NLPDove at SemEval-2020 Task 12: Improving Offensive Language Detection with Cross-lingual Transfer

Hwijeen Ahn[†] Jimin Sun[‡] Chan Young Park[§] Jungyun Seo[†]
{hwijeen, seojy}@sogang.ac.kr, jiminsun@dm.snu.ac.kr, chanyoun@cs.cmu.edu
[†]Sogang University [‡]Seoul National University [§]Language Technologies Institute, Carnegie Mellon University

Contributions

- **Competitive Results**: Our multilingual systems achieved competitive results in Greek, Danish, and Turkish at OffensEval 2020.
- **Preprocessing**: We introduce preprocessing steps tailored to social media text with practical methods to fine-tune multilingual BERT (mBERT) for offensive language identification.
- **Data Augmentation**: We investigate two data augmentation strategies. 1) using additional semi-supervised labels with different thresholds and 2) cross-lingual transfer with data selection.
- **Cross-lingual Transferability Metric**: We propose a new metric, Translation Embedding Distance (TED), to quantify cross-lingual transferability and analyze how data selection with TED increases performance.

Results

language	Rank	NLPDove F1	Best Model F1
Greek	1/37	0.85	0.85
Danish	3/39	0.79	0.81
Turkish	5/46	0.80	0.83
English	15/85	0.91	0.92
Arabic	20/53	0.80	0.90

Preprocessing

Five preprocessing methods and the number of examples modified by each method in the English training data.

Preprocessing	# of Cases
Emoji substitution	1454
Hashtag segmentation	2290
Letter casing normalization	12665
URL replacement	2140
Punctuation trimming	504
Total number of instances	12691

Proposed Model

- Multilingual BERT (mBERT) based classifier
- Ensemble of multiple mBERT models
  - Pooling method (mean, max, concat (mean;max), CNN)
  - Representation layer (last, second-to-last)
  - Training data (original, cross-lingual data transfer, augmenting with semi-supervised data)

Cross-lingual Data Transfer

- **Translation Embedding Distance (TED)** is defined as L2 distance between the mBERT representation of the original sentence (s_{tf}) and machine translated sentence into the target language ($s_{tf \to tg}$).
  - tf: transfer language, tg: target language, D_{tf}: transfer language dataset
  - Instance-level: $\texttt{TED}_{\texttt{inst}}(s_{tf}, tg) = \|\text{mBERT}(s_{tf}) - \text{mBERT}(s_{tf \to tg})\|_2$
  - Language-level: $\texttt{TED}_{\texttt{lang}}(tf, tg) = \sum_{s \in D_{tf}} \frac{\texttt{TED}_{\texttt{inst}}(s, tg)}{|D_{tf}|}$
- We hypothesize that ***transferability*** ~ ***translatability***; if a sentence is easily translated into the target language, it is more likely to be a transferable training sample.

Train data	da	$da + en_{all}$	$da + en_{top}$	$da + en_{bottom}$	$da + en_{rand}$
Macro F1	0.77	0.78	**0.84**	0.72	0.72

Danish validation set performance under different data transfer conditions

- $da + en_{all}$: all English samples in addition to Danish samples.
  $da + en_{top}$: adds 1.3K samples from English with smallest TED
  $da + en_{bottom}$: adds 1.3K samples with largest TED
  $da + en_{rand}$: adds randomly chosen 1.3K samples.

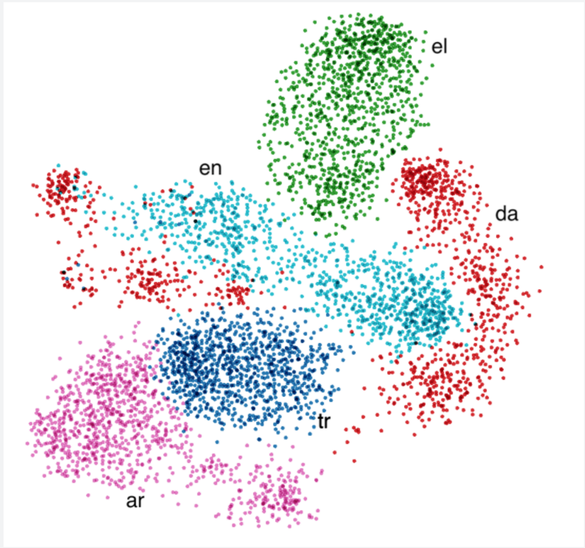
Why instance level TED works

Original (English)	**Translated (Danish)**
@USER Creepy	@USER Creepy
@USER Fool!	@USER Fool!
#RestoreHumanity #AntiFa September 22: Stop the fascist NVU in #Amsterdam: URL HT @USER	#RestoreHumanity #AntiFa September 22: Stop det fascistiske NVU in #Amsterdam: URL HT @USER
@USER I'M SO FU*KING READY	@USER I'M SO FU*KING KLAR
@USER @USER so sad #taketworeferenceswithmoniqueandchloe x	@USER @USER så trist #taketworeferenceswithmoniqueandchloe x

Samples with low TED scores

- These samples did not change significantly after translation procedure
- Often contain words used in both languages colloquially (e.g., creepy, fool)
- Short with a rather simple grammatical structure
- We posit that the simplistic nature of low TED samples made themselves more translatable, thus more transferable, and led to improvements via cross-lingual transfer.

Why language level TED works



- mBERT representation of examples form rough clusters based on their language.
- Proximity in the embedding space may explain the improvement yielded when English samples with low TED were added to the Danish dataset.

Augmenting Semi-supervised Data

- Thresholding the data with average prediction confidence and standard deviation to filter out noisy data

Threshold OFF	Threshold NOT	Threshold std	Additional data size	Macro F1
0.8	0.2	0.1	3533	0.7693
0.8	0.2	0.125	29681	**0.7793**
0.8	0.3	0.1	7956	0.7716
0.8	0.3	0.125	83782	0.7738
0.9	0.2	0.1	2769	0.7741
0.9	0.2	0.125	19590	0.7709
0.9	0.3	0.1	7192	0.7707
0.9	0.3	0.125	73691	0.7453
Baseline			0	0.7754

English validation set performance

Future Directions

- Measuring transferability without an MT module
- Reducing the model's over-sensitivity to word-level signals (offensive words), disregarding overall context