Hate Speech Detection in Low Resource Languages using Synthetic Data Generation

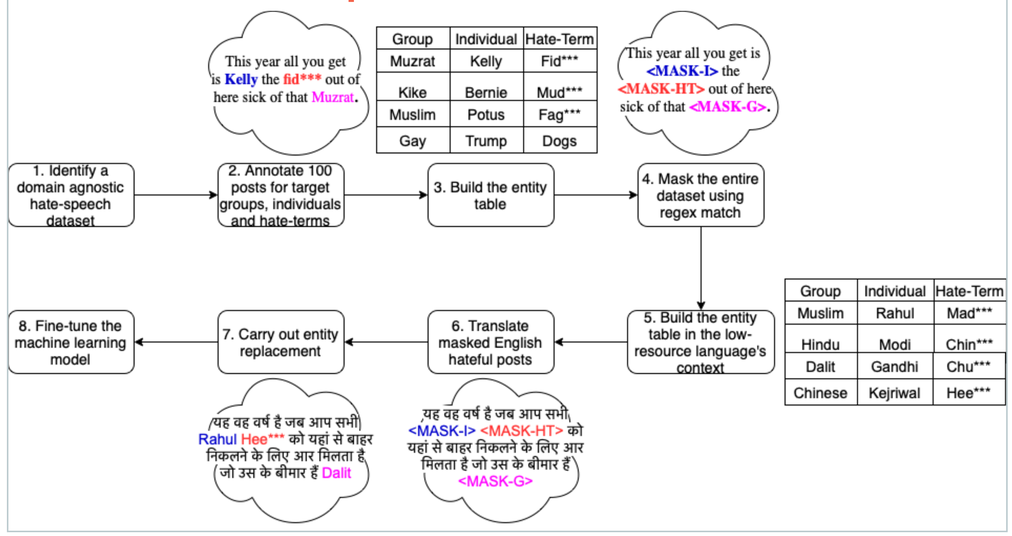
Aman Khullar, Daniel K. Nkemelu, Cuong V. Nguyen, Michael L. Best

T+ID Lab, Georgia Institute of Technology

Motivation

- We have limited or no hate speech data to train machine learning models in low-resource languages
- Can we generate synthetic data to train and boost performance of hate speech detection models?
- Possible to translate hate speech posts from a high resource language like English
- Simple translation does not take context shift into account
- How can we make synthetic data more natural and diverse?
- Can large generative models offer more respite?

Proposed Framework



Experiments

1. Augment low-resource language with translated posts
2. Augment low-resource language with contextually adapted synthetic posts
3. Augment low-resource language with large language model generated hateful posts
4. Compare the performance of fine-tuned mBERT model in Hindi and Vietnamese

Results

Hindi

Model	LR non-hateful	LR hateful	Synthetic hateful	Macro F1
No-Aug	450	100	0	84.46
Translated Aug	450	100	350	84.40
Context Translated Aug	450	100	350	**85.99**
LR Aug	450	450	0	87.61

Vietnamese

Model	LR non-hateful	LR hateful	Synthetic hateful	Macro F1
No-Aug	2500	250	0	66.42
Translated Aug	2500	250	2250	66.73
Context Translated Aug	2500	250	2250	**67.01**
LR Aug	450	2250	0	71.70

BLOOM and Diversified Test Set

BLOOM LM - Hindi

Model	LR non-hateful	LR hateful	Synthetic hateful	Macro F1
No-Aug	450	100	0	84.46
Translated Aug	450	100	100	84.07
Context Translated Aug	450	100	100	84.52
BLOOM Aug	450	100	100	**86.14**
LR Aug	450	200	0	86.85

Diversified Test Set - Hindi

Model	LR non-hateful	LR hateful	Synthetic hateful	Macro F1
Translated Aug	450	100	100	45.30
Context Translated Aug	450	100	100	45.80
BLOOM Aug	450	100	100	**46.19**
LR Aug	450	200	0	48.19

Qualitative Study

Original Hateful Post: this ugly k** c** keeps showing up on my timeline

Translated Hateful Post: इस बदसूरत कि** योनी रहता है ऊपर दिखा रहा है पर मेरे समय रेखा

Contextual Translated Hateful Post: यह कमी** बिहारी ल** मेरी टाइमलाइन पर दिखाई देता रहता है

BLOOM: अगर तुम मुसलमान हो तुम अपराधी हो, बात खतम!

Summary

Our work proposes a new way of boosting performance of hate speech detection models in low-resource languages through generation of more natural and diverse synthetic hateful posts in the low-resource language.

Acknowledgement

We would like to thank our colleagues at the T+ID Lab for their feedback and our partners for funding this work.

T+IDLab Technologies & International Development Lab · Georgia Tech · The Carter Center · Microsoft

akhullar8@gatech.edu