Emotionally-bridged cross-lingual meta-learning for chinese sexism detection

Guanlin Li, Praboda Rajapaksha, Reza Farahbakhsh, Noel Crespi

Emotionally-Bridged Cross-Lingual Meta-Learning for Chinese Sexism Detection

Guanlin Li, Praboda Rajapaksha, Reza Farahbakhsh, and Noel Crespi

Samovar, Telecom SudParis, Institut Polytechnique de Paris, 91120 Palaiseau, France
{guanlin.li,praboda.rajapaksha,reza.farahbakhsh,
noel.crespi}@telecom-sudparis.eu

Abstract. Sexism detection remains as an extremely low-resource task for most of the languages including Chinese. To address this issue, we propose a zero-shot cross-lingual method to detect sexist speech in Chinese and perform qualitative and quantitative analyses on the data we employed. The proposed method aims to explicitly model the knowledge transfer process from rich-resource language to low-resource language using metric-based meta-learning. To overcome the semantic disparity between various languages caused by language-specific biases, a common label space of emotions expressed across languages is used to integrate universal emotion features into the meta-learning framework. Experiment results show that the proposed method improves over the state-of-the-art zero-shot cross-lingual classification methods.

Keywords: Sexist Speech Detection · Cross-lingual · Meta-learning.

1 Introduction

The rise of the internet and new media calls for more attention to gender awareness and solidarity, for which automatic methods are needed to identify sexism on social media. Sexist speech is usually defined as prejudice, stereotyping or discrimination, typically against women, on the basis of sex, which can cause measurable negative impact [1]. As the volume of social media content continues to increase, it is important to detect sexist speech automatically so as to prevent the circulation of such speech on social media platforms and also to better study the related phenomenon. Previous works typically viewed sexism detection as a supervised classification problem. Waseem and Hovy [2] studied sexist speech as a category of hate speech and constructed a hate speech dataset comprised of sexism and racism classes. Two AMI (Automatic Misogyny Identification) shared tasks, IberEval2018 [3] and EVALITA2018 [4] provided datasets of misogyny-related speech in social media with multiple languages (English, Spanish and Italian) and extensive studies of automatic misogyny detection have been done based on the AMI datasets [5]. Sexist speech is not always hateful and has the forms of hostile and benevolent seixsm [6]. Based on the ambivalent sexism theory, Jha and Mamidi [7] constructed a dataset containing three categories of

sexism, namely benevolent, hostile and others and developed classification models to identify benevolent sexism. Samory et al. [1] further proposed to measure sexism in psychological scales and defined four categories of sexist content: behavioral expectations, stereotypes and comparisons, endorsements of inequality, denying inequality and rejection of feminism.

Despite the increased interest in sexist speech detection on social media, the number of studies is fewer compared to that of hate speech detection in general, and most of the sexism detection-related research works focused on Indo-European languages [8]. Thus, in this paper, we study the problem of Chinese sexism detection with the help of cross-lingual knowledge transfer. Instead of only using binary sentiment features, we propose to integrate external emotion knowledge about sexism datasets within the framework of meta-learning to explicitly model the transfer process between languages, while the heterogeneity of emotion labels in different training sources is bridged by a unified taxonomy. Multilingual language models are used as the backbone model so that the method can be generalized to other low-resource languages. To eliminate the need for auxiliary tasks and languages in the meta-learning process, machine translation is used to generate samples which are used to provide gradient during the meta-training stage. Experiments on cross-lingual datasets composed of English (resource-rich language) and Chinese sexist speech show that the proposed method improves upon previous state-of-the-art models.

2 Related Works

In this part, we briefly survey the more general topic of multilingual hate speech detection. In cross-lingual hate speech detection, resource-rich languages are used as source language to provide sexism related knowledge, and zero-shot or few-shot prediction is done on low-resource target languages. Pamungkas et al. [5] studied the features of misogynistic content and investigated misogyny detection on cross-domain and multilingual content. Furthermore, they experimented with several different methods and joint learning approach to perform multilingual misogyny detection in English and Italian [9]. Jiang and Zubiaga [10] proposed a capsule network for cross-lingual hate speech detection. The network relies on the source language and its translated counterpart in the the target language. Aluru et al. [11] compared the performance of LASER embedding and mBERT on cross-lingual hate speech detection using datasets in 9 different languages and found that in low resource setting LASER performs better.

To sum up, the models employed in cross-lingual hate speech detection can be categorized as follows: (1) Monolingual embedding and machine translation of target languages; (2) Multilingual embeddings and supervised classification model; (3) Multilingual pretrained language models (mPLM) and the combination of above models. However, the performance of such models could be strongly affected by the negative effect of non-hateful, language-specific taboo interjections [12] and data overfitting [13].

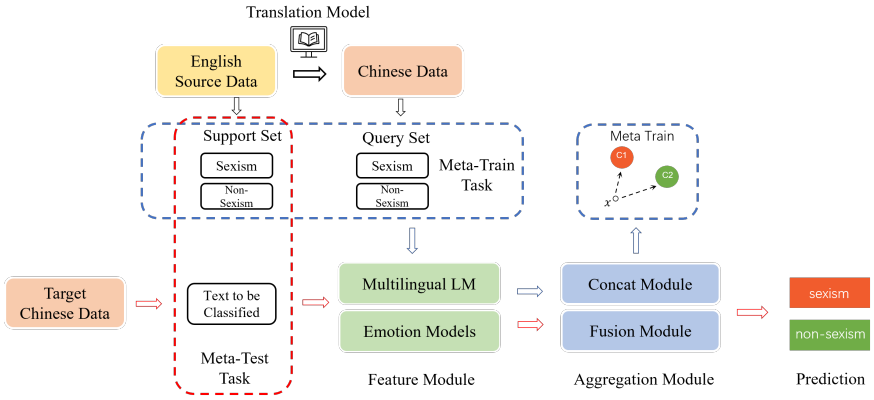

Fig. 1: Framework of the proposed method (better viewed in color). In the meta-train stage, the query sets are generated from the source data; in the meta-test stage, the query sets are composed of target data. Blue arrows are used to indicate the meta-train flow and red arrows are used to indicate the inference.

3 Methodology

In this section, we present the proposed zero-shot cross-lingual meta-learning method for sexism detection which takes advantage of external emotion knowledge. The framework of the proposed method is shown in Figure 1. We elaborate on the essential parts of the method, namely cross-lingual meta-learning, emotion analysis and emotion knowledge injection.

3.1 Cross-lingual Meta Learning

For the task sets $\mathcal{T}$ under the zero-shot cross-lingual setting, the support set is denoted as $D^s_{\ell_s} = \{X^s_{\ell_s}, Y^s_{\ell_s}\}$, and the query set is denoted as $D^q_{\ell_t} = \{X^q_{\ell_t}, Y^q_{\ell_t}\}$, where data points in support sets are sampled from the source language and data points in query sets are sampled from a different language. To enable zero-shot learning in meta-learning, previous methods require auxiliary languages to provide samples for query sets [14, 15]. In our setting, we eliminate the need for auxiliary language by using machine translation to generate data in the target language from the source language. Given a translation model M^t_s which translates from source language ℓ_s to target language ℓ_t , for a sample D^i_s from dataset in source language ℓ_s, the corresponding sample D'^i_s in target language ℓ_t is generated. Then, support sets are sampled from D_s and query sets are sampled from D'_s, such that, during the meta-training stage, only labels from the source language are used. To get universal features for support and query sets, we denote the multilingual model as f_m, the features of an example x_i in the support set S_k of class k as $f_m(x_i)$, the features of a query sample x_j in the query set Q as $f_m(x_j)$. The prototype features c_k of S_k is denoted as $1/|S_k| \sum_{(x_i,y_i)\in S_k} f_m(x_i)$.

To predict the probability p given x_j belonging to class k, we use distance function d to measure the distance between prototype c_k and x_j. We adopted the Euclidean distance function as suggested in ProtoNet [16]. For the target text x_j to be classified, we can get the probability of x_j using a softmax function over all the classes:

$$p(y = k|x_j) = \frac{exp(-d(f_m(x_j), c_k))}{\sum_{k'}^{K} exp(-d(f_m(x_j), c_{k'}))} \tag{1}$$

To train the model, during the meta-train stage, for a meta-train task $\mathcal{T}$ with K classes in support sets S and N_q query sample size in the query set Q, the loss is calculated using Formula (2):

$$Loss_T(Q) = \mathcal{L}(d(\{f_m(x_j^q)\}_{j=1}^{N_q}, \{c^k\}_{k=1}^{K}), y^q) \tag{2}$$

where:

$$d(x1, x2) = ||x1 - x2||_2 \tag{3}$$

Cross entropy loss is used for $\mathcal{L}$. We can observe that the classification of the target low-resource language is based on the distance between its features and those of different classes of the rich-resource source language. Thus, the knowledge transfer process between the source and the target language is explicitly modeled in the meta-training process where the multilingual model is trained to output representations that measure similarities between languages in terms of the degree of sexism.

3.2 Emotion Analysis

Cross-lingual hate speech detection methods suffer from unintended bias introduced by language-specific expressions and overfitting issues. We seek language-agnostic features that benefit the task while at the same time not affected by language-related bias, and emotion features serve as a good candidate to this end. Previous studies have shown the effectiveness of sentiment and emotion features in monolingual hate speech detection [17]. However, to our knowledge, no previous study has explored the effect of universal emotion features across different languages on the detection of hate speech. Although it has been reported that emotions can vary systematically in their meaning and experience across culture [18], a previous study showed that emotion semantics has a common underlying structure across languages [19], and empirical results showed that there are commonalities underlying the expression of emotion in different languages [20]. Thus, We develop a model to provide emotion classification under a common label space for multilingual sexist speeches. For each language ℓ, the emotion model f_e is trained on the emotion dataset $D_\ell = \{X_\ell, y\} = \{(x_\ell^i, y^i)\}_i^{N_\ell}$, where x_ℓ^i is the feature of sample i in the dataset, $y = \{y_1, y_2, \ldots, y_c\}$ is a common label space across the languages decided by the emotion taxonomy adopted. The model f_e is learned to minimize the binary cross entropy loss.

For a given sentence s_i in language ℓ, we first obtain its universal feature x_ℓ^i, and the model f_e provides its emotion vector with $v_i = sigmoid(f_e(x_\ell^i))$. After

training, we get a multi-label multi-class classifier which is later used to provide emotion knowledge for the sexist speeches in the corresponding languages.

3.3 Integration of Emotion Knowledge

We design an aggregation module to merge semantic and emotion features into multimodal features. Specifically, the aggregation module is composed of two parts: a feature concatenation module and a modality fusion module. The feature concatenation module works by concatenating together text embedding and emotion features indicated as in Formula (4).

$$z = Concat(v1, v2) = \{v1v2 \in \mathbb{R}^{m+n} : v1 \subset \mathbb{R}^m, v2 \subset \mathbb{R}^n\} \tag{4}$$

where m is the dimension of text features and n is the dimension of emotion features. The concatenated feature is then passed into the modality fusion module which is trained in an end-to-end way to translate the simply-concatenated vector into a joint representation. We use a convolutional layer for the fusion module as proposed in [21]. Given the multilingual model f_m, the emotion model f_e, and a sample x_i from either support set or query set, the aggregated features are produced as follows:

$$f_{agg}(x_i) = Conv(Concat(f_m(x_i), f_e(x_i))) \tag{5}$$

The joint representation is optimized with regard to the same loss given by Formula (2) :

$$Loss'_T(Q) = \mathcal{L}(d(\{f_{agg}(x_j^q)\}_{j=1}^{N_q}, \{c^k\}_{k=1}^{K}), y^q) \tag{6}$$

The Algorithm 1 illustrates the entire learning process.

4 Experiment

4.1 Datasets

We use publicly available datasets in English and Chinese. The broadly used Waseem dataset [2] is used to provide training data, and the English data of AMI EVALITA [4] is used to test the model's robustness. A recently published Chinese sexism dataset SWSR [8] is used for testing on Chinese data. For the training of emotion models, we use the GoEmotions dataset [22] which provided fine-grained emotion annotations for a large number of English texts collected from Reddit comments. We use the dataset provided by NLPCC-2013 emotion classification task [23] for Chinese emotion data. There exist other emotion datasets for Chinese, but they are either for other domains [24] or publicly not available [25]. For both Chinese datasets, the data are collected from Sina Weibo (microblog). We map emotion labels between the NLPCC-2013 dataset and GoEmotions dataset to a common label space based on emotion lexicon ontology [26].

Algorithm 1 Zero-Shot Cross-lingual Meta-learning with Emotion Features

Require: Multilingual Model f_m, Emotion Model f_e, Translation Model M_s^t, Training Set D_s with K classes in Resource-rich Language ℓ_s, Test Set D_t in Target Language ℓ_t, Aggregation Module AGG, Training Episodes Number N

$D'_s \leftarrow M_s^t(D_s)$ ▷ generate source for query sets
for i in $\{1, ..., N\}$ **do**
  for k in $\{1, ..., K\}$ **do** ▷ Iterate over training classes
    $S_i^k = D_s^i = \{(x_1, y_1), ...(x_j, y_j)\} \leftarrow RandomSample(D_s, j)$
    $Q_i^k = D_s'^i = \{(x'_1, y_1), ...(x'_q, y_q)\} \leftarrow RandomSample(D'_s, q)$
    $c_k = \frac{1}{|S_i^k|} \sum_{(x_i, y_i) \in S_i^k} f_{agg}(x_i)$
  end for
  $J \leftarrow 0$ ▷ Initiate Loss
  for k in $\{1, ..., K\}$ **do**
    $J \leftarrow J + Loss'_i(Q_i^k)$ ▷ Update Loss using Formula (6)
  end for
  Update all parameters $\theta_{f_m}, \theta_{f_{agg}}, \theta_{d'}$ w.r.t. J using gradient descent
end for
Do predictions on test set D_t using models with updated parameters.

Bias analysis of datasets Following the definition of unintended bias given by [27], we view expressions that affect the multilingual model's performance, such as language-specific taboo interjections, as false positive bias demonstrated by a disproportionate amount of samples containing certain terms. These terms appear in data labeled both as sexism and non-sexism, but the likelihood of the terms in sexism class is significantly higher than in the non-sexism class. Some of these terms may express the bias that the model should learn to distinguish between the two classes, while some may cause unintended behavior of the model, resulting in the model tending to classify some comments containing particular terms as sexism even if these terms do not convey such meaning. Besides, in the context of cross-domain dataset evaluation, the marginal distribution shift between datasets could lead to performance drop [28], which is also the case in the context of cross-lingual learning. Thus, we identify terms that distribute disproportionately in the sexism and non-sexism category and compare between datasets in English **(EN)** and Chinese **(CN)**.

We calculate the likelihood of a term w_i given the label as $p(w_i|label)$. To compute the degree of bias r, we use Formula (7):

$$r = \frac{p(w_i|sexism)}{p(w_i|nonsexism)} \tag{7}$$

Then a threshold is set to identify a set of terms that are disproportionately distributed. From terms above the threshold, we manually pick meaningful terms and analyze them qualitatively. Term analysis results are shown in Table 1.

We can observe that there are overlaps between the two datasets, mainly on terms expressing meanings related to feminism and gender. There exist non-sexist terms strongly linked with sexism in datasets of both languages, which

Table 1: Term bias analysis of EN-CN sexism datasets in terms of r (Formula 7). The term "gay", annotated with *, appears in English in the original Chinese text.

Term (EN)	r	**Term (CN)**	r
sexist	28.27	婚驴 (marriage donkey)	7.30
sport	27.91	gay*	7.19
female	17.63	男权 (patriarchy)	6.49
bitch	12.67	女拳 (negative feminism)	5.98
equal	12.01	伪女权 (fake feminism)	5.70
feminism	9.78	男人 (man)	4.76
blond	7.30	奴隶 (slave)	4.33
woman	6.61	洗脑 (brain washing)	4.25
dumb	6.48	彩礼 (bride price)	4.23
drive	5.74	女人 (woman)	4.18
man	4.57	职场 (work place)	4.03

could lead to potential unintended bias either in mono-lingual setting or cross-lingual setting. For example, the term "sport" is shown to have a significantly higher likelihood to appear in English tweets labeled as sexism, but the term is neutral and should not convey any bias. There are also language-specific terms which are more likely due to the cultural difference intrinsic to the language that can also harm the cross-lingual transfer learning performance of the model.

Emotion analysis of the dataset We analyze emotion features in the Chinese and English sexism datasets. For each sample in the datasets, we employ the prediction of the emotion model to generate an emotion feature. The emotion feature has eight dimensions as shown in Figure 2 which we use as a real-valued vector for later analysis.

We set a threshold to the emotion vector to decide if an emotion appears in the sample and count the frequencies of emotions. We normalized these frequencies to be the probability distribution, which is on a scale of 0 to 1, considering the fact that the sizes of datasets are different. The result is shown in Figure 2. To gain a better perspective, we set frequency values to be negative for negative emotions (disgust, fear, sadness, anger), and negative emotions are shown in the left part of the figure.

We observe that in both languages, non-sexist speech tends to have more positive emotions than sexist speech. In Chinese datasets, a large part of the non-sexist speeches still conveys negative emotions. The observation is consistent with the dataset's keyword-based construction method, where controversial contents are more likely to be selected. In addition, many speeches could be hateful, thus conveying more negative emotions, but they may not be sexism related. As a result, using emotion features independently for sexism or non-sexism may not be a good method to conduct cross-lingual transfer for sexist speech detection. We also observe a notable difference in emotions between sexism and non-sexism classes, which is in line with our previous assumption.

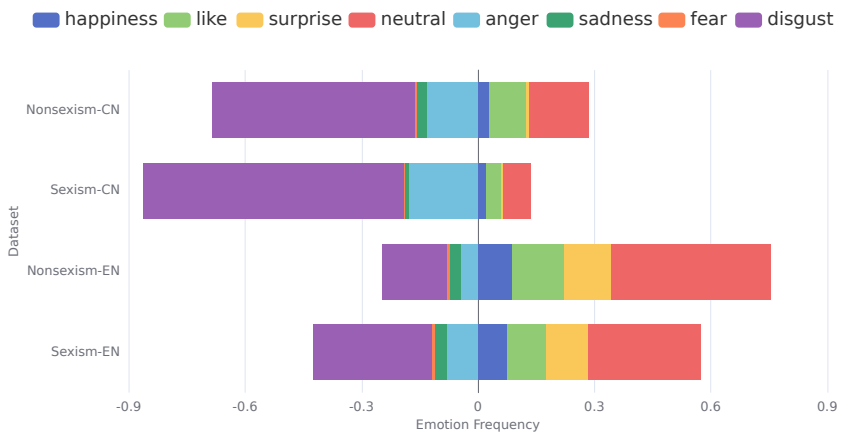

Fig. 2: Emotion Analysis of EN-CN sexism datasets. Frequency values are set to be negative for negative emotions (disgust, fear, sadness, anger) and are shown in the left part of the figure.

4.2 Experiment Settings

Baseline Models For zero-shot baselines, we re-implement previously reported best-performing models on cross-lingual hate speech detection [9, 11]: logistic regression with LASER embedding (**LASER+LR**); monolingual RoBERTa-base, with Chinese target data translated to English by machine translation API[1] (**RoBERTa-translation**); XLM-RoBERTa-base (**XLM-R**). For comparison with the fully supervised method, we implement a strong baseline for hate speech detection, **BERTCNN** [29] and also use the best model reported in the Chinese sexism dataset paper [8], a RoBERTa model trained with sexism lexicon knowledge (**RoBERTa Lexicon**).

Implementation Details We implement the meta-learning model using the PyTorch library. XLM-R is used as the backbone model to provide universal encodings. The support set and query set sample sizes N_s and N_q are set to be 32 and 10, respectively. During the meta-train phase, the number of episodes is 1000 and the learning rate is 1e-5. For model evaluation, in the fully-supervised setting, 10% of randomly sampled Chinese data is used as a test set; in the zero-shot setting, all of the Chinese data is used as the test set. The results are reported on an average of over 5 runs.

4.3 Experiment Results

Model Performance The overall performance of the baselines and the proposed models (in bold) are reported in Table 2. The results indicate that the

[1] https://cloud.tencent.com/document/api/551

Table 2: Model performances with all metrics tested on the SWSR Chinese sexism dataset. For fully supervised baselines, the models are trained using the SWSR dataset; for zero-shot models, the English Waseem dataset is used to provide source data for training, and metrics are tested on the SWSR dataset. For the robustness test, we change the source data to AMI dataset for zero-shot models. F1-sex and F1-not indicate F1 score for sexism and non-sexism class

Model	**F1-sex**	**F1-not**	**Macro F1**	**Accuracy**
Fully supervised baselines, monolingual				Source data: SWSR dataset (CN)
BERTCNN	0.721	0.834	0.778	0.792
RoBERTa Lexicon	0.707	0.853	0.780	0.804
Zero-shot cross-lingual models				Source data: Waseem dataset (EN)
LASER+LR	0.409	0.804	0.607	0.706
XLM-R	0.612	0.743	0.671	0.692
RoBERTa translation	0.398	0.782	0.591	0.681
ProtoNet	0.592	0.763	**0.681**	0.701
ProtoNet Emotion	0.602	0.788	**0.691**	0.713
Zero-shot cross-lingual models				Source data: AMI dataset (EN)
LASER+LR	0.491	0.803	0.647	0.716
XLM-R	0.538	0.785	0.662	0.707
RoBERTa translation	0.548	0.773	0.663	0.702
ProtoNet	0.637	0.737	**0.687**	0.695
ProtoNet Emotion	0.659	0.749	**0.706**	0.703

proposed meta-learning method using ProtoNet and ProtoNet with emotion improves over the previous zero-shot methods and shows a more stable performance, although there is still a drop in performance compared with the best-performing supervised models. The performance of the baseline models indicates that for the sexism detection task, even for languages strongly different from each other (Chinese-English), the zero-shot cross-lingual methods still yield comparable performance compared to previous cross-lingual settings where those focused languages are from the same or closer language family (e.g., English-Spanish, English-Italian). We observe that XLM-R demonstrates a good multilingual ability and outperforms the method using monolingual model with translation data. The proposed meta-learning method using ProtoNet alone achieves good performance, and the addition of emotion knowledge gains a marginal improvement over the ProtoNet model's overall improvement on F1 and accuracy. Specifically, we observe a notable improvement of the F1 score over non-sexism class with ProtoNet Emotion, which may explain the effect of emotion features in mitigating the unintended bias introduced by non-sexism terms strongly linked with the sexism class.

Robustness Analysis We analyze the robustness and generalization of the proposed method by changing the domain of the training dataset. Specifically, we use the English misogyny dataset provided by AMI EVALITA 2018 shared task [4] and test if the model's performance remains stable on the test data. Compared to the Waseem dataset, the AMI dataset contains fewer data points

Table 3: Examples of correctly and wrongly classified samples. The texts are translated from Chinese.

ID	Text	True Label	Predicted Label		
			XLM-R	ProtoNet	+ Emotion
#1	There's nothing wrong about feminism itself, what's wrong with pursuing gender equality?	non	sexism	sexism	non
#2	...You can't sexually harass people just because they are gay.	non	sexism	sexism	non
#3	Feminism is not against men! It's against the patriarchal society!	non	sexism	sexism	sexism
#4	The person seems to be a recidivist... and should be prosecuted!:rage: :rage:	non	sexism	non	sexism
#5	Enoki mushroom like you will only rot on the shelf even if clearly priced.	sexism	non	non	non
#6	The unmarried donkeys sounds happy hahaha	sexism	non	non	non
#7	Boys want to work with you only because they want to use you to their advantage.	sexism	non	non	non
#8	...when girls are older, it becomes more difficult to find a partner.	sexism	non	non	non

for both sexism and non-sexism classes but is more comprehensive regarding the types of sexism. The result is shown in Table 2. For the zero-shot baseline models, changing the training dataset has a large impact on the performance of the models, which suggests that these methods tend to be more easily affected by the domain and the content of the datasets. With the meta-learning method, the model's overall performance is more stable, and the F1 score and accuracy of the model are close between the two datasets.

4.4 Case Studies

To better understand the results of the cross-lingual sexism detection, we conducted a qualitative study of cases where the proposed method leads to the correct classification of previously misclassified examples and cases where the proposed method has failed to classify sexist speeches correctly. Specifically, we search with non-hateful terms that are strongly linked with sexism class as shown in table 1. The results are shown in table 3, the texts are translated from Chinese.

We observe that the common reasons that cause errors are as follows: (i) biased terms; (ii) specific expression in the target language; (iii) implicit or benevolent sexism. In a few cases, the integration of emotion knowledge helps out to correct examples wrongly classified as sexism due to biased terms, such as #1 and #2. However, we also observe that the bias mitigation effect brought about by emotion knowledge injection is limited and there are still a considerable number of examples misclassified as sexism class due to the presence of biased terms, such as #3, and in some cases, the integration of emotion knowledge harms the prediction such as #4. This may be because the emotion models are not accurate enough and the emotion label space is not expressive enough. Some specific expressions are also hard to detect and require culturally related a priori knowledge to identify, such as Internet slang terms in #5 and #6 which are sexist and appear mostly in social media content. Sexist speeches that do not contain explicit sexist terms or convey benevolent sexism are also hard to detect.

5 Conclusion

In this paper, we explore the cross-lingual method to detect Chinese sexism. We propose to use meta-learning method for zero-shot cross-lingual sexist speech detection and to integrate emotion knowledge about sexism datasets in the meta-learning framework. Our proposed method with ProtoNet and ProtoNet Emotion improves over previous cross-lingual zero-shot methods and achieve new state-of-the-art. We also observed that the proposed method is still limited in dealing with issues related to cultural factors which could be reflected by language-specific expressions such as Internet slang terms. Our proposed method can be easily extended to other low-resource languages, and in the future works we wish to experiment the method with more languages and seek to better deal with problems caused by cultural factors in cross-lingual hate speech detection.

References

1. Samory, M., Sen, I., Kohne, J., Flöck, F., Wagner, C.: Call me sexist, but...: Revisiting sexism detection using psychological scales and adversarial samples. In: Intl AAAI Conf. Web and Social Media. pp. 573–584 (2021)
2. Waseem, Z., Hovy, D.: Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In: Proceedings of the NAACL student research workshop. pp. 88–93 (2016)
3. Fersini, E., Rosso, P., Anzovino, M.: Overview of the task on automatic misogyny identification at ibereval 2018. Ibereval@ sepln **2150**, 214–228 (2018)
4. Fersini, E., Nozza, D., Rosso, P.: Overview of the evalita 2018 task on automatic misogyny identification (ami). EVALITA Evaluation of NLP and Speech Tools for Italian **12**, 59 (2018)
5. Pamungkas, E.W., Basile, V., Patti, V.: Misogyny detection in twitter: a multilingual and cross-domain study. Information Processing & Management **57**(6), 102360 (2020)
6. Glick, P., Fiske, S.T.: Ambivalent sexism. In: Advances in experimental social psychology, vol. 33, pp. 115–188. Elsevier (2001)
7. Jha, A., Mamidi, R.: When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data. In: Proceedings of the second workshop on NLP and computational social science. pp. 7–16 (2017)
8. Jiang, A., Yang, X., Liu, Y., Zubiaga, A.: Swsr: A chinese dataset and lexicon for online sexism detection. Online Social Networks and Media **27**, 100182 (2022)
9. Pamungkas, E.W., Basile, V., Patti, V.: A joint learning approach with knowledge injection for zero-shot cross-lingual hate speech detection. Information Processing & Management **58**(4), 102544 (2021)
10. Jiang, A., Zubiaga, A.: Cross-lingual capsule network for hate speech detection in social media. In: Proceedings of the 32nd ACM Conference on Hypertext and Social Media. pp. 217–223 (2021)
11. Aluru, S.S., Mathew, B., Saha, P., Mukherjee, A.: A deep dive into multilingual hate speech classification. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 423–439. Springer (2020)
12. Nozza, D.: Exposing the limits of zero-shot cross-lingual hate speech detection. In: Proceedings of the 59th Annual Meeting of the Association for Computational

Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). pp. 907–914 (2021)
13. Arango, A., Pérez, J., Poblete, B.: Hate speech detection is not as easy as you may think: A closer look at model validation. In: Proceedings of the 42nd international acm sigir conference on research and development in information retrieval. pp. 45–54 (2019)
14. Nooralahzadeh, F., Bekoulis, G., Bjerva, J., Augenstein, I.: Zero-shot cross-lingual transfer with meta learning. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 4547–4562 (2020)
15. Xu, W., Haider, B., Krone, J., Mansour, S.: Soft layer selection with meta-learning for zero-shot cross-lingual transfer. In: Proceedings of the 1st Workshop on Meta Learning and Its Applications to Natural Language Processing. pp. 11–18 (2021)
16. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. Advances in neural information processing systems **30** (2017)
17. Chiril, P., Pamungkas, E.W., Benamara, F., Moriceau, V., Patti, V.: Emotionally informed hate speech detection: a multi-target perspective. Cognitive Computation **14**(1), 322–352 (2022)
18. Mesquita, B., Boiger, M., De Leersnyder, J.: The cultural construction of emotions. Current opinion in psychology **8**, 31–36 (2016)
19. Jackson, J.C., Watts, J., Henry, T.R., List, J.M., Forkel, R., Mucha, P.J., Greenhill, S.J., Gray, R.D., Lindquist, K.A.: Emotion semantics show both cultural variation and universal structure. Science **366**(6472), 1517–1522 (2019)
20. Lamprinidis, S., Bianchi, F., Hardt, D., Hovy, D.: Universal joy: A data set and results for classifying emotions across languages. In: The 16th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics (2021)
21. Ma, L., Lu, Z., Shang, L., Li, H.: Multimodal convolutional neural networks for matching image and sentence. In: Proceedings of the IEEE international conference on computer vision. pp. 2623–2631 (2015)
22. Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., Ravi, S.: Goemotions: A dataset of fine-grained emotions. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 4040–4054 (2020)
23. Yao, Y., Wang, S., Xu, R., Liu, B., Gui, L., Lu, Q., Wang, X.: The construction of an emotion annotated corpus on microblog text. Journal of Chinese Information Processing **28**(5), 83–91 (2014)
24. Quan, C., Ren, F.: A blog emotion corpus for emotional expression analysis in chinese. Computer Speech & Language **24**(4), 726–749 (2010)
25. Li, M., Long, Y., Qin, L., Li, W.: Emotion corpus construction based on selection from hashtags. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). pp. 1845–1849 (2016)
26. Xu, L., Lin, H., Pan, Y., Ren, H., Chen, J.: Constructing the affective lexicon ontology. Journal of the China society for scientific and technical information **27**(2), 180–185 (2008)
27. Dixon, L., Li, J., Sorensen, J., Thain, N., Vasserman, L.: Measuring and mitigating unintended bias in text classification. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. pp. 67–73 (2018)
28. Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., Vaughan, J.W.: A theory of learning from different domains. Machine learning **79**(1), 151–175 (2010)
29. Mozafari, M., Farahbakhsh, R., Crespi, N.: Hate speech detection and racial bias mitigation in social media based on bert model. PloS one **15**(8), e0237861 (2020)