Vicinal Risk Minimization for Few-Shot Cross-lingual Transfer in Abusive Language Detection

Gretel Liz De la Peña Sarracén [1] Paolo Rosso [1] Robert Litschko [2]
Goran Glavaš [3] Simone Paolo Ponzetto [4]

[1]Universitat Politècnica de València, [2]MaiNLP, LMU Munich, [3]CAIDAS, University of Würzburg, [4]DWS Group, University of Mannheim

Overview

This work resorts to data augmentation and continual pre-training for domain adaptation to improve cross-lingual abusive language detection. For data augmentation, we analyze two existing techniques based on vicinal risk minimization and propose MIXAG, a data augmentation method that interpolates pairs of instances based on the angle between their representations.

Contributions

- **Dataset extension:** We extend the multilingual dataset we used in the experiments by including the corresponding Spanish dataset.
- **Few-shot cross-lingual transfer learning improvement at data-level:** We rely on Vicinal Risk Minimization principle to generate synthetic samples in the vicinity of the training samples to increase the amount of information to fine-tune the model in the target language.
- **Unsupervised language adaptation:** We simulate a fully unsupervised setup, removing the label information from the target languages. In this setting, we make a domain adaption for abusive terms via masked language modeling in the target language before a zero-shot transfer.

Vicinal Risk Minimization (VRM)

Data augmentation as an extension of the training set $D_{train} = \{(x_i, y_i)\}$ by drawing samples from a neighborhood of the existing samples [1].
The distribution $p(x, y)$ is approximated by a vicinity distribution $D_v = \{(\hat{x}_i, \hat{y}_i)\}_{i=1}^{n}$, whose instances are a function of the instances of D_{train}.
Vicinal risk (R_v) is then calculated on D_v:

$$R_v = \frac{1}{n}\sum_{i=1}^{n} l(f(\hat{x}_i), \hat{y}_i) \quad (1)$$

Techniques in NLP

- **SSMBA**: Pair of functions (Corruption and Reconstruction)
- **MIXUP**: Constructs a synthetic example in the vicinity distribution from the linear combination of examples.

Our technique: MIXAG

Constructs a synthetic example from the combination of instances with a focus on the angle (α) between their representations.

$$\hat{x} = \lambda x_i + x_j \quad (2)$$

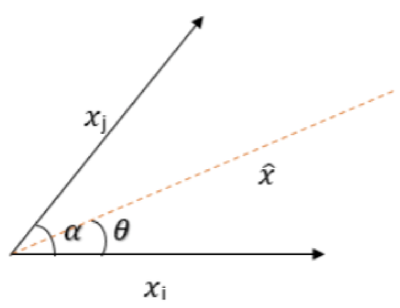
$$\hat{y} = \lambda y_i + y_j \quad (3)$$

Using the Law of Sines we express the linear combination coefficient (λ) as a function of the cosine of α.

$$\lambda = \frac{||x_j||(cos(\theta)\sqrt{1-cos(\alpha)^2} - cos(\alpha)\sqrt{1-cos(\theta)^2})}{||x_i||\sqrt{1-cos(\theta)^2}} \quad (4)$$

$$cos(\alpha) = \frac{x_i x_j}{||x_i||||x_j||} \quad (5)$$

θ is a parameter.



Experiments

Dataset: XHate-999 [2]: Several variants of abusive language detection.

- 3 domains: Fox News (GAO), Twitter/Facebook (TRAC), and Wikipedia (WUL).
- Texts in English for training, validation, and testing.
- Test instances in 5 target languages: Albanian (SQ), Croatian (HR), German (DE), Russian (RU), and Turkish (TR). We extended with test instances in Spanish (ES).

Model: mBERT
Fine-tuning and Evaluation Details: For each language

- Zero-shot experiments: 90% of the test set to evaluate.
- Few-shot experiments: 10% of the test set to fine-tune the model and 90% of the test set to evaluate.
- Unsupervised language adaptation: Two-step methodology:
  - 1. continual pre-training for domain adaptation via masked language modeling.
  - 2. zero-shot learning to detect abusive language.

Variants: **ZS:** Zero-shot, **FS:** Zero-shot, **SS:** SSMBA, **SS-HL:** SSMBA with HurtLex, **MU:** MIXUP, **MMU:** multilingual MIXUP, **MU-SS:** MIXUP with SSMBA, **MA:** MIXAG, **MMA:** multilingual MIXAG, **MA-SS:** MIXAG with SSMBA, **ZS_MLM:** language adaptation

Results

The results are reported in terms of F1 and significantly better results are underlined for each language and domain ($\alpha = .05$). Numbers in bold indicate the best results.

Multidomain results

ALL	EN	DE	RU	TR	HR	SQ	ES
ZS	0.8085	0.7156	0.6308	0.3627	0.6214	0.6127	0.6008
FS	0.8112	0.7141	0.6329	0.4063	0.6316	0.6238	0.6130
SS	0.8077	0.7253	0.7071	0.6568	0.6965	0.6990	0.6838
SS-HL	0.8097	0.7273	0.6987	0.6689	0.6725	0.6909	0.6973
MU	0.8102	0.7404	0.7013	0.6740	0.7116	0.7001	0.6878
MMU	**0.8284**	0.7500	0.7312	0.7113	0.7371	0.7128	0.7250
MU-SS	0.8176	0.7531	0.7233	0.6839	0.7186	0.6881	0.7087
MA	0.8083	0.7245	0.6757	0.5616	0.6710	0.6788	0.6508
MMA	0.8237	**0.7585**	**0.7392**	**0.7224**	**0.7523**	**0.7344**	**0.7476**
MA-SS	0.8096	0.7229	0.7193	0.6369	0.6759	0.6734	0.6713

Results in language adaptation

GAO	EN	DE	RU	TR	HR	SQ	ES
ZS	**0.6747**	0.5067	0.5205	0.5116	0.6234	0.5405	0.5263
ZS_MLM	0.6050	**0.6364**	**0.6261**	**0.6341**	**0.6290**	**0.6016**	**0.6154**
TRAC							
ZS	**0.7642**	**0.7582**	**0.6815**	**0.6777**	**0.6892**	**0.7235**	**0.7000**
ZS_MLM	0.6821	0.6480	0.6718	**0.6785**	0.6785	0.6995	0.6118
WUL							
ZS	**0.8800**	0.6698	0.5561	0.2945	0.5469	0.5556	0.4960
ZS_MLM	0.6093	**0.6765**	**0.6708**	**0.6765**	**0.6732**	**0.6675**	**0.6765**
ALL							
ZS	**0.8085**	**0.7156**	0.6308	0.3627	0.6214	0.6127	0.6008
ZS_MLM	0.6662	0.6711	**0.6637**	**0.6716**	**0.6716**	**0.6721**	**0.6419**

Findings

- VRM-based techniques improve few-shot cross-lingual transfer.
- In results by domain: Unclear difference between the performance of VRM-based techniques.
- In multidomain experiments: Multilingual MIXAG outperforms the other strategies.
- Domain adaptation can improve zero-shot cross-lingual transfer, but few-shot cross-lingual transfer with VRM-based techniques seems to be more robust.

References

[1] O. Chapelle, J. Weston, L. Bottou, and V. Vapnik. Vicinal risk minimization. *Advances in neural information processing systems*, 13, 2000.

[2] G. Glavaš, M. Karan, and I. Vulić. XHate-999: Analyzing and detecting abusive language across domains and languages. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6350–6365, Barcelona, Spain (Online), Dec. 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.559. URL `https://aclanthology.org/2020.coling-main.559`.