XED:
A Multilingual Dataset for Sentiment Analysis and Emotion Detection

Emily Öhman and **Marc Pàmies** and **Kaisla Kajava** and **Jörg Tiedemann**
University of Helsinki
firstname.lastname@helsinki.fi

Abstract

We introduce XED, a multilingual fine-grained emotion dataset. The dataset consists of human-annotated Finnish (25k) and English sentences (30k), as well as projected annotations for 30 additional languages, providing new resources for many low-resource languages. We use Plutchik's core emotions to annotate the dataset with the addition of neutral to create a multilabel multiclass dataset. The dataset is carefully evaluated using language-specific BERT models and SVMs to show that XED performs on par with other similar datasets and is therefore a useful tool for sentiment analysis and emotion detection.

1 Introduction

There is an ever increasing need for labeled datasets for machine learning. This is true for English as well as other, often under-resourced, languages. We provide a cross-lingual fine-grained sentence-level emotion and sentiment dataset. The dataset consists of parallel manually annotated data for English and Finnish, with additional parallel datasets of varying sizes for a total of 32 languages created by annotation projection. We use Plutchik's Wheel of Emotions (*anger*, *anticipation*, *disgust*, *fear*, *joy*, *sadness*, *surprise*, *trust*) (Plutchik, 1980) as our annotation scheme with the addition of *neutral* on movie subtitle data from OPUS (Lison and Tiedemann, 2016).

We perform evaluations with fine-tuned cased multilingual and language specific BERT (Bidirectional Encoder Representations from Transformers) models (Devlin et al., 2019), as well as Suport Vector Machines (SVMs). Our evaluations show that the human-annotated datasets behave on par with comparable state-of-the-art datasets such as the GoEmotions dataset (Demszky et al., 2020). Furthermore, the projected datasets have accuracies that closely resemble human-annotated data with macro f1 scores of 0.51 for the human annotated Finnish data and 0.45 for the projected Finnish data when evaluating with FinBERT (Virtanen et al., 2019).

The XED dataset can be used in emotion classification tasks and other applications that can benefit from sentiment analysis and emotion detection such as offensive language identification. The data is open source[1] licensed under a Creative Commons Attribution 4.0 International License (CC-BY).

In the following sections we discuss related work and describe our datasets. The datasets are then evaluated and the results discussed in the discussion section.

2 Background & Previous Work

Datasets created for sentiment analysis have been available for researchers since at least the early 2000s (Mäntylä et al., 2018). Such datasets generally use a binary or ternary annotation scheme (positive, negative + neutral) (e.g. Blitzer et al. (2007)) and have traditionally been based on review data such as, e.g. Amazon product reviews, or movie reviews (Blitzer et al., 2007; Maas et al., 2011; Turney, 2002). Many, if not most, emotion datasets on the other hand use Twitter as a source and individual tweets as

[1] https://github.com/Helsinki-NLP/XED

Proceedings of the 28th International Conference on Computational Linguistics, pages 6542–6552
Barcelona, Spain (Online), December 8-13, 2020

level of granularity (Schuff et al., 2017; Abdul-Mageed and Ungar, 2017; Mohammad et al., 2018). In the case of emotion datasets, the emotion taxonomies used are often based on Ekman (1971) and Plutchik (1980) (which is partially based on Ekman).

2.1 Existing Emotion Datasets

Bostan and Klinger (2018) analyze 14 existing emotion datasets of which only two are multilabel. These are AffectiveText (Strapparava and Mihalcea, 2007) and SSEC (Schuff et al., 2017). Nearly all of these datasets use an annotation scheme based on Ekman (Ekman, 1971; Ekman, 1992) with many adding a few labels often following Plutchik's theory of emotions (Plutchik, 1980). A typical emotion dataset consists of 6-8 categories. The exception Bostan and Klinger (2018) mention is CrowdFlower[2] with 14 categories, and those not mentioned in Bostan et al. are e.g. the SemEval 2018 task 1 subtask c dataset (Mohammad et al., 2018) with 11 categories, EmoNet with 24 (Abdul-Mageed and Ungar, 2017), and the GoEmotions dataset (Demszky et al., 2020) with 27 categories.

A majority of recent papers on multilabel emotion classification focus on the SemEval 2018 dataset which is based on tweets. Similarly, many of the non-multilabel classification papers use Twitter data. Twitter is a good base for emotion classification as tweets are limited in length and generally stand-alone, i.e. the reader or annotator does not need to guess the context in the majority of cases. Furthermore, hashtags and emojis are common, which further makes the emotion recognition easier for both human annotators and emotion detection and sentiment analysis models. Reddit data, as used by Demszky et al. (2020), and movie subtitles used by this paper, are slightly more problematic as they are not "self-contained". Reddit comments are typically longer than one line and therefore provide some context for annotators to go by, but often lacks the hashtags and emojis of twitter and can be quite context-dependent as Reddit comments are by definition reactions to a post or another comment. Movie subtitles annotated out of sequence have virtually no context to aid the annotator and are supposed to be accompanied by visual cues as well. However, annotating with context can reduce the accuracy of one's model by doubly weighting surrounding units of granularity (roughly 'sentences' in our case) (Boland et al., 2013). On the other hand, contextual annotations are less frustrating for the annotator and therefore, would likely provide more annotations in the same amount of time (Öhman, 2020).

In table 1 we have gathered some of the most significant emotion datasets in relation to this study. The table lists the paper in which the dataset was released (study), what the source data that was used was (source), what model was used to obtain the best evaluation scores (model), the number of categories used for annotation (cat), whether the system was multilabel or not (multi), and the macro f1 scores and accuracy score as reported by the paper (macro f1 and accuracy respectively). Some papers only reported a micro f1 and no macro f1 score. These scores have been marked with a μ.

study	source	model	cat	multi	macro f1	accuracy
Strapparava and Mihalcea (2007)	news headlines	"CLaC system"	6+val	yes	N/A	55.1%
Tokuhisa et al. (2008)	JP web corpus	k-NN	10	no	N/A	N/A
Schuff et al. (2017)	Twitter	BiLSTM	8	yes	N/A	N/A
Abdul-Mageed and Ungar (2017)	Twitter	GRNNs	6	no	N/A	95.68%
Abdul-Mageed and Ungar (2017)	Twitter	GRNNs	24	no	87.47%	N/A
Samy et al. (2018)	Twitter	C-GRU	11+neu	yes	64.8%	53.2%
Yu et al. (2018)	Twitter	DATN	11	yes	44.4%	45.7%
Jabreel and Moreno (2019)	Twitter	BiRNN/GRU	11	yes	56.4%	59%
Huang et al. (2019)	Twitter	BiLSTM/ELMO	11	yes	70.9%μ	59.2%
Liu et al. (2019)	books etc	BERT	8+neu	no	60.4%μ	N/A
Demszky et al. (2020)	Reddit	BERT	27	yes	46%	N/A
Demszky et al. (2020)	Reddit	BERT	6	no	64%	N/A
XED (English)	movie subtitles	BERT	8+neu	yes	53.6%	54.4%

Table 1: Overview of emotion datasets.

[2]CrowdFlower was created in 2016 but has since been acquired by different companies at least twice and is now hard to find. It is currently owned by Appen.

The datasets in table 1 differ from each so much in content, structure, and manner of annotation that direct comparisons are hard to make. Typically, the fewer the number of categories, the easier the classification task and the higher the evaluation scores. It stands to reason that the easier it is to detect emotions in the source data, the easier it is for annotators to identify and agree upon annotation labels and therefore it becomes easier for the system or model to correctly classify the test data as well. The outlier in these datasets is EmoNet (Abdul-Mageed and Ungar, 2017) which achieved astonishing accuracies by using 665 different hashtags to automatically categorize 1.6 million tweets into 24 categories (Plutchik's 8 at 3 different intensities), unfortunately neither the dataset or their model has been made available for closer inspection.

The downside of datasets trained on Twitter is that they are likely not that good at classifying anything other than tweets. It is plausible that datasets trained on less specific data such as XED and those created by Tokuhisa et al. (2008) and Demszky et al. (2020) are better at crossing domains at the cost of evaluation metrics.

2.2 Annotation Projection

Research shows that affect categories are quite universal (Cowen et al., 2019; Scherer and Wallbott, 1994). Therefore, theoretically they should also to a large degree retain emotion categories when translated. Annotation projection has been shown to offer reliable results in different NLP and NLU tasks (Kajava et al., 2020; Yarowsky et al., 2001; Agić et al., 2016; Rasooli and Tetreault, 2015). Projection is sometimes the only feasible way to produce resources for under-resourced languages. By taking datasets created for high-resource languages and projecting these results on the corresponding items in the under-resourced language using parallel corpora, we can create datasets in as many languages as exist in the parallel corpus. A parallel corpus for multiple languages enables the simultaneous creation of resources for multiple languages at a low cost.

Previous annotation tasks have shown that even with binary or ternary classification schemes, human annotators agree only about 70-80% of the time and the more categories there are, the harder it becomes for annotators to agree (Boland et al., 2013; Mozetič et al., 2016). For example, when creating the DENS dataset (Liu et al., 2019), only 21% of their annotations had consensus between all annotators with 73.5% having to resort to majority agreement, and a further 5.5% could not be agreed upon and were left to expert annotators to be resolved.

Some emotions are also harder to detect, even for humans. Demszky et al. (2020) show that the emotions of *admiration, approval, annoyance, gratitude* had the highest interrater correlations at around 0.6, and *grief, relief, pride, nervousness, embarrassment* had the lowest interrater correlations between 0-0.2, with a vast majority of emotions falling in the range of 0.3-0.5 for interrater correlation. Emotions are also expressed differently in text with *anger* and *disgust* expressed explicitly, and *surprise* in context (Alm et al., 2005).

Some emotions are also more closely correlated. In Plutchik's wheel (Plutchik, 1980) related emotions are placed on the same dyad so that for example for *anger* as a core emotion, there is also *rage* that is more intense, but highly correlated with anger, and *annoyance* which is less intense, but equally correlated. In this way it is also possible to map more distinct categories of emotions onto larger wholes; in this case *rage* and *annoyance* could be mapped to *anger*, or even more coarsely to *negative*. This approach has been employed by for example Abdul-Mageed and Ungar (2017).

3 The Data

In table 2 we present an overview of the English part of the XED dataset. With 24,164 emotion annotations[3] excluding *neutral* on 17,520 unique sentences[4] the XED dataset is one of the largest emotion datasets we are aware of.

[3]Note that the total number of annotations excluding *neutral* (24,164) and the combined number of annotations (22,424) differ because once the dataset was saved as a Python dictionary, identical lines were merged as one (i.e. some common movie lines like "All right then!" and "I love you" appeared multiple times from different sources).

[4]A sentence could have been annotated as containing 3 different emotions by one or more annotators. This would count as 3 annotations on one unique data point.

We used Plutchik's core emotions as our annotation scheme resulting in 8 distinct emotion categories plus neutral. The *Sentimentator* platform (Öhman and Kajava, 2018; Öhman et al., 2018) allows for the annotation of intensities resulting in what is essentially 30 emotions and sentiments, however, as the intensity score is not available for all annotations, the intensity scores were discarded. The granularity of our annotations roughly correspond to sentence-level annotations, although as our source data is movie subtitles, our shortest subtitle is *!* and the longest subtitle consists of three separate sentences.

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 2: Overview of the XED English dataset.

A majority of the subtitles for English were assigned one emotion label (78%), 17% were assigned two, and roughly 5% had three or more categories (see also Table 3).

3.1 Movie Subtitles as Multilingual Multi-Domain Proxy

We use the OPUS (Lison and Tiedemann, 2016) parallel movie subtitle corpus of subtitles collected from opensubtitles.org as a multi-domain proxy. As the movies we use for source data cover several different genres and, although scripted, represents real human language used in a multitude of situations similar to many social media platforms.

Because OPUS open subtitles is a parallel corpus we are able to evaluate our annotated datasets across languages and at identical levels of granularity. Although the subtitles might be translated using different translation philosophies (favoring e.g. meaning, mood, or idiomatic language as the prime objective) (Carl et al., 2011), we expect the translations to have aimed at capturing the sentiments and emotions originally expressed in the film based on previous studies (e.g. Cowen et al. (2019), Scherer and Wallbott (1994), Creutz (2018), Scherrer (2020) and Kajava et al. (2020)).

3.2 Data Annotation

The vast majority of the dataset was annotated by university students learning about sentiment analysis with some annotations provided by expert annotators for reliability measurements (Öhman et al., 2018). The students' annotation process was monitored and evaluated. They received only minimal instructions. These instructions included that they were to focus on the quality of annotations rather than quantity, and to annotate from the point of view of the speaker. We also asked for feedback on the annotation process to improve the user-friendliness of the platform for future use. In tables 2 and 5 the number of active annotators have been included. All in all over 100 students annotated at least some sentences with around 60 active annotators, meaning students who annotated more than 300 sentences (Öhman, 2020).

It should be noted that the annotators were instructed to annotate the subtitles without context, a task made harder by the fact that we chose subtitles that were available for all languages, which likely meant that some of the most famous movies were included thus creating recognizable context for the annotators.

The data for annotation was chosen randomly from the OPUS subtitle corpus (Lison and Tiedemann, 2016) from subtitles that were available for the maximum number of languages. We chose 30,000 individual lines to be annotated by 3 annotators. For the final dataset, some of these annotations were not annotated by all 3 annotators, as it was possible to skip difficult-to-annotate instances, but the subtitle was included if at least 2 annotators agreed on the emotion score. In some cases if the expert annotators agreed that the annotation was feasible during the pre-processing phase, subtitles annotated by a single annotator and checked by expert annotators, were also included.

3.3 Pre-processing

After the annotations were extracted from the database, the data needed to be cleaned up. The different evaluations required different pre-processing steps. Most commonly, this included the removal of superfluous characters containing no information. We tried to keep as much of the original information as possible, including keeping offensive, racist, and sexist language as is. If such information is removed, the usefulness of the data is at risk of being reduced, particularly when used for e.g. offensive language detection (Pàmies et al., 2020).

For the English data we used Stanford NER (named entity recognition) (Finkel et al., 2005) to replace names and locations with the tags: [PERSON] and [LOCATION] respectively. We kept organization names as is because we felt that the emotions and sentiments towards some large well-known organizations differ too much (cf. IRS, FBI, WHO, EU, and MIT). For the Finnish data, we replaced names and locations using the Turku NER corpus (Luoma et al., 2020).

Some minor text cleanup was also conducted, removing hyphens and quotations marks, and correcting erroneous renderings of characters (usually encoding issues) where possible.

3.4 English Dataset Description

The final dataset contained 17,520 unique emotion-annotated subtitles as shown in table 3. In addition there are some 6.5k subtitles annotated as *neutral*. The label distribution can be seen in table 3.

anger	anticipation	disgust	fear	joy	sadness	surprise	trust	Total annotations
4,182	3,660	2,442	2,585	3,139	2,635	2,635	2,886	24,164
17.31%	15.15%	10.11%	10.70%	12.99%	10.90%	10.90%	11.94%	XED percentage
15.09%	10.15%	12.80%	17.86%	8.34%	14.41%	6.46%	14.89%	EmoLex percentage

Table 3: Emotion label distribution in the XED English dataset.

The emotion labels are surprisingly balanced with the exception of *anger* and *anticipation*, which are more common than the other labels. In comparison with one of the most well-known emotion datasets using the same annotation scheme, the NRC emotion lexicon (EmoLex) (Mohammad and Turney, 2013), the distribution differs somewhat. Although *anger* is a large category in both datasets, *fear* is average in our dataset, but the largest category in EmoLex. It is hard to speculate why this is, but one possible reason is the different source data.

The number of unique label combinations is 147, including single-label. The most common label combinations beyond single-label are *anger* with *disgust* (2.4%) and *joy* with *trust* (2.1%) followed by different combinations of the positive emotions of *anticipation*, *joy*, and *trust*. These findings are in line with previous findings discussing overlapping categories (Banea et al., 2011; Demszky et al., 2020). However, these are followed by *anger* combined with *anticipation* and *sadness* with *surprise*. The first combination is possibly a reflection of the genre, as a common theme for *anger* with *anticipation* is threats. The combination of *surprise* with negative emotions (*anger, disgust, fear, sadness*) is much more common than a combination with positive emotions.

Note that the difference between total annotations excluding *neutral* (24,164) and the combined number of annotations (22,424) differ because once the dataset was saved as a Python dictionary, identical lines were merged as one (i.e. some common movie lines like ”All right then!” and ”I love you” appeared multiple times from different sources). Additionally, lines annotated as both *neutral* and an emotion were removed from the *neutral* set.

3.4.1 Crosslingual Data & Annotation projection

From our source data we can extract parallel sentences for 43 languages. For 12 of these languages we have over 10,000 sentences available for projection as per table 4. We removed some of these languages for having fewer than 950 lines, resulting in a total of 32 languages[5] including the annotated English and

[5]Arabic (AR), Bosnian (BS), Brazilian Portuguese (PT_BR), Bulgarian (BG), Croatian (HR), Czech (CS), Danish (DA), Dutch (NL), English (EN), Estonian (ET), Finnish (FI), French (FR), German (DE), Greek (EL), Hebrew (HE), Hungarian

Finnish data. We have made all 32 datasets available on GitHub plus the raw data for all 43 languages including the 11 datasets that had fewer than 950 lines.

IT	FI	FR	CS	PT	PL	SR	TR	EL	RO	ES	PT_BR
10,582	11,128	11,503	11,885	12,559	12,836	14,831	15,712	15,713	16,217	16,608	22,194

Table 4: Languages (ISO code) with over 10k parallel sentences with our annotated English data.

To test how well our data is suited for emotion projection, we projected the English annotations onto our Finnish unannotated data using OPUS tools (Aulamo et al., 2020). We chose Finnish as our main test language as we also have some annotated data for it to use as a test set. The manually annotated Finnish data consists of nearly 20k individual annotations and almost 15k unique annotated sentences plus an additional 7,536 sentences annotated as *neutral* [6]. The criteria for the inclusion of an annotation was the same as for English. The distribution of the number of labels and the labels themselves are quite similar to that of the English data. Relatively speaking there is a little less *anticipation* in the Finnish data, but *anger* is the biggest category in both languages.

Number of annotations:	21,984 (/w neutral)
Number of unique data points:	14,449 + 7,536 neutral
Number of emotions:	8 (+neu)
Number of annotators:	40 active
Number of labels per data point:	1: 77% 2: 18% 3: 5% 4+: 1.2%

Table 5: Overview of the XED Finnish dataset.

The distribution of the number of emotions (table 5) and the distribution of emotions (table 6) are similar to their corresponding distributions in the English dataset.

anger	anticipation	disgust	fear	joy	sadness	surprise	trust	Total annotations
3,345	2,496	2,373	2,186	2,559	2,184	1,982	2,404	19,529
17.13%	12.78%	12.15%	11.19%	13.10%	11.18%	10.15%	12.31%	100%

Table 6: Emotion label distribution in the XED Finnish dataset.

We used the 11,128 Finnish sentences for which directly parallel sentences existed and projected the English annotations on them using the unique alignment IDs for both languages as guide. Some of those parallel sentences were part of our already annotated data and were discarded as training data. This served as a useful point of comparison. The average annotation correlation using Cohen's kappa is 0.44 (although accuracy by percentage is over 90%), and highest for *joy* at 0.65, showing that annotation projection differs from human annotation to a similar degree as human annotations differ from each other.

4 Evaluation

A dataset for classification tasks is useful only if the accuracy of its annotations can be confirmed. To this end we use BERT to evaluate our annotations as it has consistently outperformed other models in recent classification tasks (see e.g Zampieri et al. (2020)), and Support Vector Machines for its simplicity and effectiveness. We use a stratified split of 70:20:10 for training, dev, and test data.

(HU), Icelandic (IS), Italian (IT), Macedonian (MK), Norwegian (NO), Polish (PL), Portuguese (PT), Romanian (RO), Russian (RU), Serbian (SR), Slovak (SK), Slovenian (SL), Spanish (ES), Swedish (SV), Turkish (TR) and Vietnamese (VI)

[6]The same calculations apply here as for English. Annotations are calculated as labels which can be more than one for each line, and unique data points refer to the number of lines that had 1 or more annotations.

We use a fine-tuned English uncased BERT, with a batch size of 96. The learning rate of Adam optimizer was set to 2e-5 and the model was trained for 3 epochs. The sequence length was set to 48. We perform a 5-fold cross validation.

We also use an SVM classifier with linear kernel and regularization parameter of 1. Word unigrams, bigrams and trigrams were used as features in this case. Implementation was done using the LinearSVC class from the scikit-learn library (Pedregosa et al., 2011).

Binary refers to *positive* and *negative*, and *ternary* refers to *positive, negative*and *neutral*. For binary evaluations we categorized *anger*, *disgust*, *fear*, and *sadness* as *negative*, and *anticipation*, *joy*, and *trust* as *positive*. *Surprise* was either discarded or included as a separate category (see table 7). For this classification task BERT achieved macro f1 scores of 0.536 and accuracies of 0.544. This is comparable to other similar datasets when classes are merged (e.g. Demszky et al. (2020)).

4.1 Evaluation Metrics

We achieve macro f1 scores of 0.54 for our multilabel classification with a fine-tuned BERT model. Using named-entity recognition increases the accuracy slightly. For binary data mapped from the emotion classifications onto positive and negative (non-multilabel classification) our model achieves a macro f1 score of 0.838 and accuracy of 0.840. Our linear SVM classifier using one-vs-rest achieves an f1 score of 0.502 with per class f1 scores between 0.8073 (anger) and 0.8832 (fear & trust) (see tables 7 and 8).

data	f1	accuracy
English without NER, BERT	0.530	0.538
English with NER, BERT	0.536	0.544
English NER with neutral, BERT	0.467	0.529
English NER binary with surprise, BERT	0.679	0.765
English NER true binary, BERT	0.838	0.840
Finnish anno., FinBERT	0.507	0.513
English NER, one-vs-rest SVM (LinearSVC)[7]	0.746	

Table 7: Evaluation results of the XED English dataset.

SVM per class f1	emotion
0.8073	anger
0.8296	anticipation
0.8832	disgust
0.8763	fear
0.8819	joy
0.8762	sadness
0.8430	surprise
0.8832	trust

Table 8: SVM per class f1 scores.

The confusion matrix (see Figure 1) reveals that *disgust* is often confused with *anger*, and to some extent this is true in the other direction as well. This relation between labels can also be observed in the correlation matrix (see Figure 2), where *anger* and *disgust* appear as one of the most highly correlated pair of categories, only behind *joy* and *trust*. On the other hand, the least correlated pair is *joy* and *anger*, closely followed by *trust* and *anger*. *Disgust* is also the hardest emotion to categorize correctly. In fact, it is more often classified as *anger* than *disgust*. *Joy, anger* and *anticipation* are the categories that are categorized correctly the most.

The correlation matrix for the multilabel English evaluation shows (table 2) how closely correlated the emotions of *anger* and *disgust*, and *joy* and *trust* in particular are.

4.2 Evaluating Annotation Projection

With the same parameters as for English, we used language-specific BERT models from Huggingface transformers (Wolf et al., 2019) for the Arabic, Chinese, Dutch, Finnish, German and Turkish datasets with 5-fold cross-validation. The annotated Finnish dataset achieves an f1 score of 0.51. The projected annotations achieve slightly worse f1 scores than the annotated dataset at 0.45 for Finnish (see table 9). The other datasets achieve similar f1 scores, with the Germanic languages of German and Dutch achieving almost as high scores as the original English dataset. This is likely a reflection of typological, cultural, and linguistic similarities between the languages making the translation to begin with more similar to the original and therefore minimizing information loss.

We also evaluated all the projected datasets using a linear SVC classifier. In most cases the linear SVC classifier performs better than language-specific BERT. We speculate this is related to the size of

[7] The SVM evaluation was performed on the student annotations only in order to be fully comparable to the projections. The BERT evaluations also contain additional data from the Sentimentator and cynarr GitHub repos. These are linked to from the main XED repo.

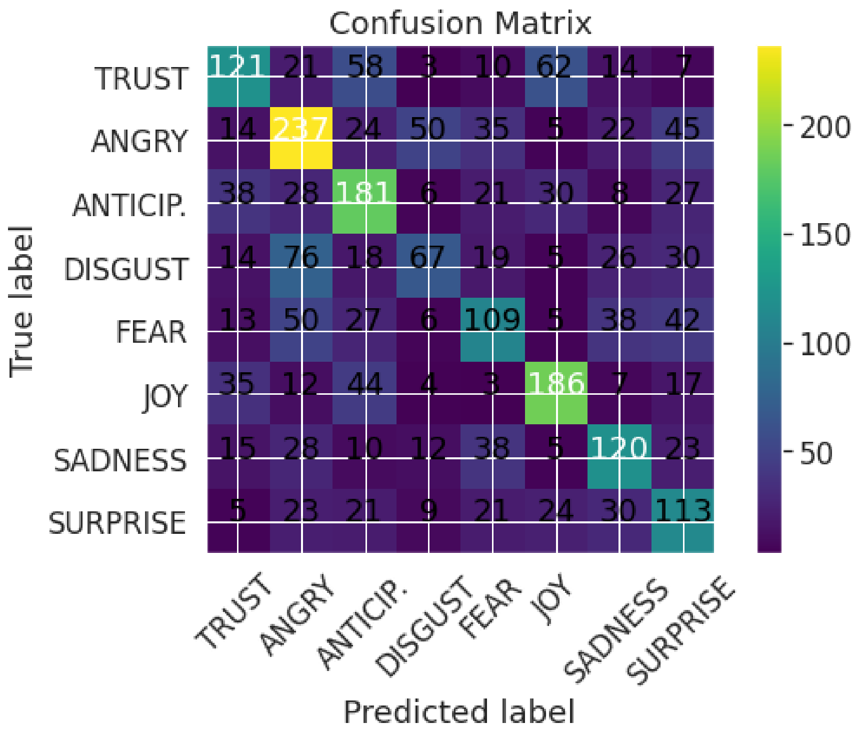

Figure 1: Confusion matrix for the XED English dataset.

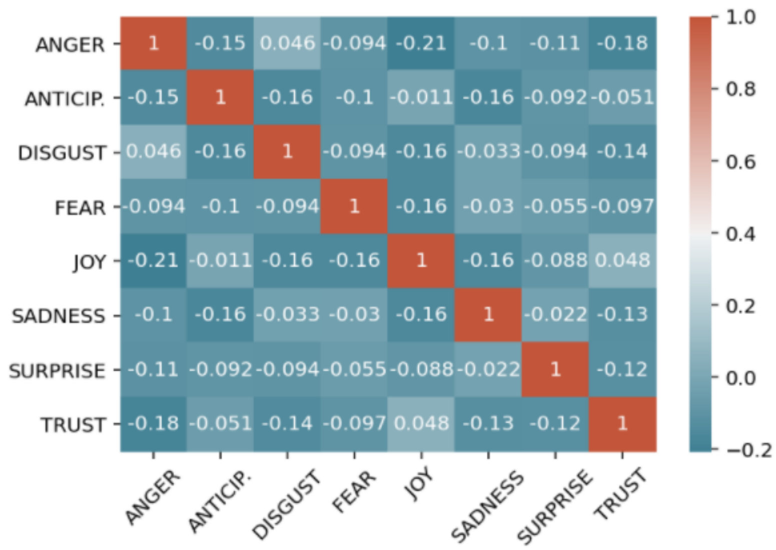

Figure 2: Correlation matrix for the XED English dataset.

data	f1	accuracy
Finnish projected	0.4461	0.4542
Turkish projected	0.4685	0.5257
Arabic projected	0.4627	0.5339
German projected	0.5084	0.5737
Dutch projected	0.5155	0.5822
Chinese projected	0.4729	0.5247

Table 9: Annotation projection evaluation results using language-specific BERT models.

the datasets.

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 10: SVM's macro f1 scores for all projected languages.

5 Discussion

The results from the dataset evaluations show that the XED is on par with other similar datasets, but they also stress that reliable emotion detection is still a very challenging task. It is not necessarily an issue with natural language processing and understanding as these types of tasks are challenging for human annotators alike. If human annotators cannot agree on labels, it is not reasonable to think computers can do any better regardless of annotation scheme or model used since these models are restricted by human performance. The best accuracies are those that are in line with annotator agreement.

XED is a novel state-of-the-art dataset that provides a new challenge in fine-grained emotion detection with previously unavailable language coverage. What makes the XED dataset particularly valuable is the large number of annotations at high granularity, as most other similar datasets are annotated at a much coarser granularity. The use of movie subtitles as source data means that it is possible to use the XED dataset across multiple domains (e.g. social media) as the source data is representative of other domains and not as restricted to the domain of the source data (movies) as many other datasets. Perhaps the greatest contribution of all is that, for the first time, many under-resourced languages have emotion datasets that can be used in other possible downstream applications as well.

References

Muhammad Abdul-Mageed and Lyle Ungar. 2017. Emonet: Fine-grained emotion detection with gated recurrent neural networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 718–728.

Željko Agić, Anders Johannsen, Barbara Plank, Héctor Martínez Alonso, Natalie Schluter, and Anders Søgaard. 2016. Multilingual projection for parsing truly low-resource languages. *Transactions of the Association for Computational Linguistics*, 4:301–312.

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005. Emotions from text: Machine learning for text-based emotion prediction. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 579–586. Association for Computational Linguistics.

Mikko Aulamo, Umut Sulubacak, Sami Virpioja, and Jörg Tiedemann. 2020. OpusTools and Parallel Corpus Diagnostics. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3782–3789.

Carmen Banea, Rada Mihalcea, and Janyce Wiebe. 2011. Multilingual Sentiment and Subjectivity Analysis. *Multilingual natural language processing*, 6:1–19. Accessed October 16, 2018.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 440–447.

Katarina Boland, Andias Wira-Alam, and Reinhard Messerschmidt. 2013. *Creating an Annotated Corpus for Sentiment Analysis of German Product Reviews*, volume 2013/05 of *GESIS-Technical Reports*. GESIS - Leibniz-Institut für Sozialwissenschaften, Mannheim.

Laura Ana Maria Bostan and Roman Klinger. 2018. An analysis of annotated corpora for emotion classification in text. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2104–2119.

Michael Carl, Barbara Dragsted, and Arnt Lykke Jakobsen. 2011. A taxonomy of human translation styles. *Translation journal*, 16(2):155–168.

Alan S. Cowen, Petri Laukka, Hillary Anger Elfenbein, Runjing Liu, and Dacher Keltner. 2019. The primacy of categories in the recognition of 12 emotions in speech prosody across two cultures. *Nature human behaviour*, 3:369 – 382.

Mathias Creutz. 2018. Open Subtitles Paraphrase Corpus for Six Languages. *arXiv preprint arXiv:1809.06142*. arXiv:1809.06142 [cs.CL]. Accessed November 9 2018.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A Dataset of Fine-Grained Emotions.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Paul Ekman. 1971. Universals and cultural differences in facial expressions of emotion. In *Nebraska symposium on motivation*. University of Nebraska Press.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, page 363–370, USA. Association for Computational Linguistics.

Chenyang Huang, Amine Trabelsi, Xuebin Qin, Nawshad Farruque, and Osmar R. Zaïane. 2019. Seq2Emo for Multi-label Emotion Classification Based on Latent Variable Chains Transformation. *arXiv preprint arXiv:1911.02147*.

Mohammed Jabreel and Antonio Moreno. 2019. A deep learning-based approach for multi-label emotion classification in tweets. *Applied Sciences*, 9:1123, 03.

Kaisla SA Kajava, Emily Sofi Öhman, Piao Hui, and Jörg Tiedemann. 2020. Emotion Preservation in Translation: Evaluating Datasets for Annotation Projection. In *Digital Humanities in the Nordic Countries 2020*. CEUR Workshop Proceedings.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *LREC*. European Language Resources Association (ELRA).

Chen Liu, Muhammad Osama, and Anderson de Andrade. 2019. DENS: A Dataset for Multi-class Emotion Analysis. *ArXiv*, abs/1910.11769.

Jouni Luoma, Miika Oinonen, Maria Pyykönen, Veronika Laippala, and Sampo Pyysalo. 2020. A Broad-coverage Corpus for Finnish Named Entity Recognition. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4615–4624.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June. Association for Computational Linguistics.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 task 1: Affect in tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana, June. Association for Computational Linguistics.

Igor Mozetič, Miha Grčar, and Jasmina Smailović. 2016. Multilingual twitter sentiment classification: The role of human annotators. *PloS one*, 11(5):e0155036.

Mika V. Mäntylä, Daniel Graziotin, and Miikka Kuutila. 2018. The evolution of sentiment analysis—a review of research topics, venues, and top cited papers. *Computer Science Review*, 27:16 – 32.

Emily Öhman and Kaisla Kajava. 2018. Sentimentator: Gamifying Fine-grained Sentiment Annotation. In *Digital Humanities in the Nordic Countries 2018*. CEUR Workshop Proceedings. Accessed October 16 2018.

Emily Öhman, Jörg Tiedemann, Timo Honkela, and Kaisla Kajava. 2018. Creating a Dataset for Multilingual Fine-grained Emotion-detection Using Gamification-based Annotation. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Association for Computational Linguistics.

Emily Öhman. 2020. Challenges in Annotation: Annotator Experiences from a Crowdsourced Emotion Annotation Task. In *Digital Humanities in the Nordic Countries 2020*. CEUR Workshop Proceedings.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. *Theories of emotion*, 1:3–31.

Marc Pàmies, Emily Öhman, Kaisla Kajava, and Jörg Tiedemann. 2020. LT@Helsinki at SemEval-2020 Task 12: Multilingual or language-specific BERT? In *Proceedings of the 14th International Workshop on Semantic Evaluation*.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. version 2.

Ahmed E. Samy, Samhaa R. El-Beltagy, and Ehab Hassanien. 2018. A context integrated model for multi-label emotion detection. *Procedia Computer Science*, 142:61 – 71. Arabic Computational Linguistics.

Klaus R. Scherer and Harald Günter Wallbott. 1994. Evidence for universality and cultural variation of differential emotion response patterning. *Journal of personality and social psychology*, 66 2:310–28.

Yves Scherrer. 2020. TaPaCo: A Corpus of Sentential Paraphrases for 73 Languages. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 6868–6873.

Hendrik Schuff, Jeremy Barnes, Julian Mohme, Sebastian Padó, and Roman Klinger. 2017. Annotation, modelling and analysis of fine-grained emotions on a stance and sentiment detection corpus. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 13–23.

Carlo Strapparava and Rada Mihalcea. 2007. Semeval-2007 task 14: Affective text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 70–74.

Ryoko Tokuhisa, Kentaro Inui, and Yuji Matsumoto. 2008. Emotion classification using massive examples extracted from the web. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 881–888. Association for Computational Linguistics.

Peter D. Turney. 2002. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 417–424, Stroudsburg, PA, USA. Association for Computational Linguistics.

Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: Bert for finnish. *arXiv preprint arXiv:1912.07076*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, pages arXiv–1910.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the First International Conference on Human Language Technology Research*.

Jianfei Yu, Luís Marujo, Jing Jiang, Pradeep Karuturi, and William Brendel. 2018. Improving multi-label emotion classification via sentiment classification with dual attention transfer network. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1097–1102, Brussels, Belgium, October-November. Association for Computational Linguistics.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020). *arXiv preprint arXiv:2006.07235*.