XED

A Multilingual Dataset for Sentiment Analysis and Emotion Detection

Emily Öhman Marc Pàmies Kaisla Kajava Jörg Tiedemann

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI
HUMANISTINEN TIEDEKUNTA
HUMANISTISKA FAKULTETEN
FACULTY OF ARTS

➔ **Manually annotated multilabel emotion detection datasets for English (30k) and Finnish (25k).**

➔ **Projected annotations for 30 additional languages.**

➔ Annotated English and Finnish movie subtitles from OPUS (Lison & Tiedemann, 2016).

➔ The 8 core emotions from Plutchik's (1980) wheel of emotions as the basis for our annotation scheme.

➔ 30k English and 20k Finnish manual annotations.

➔ Projected annotations for 30 additional languages with over 950 lines for all and over 10k for some (Figure 2).

Number of annotations:	24,164 + 9,384 neutral
Number of unique data points:	17,520 + 6,420 neutral
Number of emotions:	8 (+pos, neg, neu)
Number of annotators:	108 (63 active)
Number of labels per data point:	1: 78% 2: 17% 3: 4% 4+: 0.8%

Table 1. Statistics for the English annotations.

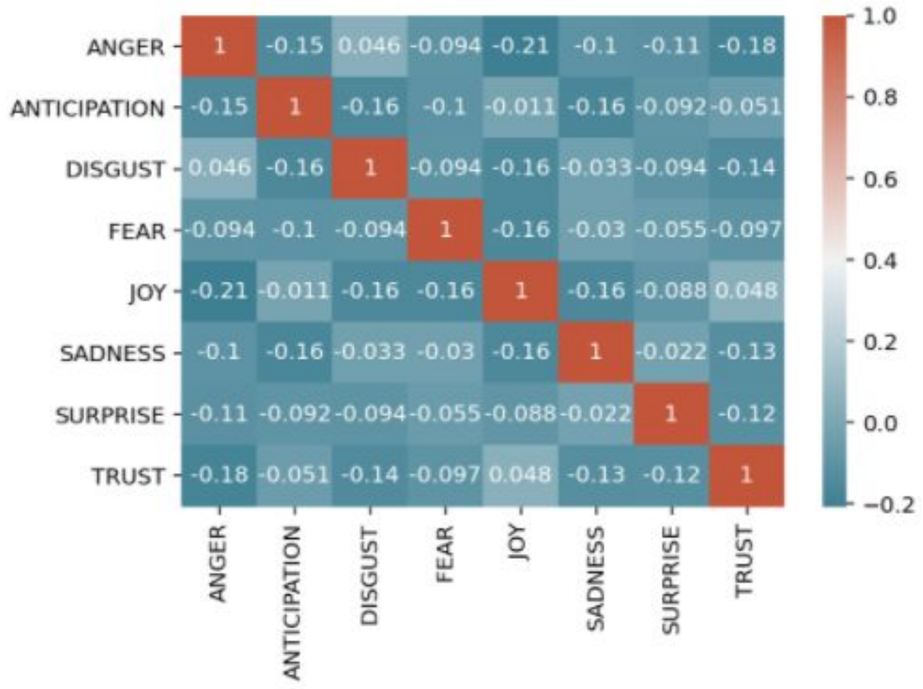

Figure 1. Correlation matrix for the English dataset.

➔ We use movie subtitles as a multi-domain proxy in the hopes that this would enable cross-domain use of XED.

➔ Comparisons with other similar datasets and lexicons suggest the source data (movie subtitles) influences the emotion label distribution to some degree.

➔ Some emotions are more likely to occur together than others (see also Figure 1):

- ➢ *anger* and *disgust*
- ➢ *joy*, *anticipation*, and *trust* in all permutations
- ➢ *anger* & *anticipation*, *sadness* & *surprise*, *fear* & *sadness*

AR	BG	BS	CN	CS	DA	DE	EL	ES	ET	
0.5729	0.6069	0.5854	0.5004	0.6263	0.5989	0.6059	0.6192	0.6760	0.5449	
FI	FR	HE	HR	HU	IS	IT	MK	NL	NO	
0.5859	0.6257	0.5980	0.6503	0.5978	0.5416	0.6907	0.4961	0.6140	0.5771	
PL	PT	PT_BR	RO	RU	SK	SL	SR	SV	TR	VI
0.6233	0.6203	0.6726	0.6387	0.6976	0.5305	0.6015	0.6566	0.6218	0.6080	0.5594

Table 2. Baseline Linear SVC classification of projected datasets have macro f1 scores between 0.496 - 0.691.

data	f1
English without NER, BERT	0.530
English with NER, BERT	0.536
English NER with neutral, BERT	0.467
English NER binary + surprise, BERT	0.679
English NER true binary, BERT	0.838
English NER, one-vs-rest SVM*	0.746

Table 3. Evaluation of annotated English data

FinBERT	f1
annotated	0.507
projected	0.446

➔ Evaluations using BERT for the annotated English dataset, shows that NER improves f1 scores whereas the addition of neutral lowers accuracy. Mapping emotions into coarser categories improves accuracy.

➔ Finnish projected vs. annotated datasets evaluated using FinBERT showed that the projected annotations fared slightly worse than the manual annotations.

➔ Reliable emotion detection is a challenging task. It is not necessarily an issue with Natural Language Processing and Understanding as these types of tasks are challenging for human annotators as well. If human annotators cannot agree on labels, it is unreasonable to think computers can do any better regardless of the annotation scheme or model used since these models are restricted by human performance.

➔ XED is the first emotion detection dataset for many under-resourced languages.

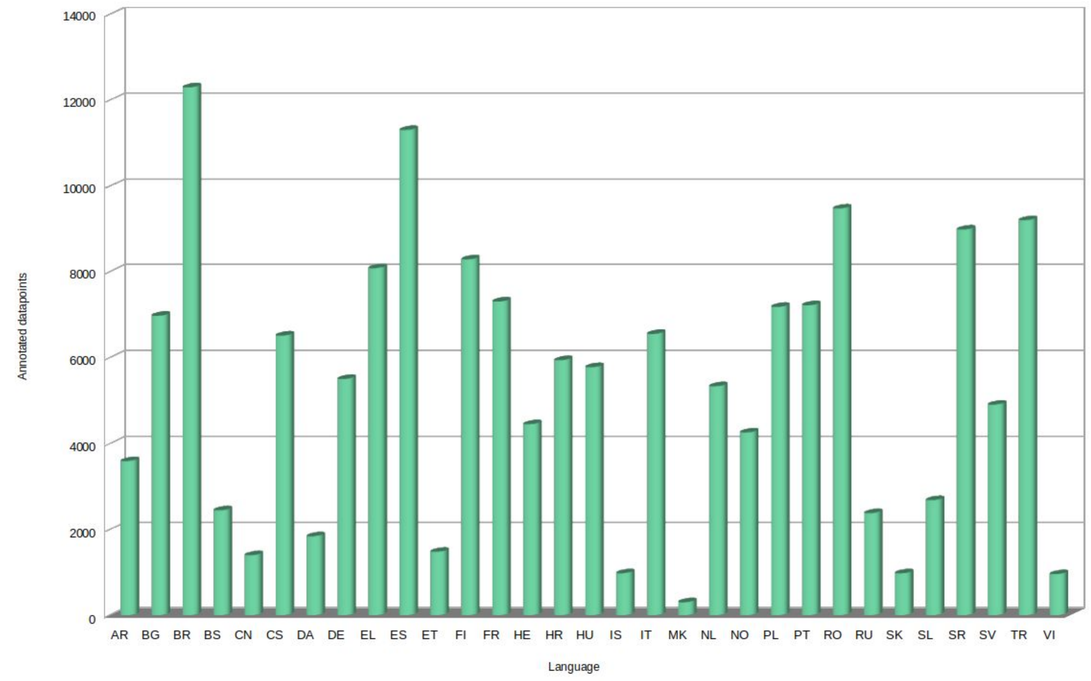

Figure 2. Projected dataset sizes for different languages.

References:

- Lison, P. and Tiedemann, J., 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles.
- Plutchik, R., 1980. A general psychoevolutionary theory of emotion. In Theories of emotion (pp. 3-33). Academic press.