/W17-3006

[40] L. Gao and R. Huang, "Detecting online hate speech using context aware models," in *Proc. Recent Adv. Natural Lang. Process. Meet Deep Learn. (RANLP)*, Varna, Bulgaria, Nov. 2017, pp. 260–266, doi: 10.26615/978-954-452-049-6_036.

[41] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, New Orleans, LA, USA, vol. 1, Jun. 2018, pp. 2227–2237. [Online]. Available: https://www.aclweb.org/anthology/N18-1202

[42] M. Mozafari, R. Farahbakhsh, and N. Crespi, "A BERT-based transfer learning approach for hate speech detection in online social media," in *Complex Networks and Their Applications VIII*, H. Cherifi, S. Gaito, J. F. Mendes, E. Moro, and L. M. Rocha, Eds. Cham, Switzerland: Springer, 2020, pp. 928–940.

[43] P. Alonso, R. Saini, and G. Kovács, "Hate speech detection using transformer ensembles on the HASOC dataset," in *Speech and Computer*, A. Karpov and R. Potapova, Eds. Cham, Switzerland: Springer, 2020, pp. 13–21.

[44] T. Ranasinghe and M. Zampieri, "Multilingual offensive language identification with cross-lingual embeddings," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 5838–5844. [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-main.470

[45] Z. Pitenis, M. Zampieri, and T. Ranasinghe, "Offensive language identification in Greek," in *Proc. LREC*, 2020, pp. 5113–5119.

[46] N. Ousidhoum, Z. Lin, H. Zhang, Y. Song, and D.-Y. Yeung, "Multilingual and multi-aspect hate speech analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, Hong Kong, 2019, pp. 4675–4684. [Online]. Available: https://www.aclweb.org/anthology/D19-1474

[47] S. Tulkens, L. Hilte, E. Lodewyckx, B. Verhoeven, and W. Daelemans, "A dictionary-based approach to racism detection in Dutch social media," 2016, *arXiv:1608.08738*.

[48] P. Fortuna, J. Rocha da Silva, J. Soler-Company, L. Wanner, and S. Nunes, "A hierarchically-labeled Portuguese hate speech dataset," in *Proc. 3rd Workshop Abusive Lang. Online*, Florence, Italy, 2019, pp. 94–104. [Online]. Available: https://www.aclweb.org/anthology/W19-3510

[49] M. O. Ibrohim and I. Budi, "Multi-label hate speech and abusive language detection in Indonesian Twitter," in *Proc. 3rd Workshop Abusive Lang. Online*, Florence, Italy, 2019, pp. 46–57. [Online]. Available: https://www.aclweb.org/anthology/W19-3506

[50] E. Fersini, D. Nozza, and P. Rosso, "Overview of the Evalita 2018 task on automatic misogyny identification (AMI)," in *Proc. 6th Eval. Campaign Natural Lang. Process. Speech Tools Italian Final Workshop (EVALITA) Co-Located 5th Italian Conf. Comput. Linguistics (CLiC)* (CEUR Workshop Proceedings), vol. 2263. T. Caselli, N. Novielli, V. Patti, and P. Rosso, Eds. Turin, Italy, 2018, p. 282. [Online]. Available: http://ceur-ws.org/Vol-2263/paper009.pdf

[51] E. Fersini, P. Rosso, and M. Anzovino, "Overview of the task on automatic misogyny identification at IberEval 2018," in *Proc. 3rd Workshop Eval. Hum. Lang. Technol. Iberian Lang. (IberEval) Co-Located 34th Conf. Spanish Soc. Natural Lang. Process. (SEPLN)* (CEUR Workshop Proceedings), vol. 2150. P. Rosso, J. Gonzalo, R. Martínez, S. Montalvo, and J. C. de Albornoz, Eds. Barcelona, Spain, Sep. 2018, pp. 214–228. [Online]. Available: http://ceur-ws.org/Vol-2150/overview-AMI.pdf

[52] R. Kumar, A. K. Ojha, B. Lahiri, M. Zampieri, S. Malmasi, V. Murdock, and D. Kadar, Eds., *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*. Marseille, France: Association for Computational Linguistics, May 2020. [Online]. Available: https://www.aclweb.org/anthology/2020.trac-1.0

[53] S. Akiwowo, B. Vidgen, V. Prabhakaran, and Z. Waseem, Eds., *Proceedings of the Fourth Workshop on Online Abuse and Harms*. Stroudsburg, PA, USA: Association for Computational Linguistics, Nov. 2020. [Online]. Available: https://www.aclweb.org/anthology/2020.alw-1.0

[54] M. O. Ibrohim and I. Budi, "Translated vs non-translated method for multilingual hate speech identification in Twitter," *Int. J. Adv. Sci., Eng. Inf. Technol.*, vol. 9, no. 4, pp. 1116–1123, 2019. [Online]. Available: http://ijaseit.insightsociety.org/index.php?option=com_content&view=article&id=9&Itemid=1&article_id=8123

[55] M. Corazza, S. Menini, E. Cabrio, S. Tonelli, and S. Villata, "A multilingual evaluation for online hate speech detection," *ACM Trans. Internet Technol.*, vol. 20, no. 2, pp. 1–22, May 2020, doi: 10.1145/3377323.

[56] N. Vashistha and A. Zubiaga, "Online multilingual hate speech detection: Experimenting with Hindi and English social media," *Information*, vol. 12, no. 1, p. 5, Dec. 2020. [Online]. Available: https://www.mdpi.com/2078-2489/12/1/5

[57] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A lite BERT for self-supervised learning of language representations," in *Proc. 8th Int. Conf. Learn. Represent. (ICLR)*, Addis Ababa, Ethiopia, Apr. 2020, pp. 1–17. [Online]. Available: https://openreview.net/forum?id=H1eA7AEtvS

[58] S. Wang, J. Liu, X. Ouyang, and Y. Sun, "Galileo at SemEval-2020 task 12: Multi-lingual learning for offensive language identification using pre-trained language models," in *Proc. 14th Workshop Semantic Eval. (SemEval@COLING)*, A. Herbelot, X. Zhu, A. Palmer, N. Schneider, J. May, and E. Shutova, Eds. Barcelona, Spain: International Committee for Computational Linguistics, 2020, pp. 1448–1455. [Online]. Available: https://aclanthology.org/2020.semeval-1.189/

[59] G. Wiedemann, S. M. Yimam, and C. Biemann, "UHH-LT at SemEval-2020 task 12: Fine-tuning of pre-trained transformer networks for offensive language detection," in *Proc. 14th Workshop Semantic Eval., (SemEval@COLING)*, A. Herbelot, X. Zhu, A. Palmer, N. Schneider, J. May, and E. Shutova, Eds. Barcelona, Spain: International Committee for Computational Linguistics, 2020, pp. 1638–1644. [Online]. Available: https://aclanthology.org/2020.semeval-1.213/

[60] L. Stappen, F. Brunn, and B. Schuller, "Cross-lingual Zero- and few-shot hate speech detection utilising frozen transformer language models and AXEL," 2020, *arXiv:2004.13850*.

[61] Z.-Y. Dou, K. Yu, and A. Anastasopoulos, "Investigating meta-learning algorithms for low-resource natural language understanding tasks," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, Hong Kong, 2019, pp. 1192–1197. [Online]. Available: https://www.aclweb.org/anthology/D19-1112

[62] F. Nooralahzadeh, G. Bekoulis, J. Bjerva, and I. Augenstein, "Zero-shot cross-lingual transfer with meta learning," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 4547–4562. [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-main.368

[63] I. Tarunesh, S. Khyalia, V. Kumar, G. Ramakrishnan, and P. Jyothi, "Meta-learning for effective multi-task and multilingual modelling," in *Proc. 16th Conf. Eur. Chapter Assoc. Comput. Linguistics: Main*, P. Merlo, J. Tiedemann, and R. Tsarfaty, Eds. Stroudsburg, PA, USA: Association for Computational Linguistics, 2021, pp. 3600–3612. [Online]. Available: https://aclanthology.org/2021.eacl-main.314/

[64] D. Sui, Y. Chen, B. Mao, D. Qiu, K. Liu, and J. Zhao, "Knowledge guided metric learning for few-shot text classification," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2021, pp. 3266–3271. [Online]. Available: https://aclanthology.org/2021.naacl-main.261

[65] S. Thrun and L. Pratt, *Learning to Learn: Introduction and Overview*. Boston, MA, USA: Springer, 1998, pp. 3–17, doi: 10.1007/978-1-4615-5529-2_1.

[66] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *Proc. ICML Deep Learn. Workshop*, 2015, pp. 1–30.

[67] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2017, pp. 1–11. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/cb8da6767461f2812ae4290eac7cbc42-Paper.pdf

[68] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, "Meta-learning with memory-augmented neural networks," in *Proc. 33rd Int. Conf. Mach. Learn.* (Proceedings of Machine Learning Research), vol. 48, M. F. Balcan and K. Q. Weinberger, Eds. New York, NY, USA: PMLR, 20–22, Jun. 2016, pp. 1842–1850. [Online]. Available: http://proceedings.mlr.press/v48/santoro16.html

[69] S. Ravi and H. Larochelle, "Optimization as a model for few-shot learning," in *Proc. 5th Int. Conf. Learn. Represent. (ICLR)*, Toulon, France, Apr. 2017, pp. 1–11. [Online]. Available: https://openreview.net/forum?id=rJY0-Kcll

[70] A. Antoniou, H. Edwards, and A. Storkey, "How to train your MAML," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–11. [Online]. Available: https://openreview.net/forum?id=HJGven05Y7

[71] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," in *Proc. 5th Int. Conf. Learn. Represent. (ICLR)*, Toulon, France, Apr. 2017, pp. 1–16. [Online]. Available: https://openreview.net/forum?id=Skq89Scxx

[72] H. Mulki, H. Haddad, C. Bechikh Ali, and H. Alshabani, "L-HSAB: A levantine Twitter dataset for hate speech and abusive language," in *Proc. 3rd Workshop Abusive Lang. Online*, Florence, Italy, 2019, pp. 111–118. [Online]. Available: https://www.aclweb.org/anthology/W19-3512

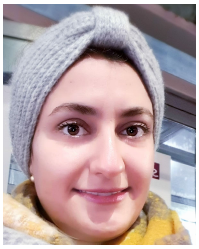
[73] J. C. Pereira-Kohatsu, L. Quijano-Sánchez, F. Liberatore, and M. Camacho-Collados, "Detecting and monitoring hate speech in Twitter," *Sensors*, vol. 19, no. 21, p. 4654, Oct. 2019. [Online]. Available: https://www.mdpi.com/1424-8220/19/21/4654

[74] I. Alfina, R. Mulia, M. I. Fanany, and Y. Ekanata, "Hate speech detection in the Indonesian language: A dataset and preliminary study," in *Proc. Int. Conf. Adv. Comput. Sci. Inf. Syst. (ICACSIS)*, Oct. 2017, pp. 233–238.

[75] K. Kann, K. Cho, and S. R. Bowman, "Towards realistic practices in low-resource natural language processing: The development set," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, Hong Kong, 2019, pp. 3342–3349. [Online]. Available: https://www.aclweb.org/anthology/D19-1329

[76] F. D. Vigna, A. Cimino, F. Dell'Orletta, M. Petrocchi, and M. Tesconi, "Hate me, hate me not: Hate speech detection on Facebook," in *Proc. 1st Italian Conf. Cybersecurity (ITASEC)* (CEUR Workshop Proceedings), vol. 1816, A. Armando, R. Baldoni, and R. Focardi, Eds. Venice, Italy, 2017, pp. 86–95. [Online]. Available: http://ceur-ws.org/Vol-1816/paper-09.pdf

[77] M. Sanguinetti, F. Poletto, C. Bosco, V. Patti, and M. Stranisci, "An Italian Twitter corpus of hate speech against immigrants," in *Proc. 11th Int. Conf. Lang. Resour. Eval. (LREC)*. Miyazaki, Japan: European Language Resources Association (ELRA), May 2018, pp. 1–8. [Online]. Available: https://www.aclweb.org/anthology/L18-1443



MARZIEH MOZAFARI (Student Member, IEEE) received the M.Sc. degree in computer science from the University of Tehran, Iran, in 2017, and the Ph.D. degree in computer science from the Institut Polytechnique de Paris, France, in 2021. She is currently a Researcher with the Institut Polytechnique de Paris. Her research interests include NLP, hate speech detection, transfer learning, machine learning, deep learning, and online social networks analysis.

REZA FARAHBAKHSH (Member, IEEE) received the Ph.D. degree from Paris VI (UPMC) jointly with the Institut-Mines Telecom, Telecom Sud-Paris (CNRS Lab UMR5157), in 2015. He is currently an Adjunct Associate Professor with the Institut Polytechnique de Paris, Telecom SudParis, and a Data Scientist at TOTAL SA. He is actively involved in international collaborative projects. His research interests include NLP, language modeling, online social networks, and the IoT.

NOEL CRESPI (Member, IEEE) received the master's degree from the Universities of Orsay and Canterbury, the Diplome d'ingénieur degree from Telecom ParisTech, and the Ph.D. and Habilitation degrees from Paris VI University. He joined the Institut Mines-Telecom in 2002. He is currently a Professor and the M.Sc. Program Director at Institut Polytechnique de Paris, Telecom SudParis, where he is leading the Service Architecture Laboratory. He coordinates the standardization activities with Institute Telecom at ETSI, 3GPP, and ITU-T. He is also an Adjunct Professor at KAIST, South Korea, an Affiliate Professor at Concordia University, Canada, and a Guest Researcher at the University of Goettingen, Germany. He is the Scientific Director of the French-Korean Laboratory ILLUMINE. He is the author/coauthor of 400 articles and contributions in standardization. His current research interests include data analytics, the Internet of Things, and softwarization.