# Learning and Evaluating Emotion Lexicons for 91 Languages

**Sven Buechel, Susanna Rücker, and Udo Hahn**
{sven.buechel|susanna.ruecker|udo.hahn}@uni-jena.de
Jena University Language and Information Engineering (JULIE) Lab
Friedrich-Schiller-Universität Jena, Jena, Germany
https://julielab.de

## Abstract

Emotion lexicons describe the affective meaning of words and thus constitute a centerpiece for advanced sentiment and emotion analysis. Yet, manually curated lexicons are only available for a handful of languages, leaving most languages of the world without such a precious resource for downstream applications. Even worse, their coverage is often limited both in terms of the lexical units they contain and the emotional variables they feature. In order to break this bottleneck, we here introduce a methodology for creating almost arbitrarily large emotion lexicons for any target language. Our approach requires nothing but a source language emotion lexicon, a bilingual word translation model, and a target language embedding model. Fulfilling these requirements for 91 languages, we are able to generate representationally rich high-coverage lexicons comprising eight emotional variables with more than 100k lexical entries each. We evaluated the automatically generated lexicons against human judgment from 26 datasets, spanning 12 typologically diverse languages, and found that our approach produces results in line with state-of-the-art *monolingual* approaches to lexicon creation and even *surpasses human reliability* for some languages and variables. Code and data are available at github.com/JULIELab/MEmoLon archived under DOI 10.5281/zenodo.3779901.

## 1 Introduction

An emotion lexicon is a lexical repository which encodes the affective meaning of individual words (lexical entries). Most simply, affective meaning can be encoded in terms of *polarity*, i.e., the distinction whether an item is considered as positive, negative, or neutral. This is the case for many well-known resources such as WORDNET-AFFECT (Strapparava and Valitutti, 2004), SENTIWORDNET (Baccianella et al., 2010), or VADER (Hutto and Gilbert, 2014). Yet, an increasing number of researchers focus on more expressive encodings for affective states inspired by distinct lines of work in psychology (Yu et al., 2016; Buechel and Hahn, 2017; Sedoc et al., 2017; Abdul-Mageed and Ungar, 2017; Bostan and Klinger, 2018; Mohammad, 2018; Troiano et al., 2019).

Psychologists, on the one hand, value such lexicons as a controlled set of stimuli for designing experiments, e.g., to investigate patterns of lexical access or the structure of memory (Hofmann et al., 2009; Monnier and Syssau, 2008). NLP researchers, on the other hand, use them to augment the emotional loading of word embeddings (Yu et al., 2017; Khosla et al., 2018), as additional input to sentence-level emotion models so that the performance of even the most sophisticated neural network gets boosted (Mohammad and Bravo-Marquez, 2017; Mohammad et al., 2018; De Bruyne et al., 2019), or rely on them in a keyword-spotting approach when no training data is available, e.g., for studies dealing with historical language stages (Buechel et al., 2016).

As with any kind of manually curated resource, the availability of emotion lexicons is heavily restricted to only a few languages whose exact number varies depending on the variables under scrutiny. For example, we are aware of lexicons for 15 languages that encode the emotional variables of Valence, Arousal, and Dominance (see Section 2). This number leaves the majority of the world's (less-resourced) languages without such a dataset. In case such a lexicon exists for a particular language, it is often severely limited in size, sometimes only comprising some hundreds of entries (Davidson and Innes-Ker, 2014). Yet, even the largest lexicons typically cover only some ten thousands of words, still leaving out major portions of the emotion-carrying vocabulary. This is especially true for languages with complex morphology or



*Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1202–1217

productive compounding, such as Finnish, Turkish, Czech, or German. Finally, the diversity of emotion representation schemes adds another layer of complexity. While psychologists and NLP researchers alike find that different sets of emotional variables are complementary to each other (Stevenson et al., 2007; Pinheiro et al., 2017; Barnes et al., 2019; De Bruyne et al., 2019), *manually* creating emotion lexicons for every language and every emotion representation scheme is virtually impossible.

We here propose an approach based on cross-lingual distant supervision to generate almost arbitrarily large emotion lexicons for any target language and emotional variable, provided the following requirements are met: a source language emotion lexicon covering the desired variables, a bilingual word translation model, and a target language embedding model. By fulfilling these preconditions, we can *automatically* generate emotion lexicons for 91 languages covering ratings for eight emotional variables and hundreds of thousands of lexical entries each. Our experiments reveal that our method is on a par with state-of-the-art monolingual approaches and compares favorably with (sometimes even outperforms) human reliability.

## 2 Related Work

**Representing Emotion.** Whereas research in NLP has focused for a very long time almost exclusively on *polarity*, more recently, there has been a growing interest in more informative representation structures for affective states by including different groups of emotional variables (Bostan and Klinger, 2018). Borrowing from distinct schools of thought in psychology, these variables can typically be subdivided into *dimensional* vs. *discrete* approaches to emotion representation (Calvo and Mac Kim, 2013). The *dimensional* approach assumes that emotional states can be *composed* out of several foundational factors, most noticeably *Valence* (corresponding to polarity), *Arousal* (measuring calmness vs. excitement), and *Dominance* (the perceived degree of control in a social situation); VAD, for short (Bradley and Lang, 1994). Conversely, the *discrete* approach assumes that emotional states can be *reduced* to a small, evolutionary motivated set of basic emotions (Ekman, 1992). Although the exact division of the set has been subject of hot debates, recently constructed datasets (see Section 4) most often cover the categories of *Joy*, *Anger*, *Sadness*, *Fear*, and *Disgust*; BE5, for short. Plutchik's Wheel of Emotion takes a middle ground between those two positions by postulating emotional categories which are yet grouped into opposite pairs along different levels of intensity (Plutchik, 1980).

Another dividing line between representational approaches is whether target variables are encoded in terms of (strict) class-membership or scores for numerical strength. In the first case, emotion analysis translates into a (multi-class) classification problem, whereas the latter turns it into a regression problem (Buechel and Hahn, 2016). While our proposed methodology is agnostic towards the chosen emotion format, we will focus on the VAD and BE5 formats here, using numerical ratings (see the examples in Table 1) due to the widespread availability of such data. Accordingly, this paper treats word emotion prediction as a regression problem.

| | Val | Aro | Dom | Joy | Ang | Sad | Fea | Dis |
|---|---|---|---|---|---|---|---|---|
| *sunshine* | 8.1 | 5.3 | 5.4 | 4.2 | 1.2 | 1.3 | 1.3 | 1.2 |
| *terrorism* | 1.6 | 7.4 | 2.7 | 1.2 | 2.9 | 3.3 | 3.9 | 2.5 |
| *nuclear* | 4.3 | 7.3 | 4.1 | 1.4 | 2.2 | 1.9 | 3.2 | 1.6 |
| *ownership* | 5.9 | 4.4 | 7.5 | 2.1 | 1.4 | 1.2 | 1.4 | 1.3 |

Table 1: Sample entries from our English source lexicon described via eight emotional variables: **Val**ence, **Aro**usal, **Dom**inance [VAD], and **Joy**, **Ang**er, **Sad**ness, **Fea**r, and **Dis**gust [BE5]. VAD uses 1-to-9 scales ("5" encodes the neutral value) and BE5 1-to-5 scales ("1" encodes the neutral value).

**Building Emotion Lexicons.** Usually, the ground truth for affective word ratings (i.e., the assignment of emotional values to a lexical item) is acquired in a questionnaire study design where subjects (annotators) receive lists of words which they rate according to different emotion variables or categories. Aggregating individual ratings of multiple annotators then results in the final emotion lexicon (Bradley and Lang, 1999). Recently, this workflow has often been enhanced by crowdsourcing (Mohammad and Turney, 2013) and best-worst scaling (Kiritchenko and Mohammad, 2016).

As a viable alternative to manual acquisition, such lexicons can also be created by automatic means (Bestgen, 2008; Köper and Schulte im Walde, 2016; Shaikh et al., 2016), i.e., by learning to predict emotion labels for unseen words. Researchers have worked on this prediction problem for quite a long time. Early work tended to focus on word statistics, often in combination with linguistic rules (Hatzivassiloglou and McKeown,

1997; Turney and Littman, 2003). More recent approaches focus heavily on word embeddings, either using semi-supervised graph-based approaches (Wang et al., 2016; Hamilton et al., 2016; Sedoc et al., 2017) or fully supervised methods (Rosenthal et al., 2015; Li et al., 2017; Rothe et al., 2016; Du and Zhang, 2016). Most important for this work, Buechel and Hahn (2018b) report on near-human performance using a combination of FASTTEXT vectors and a multi-task feed-forward network (see Section 4). While this line of work can add new words, it does not extend lexicons to other emotional variables or languages.

A relatively new way of generating novel labels is *emotion representation mapping* (ERM), an annotation projection that translates ratings from one emotion format into another, e.g., mapping VAD labels into BE5, or vice versa (Hoffmann et al., 2012; Buechel and Hahn, 2016, 2018a; Alarcão and Fonseca, 2017; Landowska, 2018; Zhou et al., 2020; Park et al., 2019). While our work uses ERM to add additional emotion variables to the source lexicon, ERM alone can neither increase the coverage of a lexicon, nor adapt it to another language.

**Translating Emotions.** The approach we propose is strongly tied to the observation by Leveau et al. (2012) and Warriner et al. (2013) who found—comparing a large number of existing emotion lexicons of different languages—that translational equivalents of words show strong stability and adherence to their emotional value. Yet, their work is purely descriptive. They do not exploit their observation to create new ratings, and only consider manual rather than automatic translation.

Making indirect use of this observation, Mohammad and Turney (2013) offer machine-translated versions of their *NRC Emotion Lexicon*. Also, many approaches in cross-lingual sentiment analysis (on the sentence-level) rely on translating polarity lexicons (Abdalla and Hirst, 2017; Barnes et al., 2018). Perhaps most similar to our work, Chen and Skiena (2014) create (polarity-only) lexicons for 136 languages by building a multilingual word graph and propagating sentiment labels through that graph. Yet, their method is restricted to high frequency words—their lexicons cover between 12 and 4,653 entries, whereas our approach exceeds this limit by more than two orders of magnitude.

Our methodology also resembles previous work which models word emotion for historical language stages (Cook and Stevenson, 2010; Hamilton et al., 2016; Hellrich et al., 2018; Li et al., 2019). Work in this direction typically comes up with a set of seed words with assumingly *temporally stable* affective meaning (our work assumes stability against translation) and then uses distributional methods to derive emotion ratings in the target language stage. However, gold data for the target language (stage) is usually inaccessible, often preventing evaluation against human judgment. In contrast, we here propose several alternative evaluation set-ups as an integral part of our methodology.

## 3 A Novel Approach to Lexicon Creation

Our methodology integrates (1) cross-lingual generation and expansion of emotion lexicons and (2) their evaluation against gold and silver standard data. Consequently, a key aspect of our workflow design is how data is split into train, dev, and test sets at different points of the generation process. Figure 1 gives an overview of our framework including a toy example for illustration.

**Lexicon Generation.** We start with a lexicon (`Source`) of arbitrary size, emotion format[1] and source language which is partitioned into train, dev, and test splits denoted by `Source-train`, `Source-dev`, and `Source-test`, respectively. Next, we leverage a bilingual word translation model between source and desired target language to build the first target-side emotion lexicon denoted as `TargetMT`. Source words are translated according to the model, whereas target-side emotion labels are simply copied from the source to the target (see Section 2). Entries are assigned to train, dev, or test set according to their source-side assignment (cf. Figure 1). The choice of our translation service (see below) ensures that each source word receives exactly one translation.

`TargetMT` is then used as the distant supervisor to train a model that predicts word emotions based on target-side word embeddings. `TargetMT-train` and `TargetMT-dev` are used to fit model parameters and optimize hyperparameters, respectively, whereas `TargetMT-test` is held out for later evaluation. Once finalized, the model is used to predict *new labels* for the words in `TargetMT`, resulting in a second target-side emotion lexicon denoted `TargetPred`. Our rationale for doing so is that a reasonably trained model should generalize well

[1] This encompasses not only VA(D) and BE5, but also any sort of (real-valued) polarity encodings.

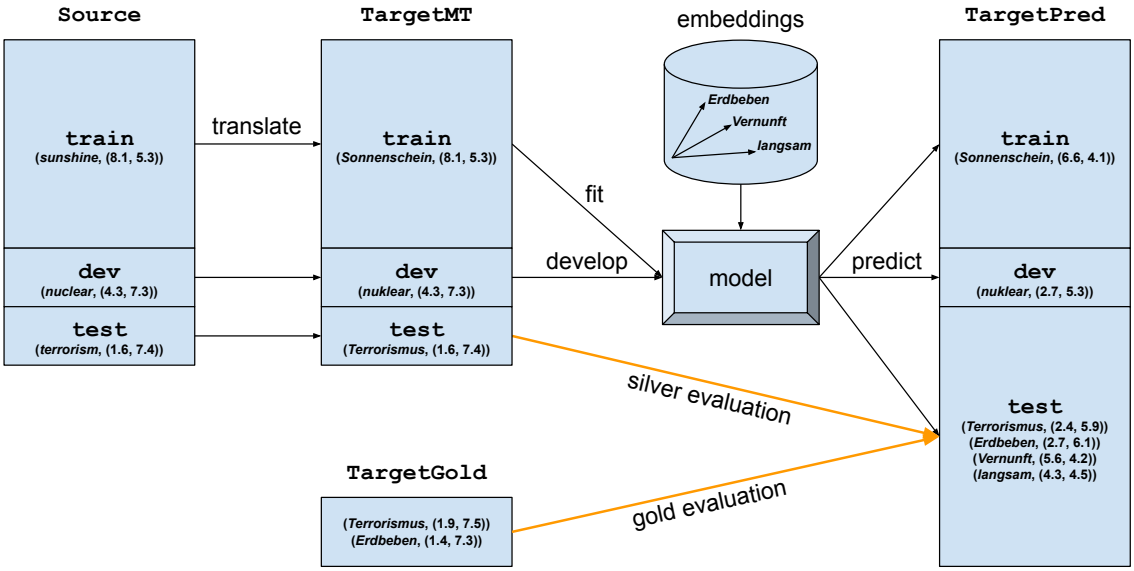


Figure 1: Schematic view on the methodology for generating and evaluating an emotion lexicon for a given target language based on source language supervision. Included is a toy example starting with an English VA lexicon (*sunshine, nuclear, terrorism* and the associated numerical scores for Valence and Arousal) and resulting in an extended German lexicon which incorporates translated entries with altered VA scores and additional entries originating from the embedding model with newly learned scores.

over the entire `TargetMT` lexicon because it has access to the target-side embedding vectors. Hence, it may mitigate some of the errors which were introduced in previous steps, either by machine translation or by assuming that source- and target-side emotion are always identical. We validate this assumption in Section 6. We also predict ratings for *all* the words in the embedding model, leading to a large number of new entries.

The splits are defined as follows: let $MT_{\text{train}}$, $MT_{\text{dev}}$, and $MT_{\text{test}}$ denote the set of words in train, dev, and test split of `TargetMT`, respectively. Likewise, let $P_{\text{train}}$, $P_{\text{dev}}$, and $P_{\text{test}}$ denote the splits of `TargetPred` and let $E$ denote the set of words in the embedding model. Then

$$\begin{aligned} P_{\text{train}} &:= MT_{\text{train}} \\ P_{\text{dev}} &:= MT_{\text{dev}} \setminus MT_{\text{train}} \\ P_{\text{test}} &:= (MT_{\text{test}} \cup E) \setminus (MT_{\text{dev}} \cup MT_{\text{train}}) \end{aligned}$$

The above definitions help clarify the way we address polysemy.[2] Ambiguity on the target-side may result in multiple source entries translating to the same target-side word.[3] This circumstance leads to "partial duplicates" in `TargetMT`, i.e., groups of entries with the same word type but different emotion values (because they were derived from distinct `Source` entries). Such overlap could do harm to the integrity of our evaluation since knowledge may "leak" from training to validation phase, i.e., by testing the model on words it has already seen during training, although with distinct emotion labels. The proposed data partitioning eliminates such distortion effects. Since partial duplicates receive the same embedding vector, the prediction model assigns the same emotion value to both, thus merging them in `TargetPred`.

**Evaluation Methodology.** The main advantage of the above generation method is that it allows us to create large-scale emotion lexicons for languages

[2] In short, our work evades this problem by dealing with lexical entries exclusively on the type- rather than the sense-level. From a lexicological perspective, this may seem like a strong assumption. From a modeling perspective, however, it appears almost obvious as it aligns well with the major components of our methodology, i.e., lexicons, embeddings, and translation. The lexicons we work with follow the design of behavioral experiments: a stimulus (word type) is given to a subject and the response (rating) is recorded. The absence of sense-level annotation simplifies the mapping between lexicon and embedding entries. While sense embeddings form an active area of research (Camacho-Collados and Pilehvar, 2018; Chi and Chen, 2018), to the best of our knowledge, type-level embeddings yield state-of-the-art performance in downstream applications.

[3] Source-side polysemy, in contrast to its target-side counterpart, is less of a problem, because we receive only a single candidate during translation. This may result in cases where the translation misaligns with the copied emotion value in `TargetMT`. Yet, the prediction step partly mitigates such inconsistencies (see Section 6).

for which gold data is lacking. But if that is the case, how can we assess the quality of the generated lexicons? Our solution is to propose two different evaluation scenarios—a *gold evaluation* which is a strict comparison against human judgment, meaning that it is limited to languages where such data (denoted `TargetGold`) is available, and a *silver evaluation* which substitutes human judgments by automatically derived ones (silver standard) which is feasible for any language in our study. The rationale is that if both, gold and silver evaluation, strongly agree with each other, we can use one as proxy for the other when no target-side gold data exists (examined in Section 6).

Note that our lexicon generation approach consists of two major steps, *translation* and *prediction*. However, these two steps are not equally important for each generated entry in `TargetPred`. Words, such as German *Sonnenschein* for which a translational equivalent already exists in the `Source` ("sunshine"; see Figure 1), mainly rely on translation, while the prediction step acts as an optional refinement procedure. In contrast, the prediction step is crucial for words, such as *Erdbeben*, whose translational equivalents ("earthquake") are missing in the `Source`. Yet, these words also depend on the translation step for producing training data.

These considerations are important for deciding which words to evaluate on. We may choose to base our evaluation on the full `TargetPred` lexicon, including words from the training set—after all, the word emotion model does not have access to *any* target-side gold data. The problem with this approach is that it merges words that mainly rely on *translation*, because their equivalents are in the `Source`, and those which largely depend on *prediction*, because they are taken from the embedding model. In this case, generalizability of evaluation results becomes questionable.

Thus, our evaluation methodology needs to fulfill the following two requirements: (1) evaluation must not be performed on translational equivalents of the `Source` entries to which the model already had access during training (e.g., *Sonnenschein* and *nuklear* in our example from Figure 1); but, on the other hand, (2) a reasonable number of instances must be available for evaluation (ideally, as many as possible to increase reliability). The intricate cross-lingual train-dev-test set assignment of our generation methodology is in place so that we meet these two requirements.

| ID | Encoding | Size | Citation |
|---|---|---|---|
| en1 | VAD | 1032 | Warriner et al. (2013) |
| en2 | VAD | 1034 | Bradley and Lang (1999) |
| en3 | BE5 | 1034 | Stevenson et al. (2007) |
| es1 | VAD | 1034 | Redondo et al. (2007) |
| es2 | VA | 14031 | Stadthagen-González et al. (2017) |
| es3 | VA | 875 | Hinojosa et al. (2016) |
| es4 | BE5 | 875 | Hinojosa et al. (2016) |
| es5 | BE5 | 10491 | Stadthagen-González et al. (2018) |
| es6 | BE5 | 2266 | Ferré et al. (2017) |
| de1 | VAD | 1003 | Schmidtke et al. (2014) |
| de2 | VA | 2902 | Võ et al. (2009) |
| de3 | VA | 1000 | Kanske and Kotz (2010) |
| de4 | BE5 | 1958 | Briesemeister et al. (2011) |
| pl1 | VAD | 4905 | Imbir (2016) |
| pl2 | VA | 2902 | Riegel et al. (2015) |
| pl3 | BE5 | 2902 | Wierzba et al. (2015) |
| zh1 | VA | 2794 | Yu et al. (2016) |
| zh2 | VA | 1100 | Yao et al. (2017) |
| it | VAD | 1121 | Montefinese et al. (2014) |
| pt | VAD | 1034 | Soares et al. (2012) |
| nl | VA | 4299 | Moors et al. (2013) |
| id | VAD | 1487 | Sianipar et al. (2016) |
| el | VAD | 1034 | Palogiannidi et al. (2016) |
| tr1 | VA | 2029 | Kapucu et al. (2018) |
| tr2 | BE5 | 2029 | Kapucu et al. (2018) |
| hr | VA | 3022 | Ćoso et al. (2019) |

Table 2: Lexicons used for gold evaluation. **ID**s consist of the respective ISO 639-1 language code plus a cardinal number to distinguish different datasets, if needed; the format of emotion **Encoding** is specified and **Size** gives the number of lexical entries per lexicon.

In particular, for our silver evaluation, we intersect `TargetMT-test` with `TargetPred-test` and compute the correlation of these two sets individually for each emotion variable. Pearson's $r$ will be used as correlation measure throughout this paper. Establishing a test set at the very start of our workflow, `Source-test`, assures that there is a relatively large overlap between the two sets and, by extension, that our requirements for the evaluation are met.

The gold evaluation is a somewhat more challenging case, because we can, in general, not guarantee that the overlap of a `TargetGold` lexicon with `TargetPred-test` will be of any particular size. For this reason, the words of the embedding model are added to `TargetPred-test` (see above), maximizing the expected overlap with `TargetGold`. In practical terms, we intersect `TargetGold` with `TargetPred-test` and compute the variable-wise correlation between these sets, in parallel to the silver evaluation. A complementary strategy for maximizing overlap, by exploiting dependencies between published lexicons, is described below.

## 4 Experimental Setup

**Gold Lexicons and Data Splits.** We use the English emotion lexicon from Warriner et al. (2013) as first part of our `Source` dataset. This popular resource comprises about 14k entries in VAD format collected via crowdsourcing. Since manually gathered BE5 ratings are available only for a subset of this lexicon (Stevenson et al., 2007), we add BE5 ratings from Buechel and Hahn (2018a) who used emotion representation mapping (see Section 2) to convert the existing VAD ratings, showing that this is about as reliable as human annotation.

As apparent from the previous section, a crucial aspect for applying our methodology is the design of the train-dev-test split of the `Source` because it directly impacts the amount of words we can test our lexicons on during gold evaluation. In line with these considerations, we choose the lexical items which are already present in ANEW (Bradley and Lang, 1999) as `Source-test` set. ANEW is the precursor to the version later distributed by Warriner et al. (2013); it is widely used and has been adapted to a wide range of languages. With this choice, it is likely that a resulting `TargetPred-test` set has a large overlap with the respective `TargetGold` lexicon. As for the `TargetGold` lexicons, we included every VA(D) and BE5 lexicon we could get hold of with more than 500 entries. This resulted in 26 datasets covering 12 quite diverse languages (see Table 2). Note that we also include English lexicons in the gold evaluation. In these cases, no translation will be carried out (`Source` is identical to `TargetMT`) so that only the expansion step is validated. Appendix A.1 gives further details on data preparation.

**Translation.** We used the GOOGLE CLOUD TRANSLATION API[4] to produce word-to-word translation tables. This is a commercial service, total translation costs amount to 160 EUR. API calls were performed in November 2019.

**Embeddings.** We use the `fastText` embedding models from Grave et al. (2018) trained for 157 languages on the respective WIKIPEDIA and the respective part of COMMONCRAWL. These resources not only greatly facilitate our work but also increase comparability across languages. The restriction to "only" 91 languages comes from intersecting the ones covered by the vectors with the languages covered by the translation service.

[4] https://cloud.google.com/translate/

**Models.** Since our proposed methodology is agnostic towards the chosen word emotion model, we will re-use models from the literature. In particular, we will rely on the multi-task learning feed-forward network (MTLFFN) worked out by Buechel and Hahn (2018b). This network constitutes the current state of the art for *monolingual* emotion lexicon creation (expanding an existing lexicon for a given language) for many of the datasets in Table 2.

The MTLFFN has two hidden layers of 256 and 128 units, respectively, and takes pre-trained embedding vectors as input. Its distinguishing feature is that hidden layer parameters are shared between the different emotion target variables, thus constituting a mild form of multi-task learning (MTL). We apply MTL to VAD and BE5 variables individually (but not between both groups), thus training two *distinct* emotion models per language, following the outcome of a development experiment. Details are given in Appendix A.2 together with the remainder of the model specifications.

Being aware of the infamous instability of neural approaches (Reimers and Gurevych, 2017), we also employ a ridge regression model, an $L_2$ regularized version of linear regression, as a more robust, yet also powerful baseline (Li et al., 2017).

## 5 Results

The size of the resulting lexicons (a complete list is provided in Table 8 in the Appendix) ranges from roughly 100k to more than 2M entries mainly depending on the vocabulary of the respective embeddings. We want to point out that not every single entry should be considered meaningful because of noise in the embedding vocabulary caused by typos and tokenization errors. However, choosing the "best" size for an emotion lexicon necessarily translates into a quality-coverage trade-off for which there is no general solution. Instead, we release the full-size lexicons and leave it to prospective users to apply any sort of filtering they deem appropriate.

**Silver Evaluation.** Figure 2 displays the results of our silver evaluation. Languages (x-axis) are sorted by their average performance over all variables (not shown in the plot; tabular data given in the Appendix). As can be seen, the evaluation results for English are markedly better than for any other language. This is not surprising since no (potentially error-prone) machine translation was performed. Apart from that, performance remains relatively stable across most of the languages and

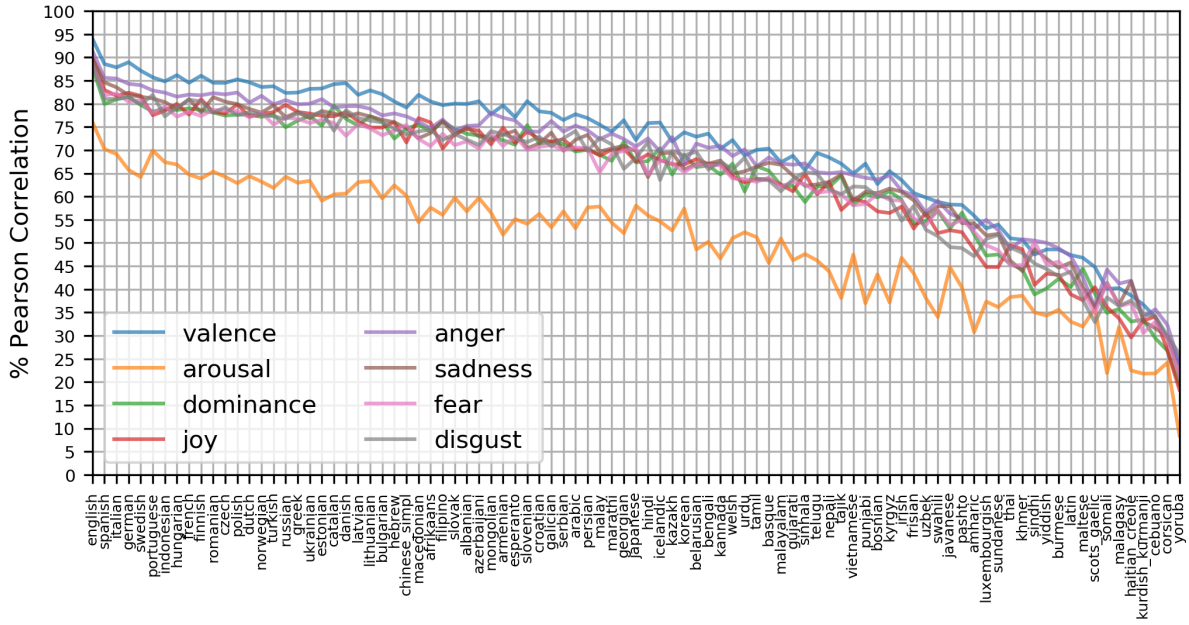


Figure 2: Silver evaluation results in Pearson's $r$. Languages (x-axis) are sorted according to mean correlation.

starts degrading more quickly only for the last third of them. In particular, for Valence—typically the easiest variable to predict—we achieve a strong performance of $r > .7$ for 56 languages. On the other hand, for Arousal—typically, the most difficult one to predict—we achieve a solid performance of $r > .5$ for 55 languages. Dominance and the discrete emotion variables show performance trajectories swinging between these two extremes. We assume that the main factors for explaining performance differences between languages are the quality of the translation and embedding models which, in turn, both depend on the amount of available text data (parallel or monolingual, respectively).

Comparing MTLFFN and ridge baseline, we find that the neural network reliably outperforms the linear model. On average over all languages and variables, the MTL models achieve 6.7%-points higher Pearson correlation. Conversely, ridge regression outperforms MTLFFN in only 15 of the total 728 cases (91 languages $\times$ 8 variables).

**Gold Evaluation.** Results for VAD variables on gold data are given in Table 3. As can be seen, our lexicons show a good correlation with human judgment and do so robustly, even for less-resourced languages, such as Indonesian (id), Turkish (tr), or Croatian (hr), and across affective variables. Perhaps the strongest negative outliers are the Arousal results for the two Chinese datasets (zh), which are likely to result from the low reliability of the gold ratings (see below).

| ID | Shared | (%) | Val | Aro | Dom |
|---|---|---|---|---|---|
| en1 | 1032 | 100 | **.94** (.87) | **.76** (.67) | **.88** (.76) |
| en2 | 1034 | 100 | **.92** (.92) | .71 (**.73**) | .78 (**.82**) |
| es1 | 612 | 59 | **.91** (.88) | **.71** (.70) | .82 (**.83**) |
| es2 | 7685 | 54 | .79 (**.82**) | .64 (**.74**) | — |
| es3 | 363 | 41 | .91 | .73 | — |
| de1 | 677 | 67 | **.89** (.87) | .78 (**.80**) | .68 (**.74**) |
| de2 | 2329 | 80 | .75 | .64 | — |
| de3 | 916 | 91 | .80 | .67 | — |
| pl1 | 2271 | 46 | **.83** (.74) | **.74** (.70) | .60 (**.69**) |
| pl2 | 1381 | 47 | .82 | .61 | — |
| zh1 | 1685 | 60 | .84 (**.85**) | .56 (**.63**) | — |
| zh2 | 701 | 63 | .84 | .44 | — |
| it | 660 | 58 | **.89** (.86) | .63 (**.65**) | **.76** (.75) |
| pt | 645 | 62 | **.89** (.86) | .71 (**.71**) | **.75** (.73) |
| nl | 2064 | 48 | **.85** (.79) | .58 (**.74**) | — |
| id | 696 | 46 | **.84** (.80) | **.64** (.60) | **.63** (.58) |
| el | 633 | 61 | .86 | .50 | .74 |
| tr1 | 721 | 35 | .75 | .57 | — |
| hr | 1331 | 44 | .81 | .66 | — |
| **Mn** (all) | | | **.85** | **.65** | **.74** |
| **Mn** (vs. monolingual) | | | **.87** (.84) | .68 (**.70**) | **.74** (.74) |

Table 3: Gold evaluation results for VAD (**Val**ence, **Aro**usal, **Dom**inance) in Pearson's $r$. Parentheses give comparative monolingual results from Buechel and Hahn (2018b). **Shared** words between `TargetGold` and `TargetPred-test`; **(%)**: percentage relative to `TargetGold`; **Mn** (all): mean over all datasets; **Mn** (vs. monolingual): mean over datasets with comparative results.

We compare these results against those from Buechel and Hahn (2018b) which were acquired on the respective `TargetGold` dataset in a monolingual fashion using 10-fold cross-validation (10-

| ID | Shared | (%) | Joy | Ang | Sad | Fea | Dis |
|---|---|---|---|---|---|---|---|
| en3 | 1033 | 99 | .89 | .83 | .80 | .82 | .78 |
| es4 | 363 | 41 | .86 | .84 | .84 | .84 | .76 |
| es5 | 6096 | 58 | .64 | .72 | .72 | .72 | .63 |
| es6 | 992 | 43 | .80 | .74 | .71 | .72 | .68 |
| de4 | 848 | 43 | .80 | .66 | .52 | .68 | .42 |
| pl3 | 1381 | 47 | .78 | .71 | .66 | .69 | .71 |
| tr2 | 721 | 35 | .77 | .69 | .71 | .70 | .65 |
| **Mean** | | | **.79** | **.74** | **.71** | **.74** | **.66** |

Table 4: Gold evaluation results for BE5 (**Joy**, **Ang**er, **Sad**ness, **Fear**, **Dis**gust) in Pearson's $r$. **Shared** words between `TargetGold` and `TargetPred-test`; **(%)**: percentage relative to `TargetGold`; **Mean** over all datasets.

CV). We admit that those results are not fully comparable to those presented here because we use fixed splits rather than 10-CV. Nevertheless, we find that the results of our cross-lingual set-up are more than competitive, outperforming the monolingual results from Buechel and Hahn (2018b) in 17 out of 30 cases (mainly for Valence and Dominance, less often for Arousal). This is surprising since we use an otherwise identical model and training procedure. We conjecture that the large size of the English `Source` lexicon, compared to most `TargetGold` lexicons, more than compensates for error-prone machine translation.

Table 4 shows the results for BE5 datasets which are in line with the VAD results. Regarding the ordering of the emotional variables, again, we find Valence to be the easiest one to predict, Arousal the hardest, whereas basic emotions and Dominance take a middle ground.

**Comparison against Human Reliability.** We base this analysis on *inter-study reliability* (ISR), a rather strong criterion for human performance. ISR is computed, per variable, as the correlation between the ratings from two distinct annotation studies (Warriner et al., 2013). Hence, this analysis is restricted to languages where more than one gold lexicon exists per emotion format. We intersect the entries from both gold standards as well as the respective `TargetPred-test` set and compute the correlation between all three pairs of lexicons. If our lexicon agrees more with one of the gold standards than the two gold standards agree with each other, we consider this as an indicator for *superhuman* reliability (Buechel and Hahn, 2018b).

As shown in Table 5, our lexicons are often competitive with human reliability for Valence (especially for English and Chinese), but outperform

| Gold1 | Gold2 | Shared | Emo | G1vsG2 | G1vsPr | G2vsPr |
|---|---|---|---|---|---|---|
| en1 | en2 | 1032 | V | **.953** | .941 | .922 |
| | | | A | .760 | **.761** | .711 |
| | | | D | .794 | **.879** | .782 |
| es1 | es2 | 610 | V | **.976** | .905 | .912 |
| | | | A | **.758** | .714 | .725 |
| es2 | es3 | 222 | V | **.976** | .906 | .907 |
| | | | A | .710 | **.724** | .691 |
| de2 | de3 | 498 | V | **.963** | .806 | .812 |
| | | | A | **.760** | .721 | .663 |
| pl1 | pl2 | 445 | V | **.943** | .838 | .852 |
| | | | A | .725 | **.764** | .643 |
| zh1 | zh2 | 140 | V | **.932** | .918 | .898 |
| | | | A | .482 | **.556** | .455 |

Table 5: Comparison against human performance. Correlation between two gold standards, **Gold1** and **Gold2**, with each other (**G1vsG2**), as well as with our lexicons `TargetPred-test` (**G1vsPr** and **G2vsPr**) relative to **Emo**tional variable and **Shared** number of words.

human reliability in 4 out of 6 cases for Arousal, and in the single test case for Dominance. There are no cases of overlapping gold standards for BE5.

## 6 Methodological Assumptions Revisited

This section investigates patterns in prediction quality *across* languages, validating design decisions of our methodology.

**Translation vs. Prediction.** Is it beneficial to predict new ratings for the words in `TargetMT` rather than using them as final lexicon entries straight away? For each `TargetGold` lexicon (cf. Table 2), we intersect its word material with that in `TargetMT` and `TargetPred`. Then, we compute the correlation between `TargetPred` and `TargetMT` with the gold standard. This analysis was done on the respective *train* sets because using `TargetMT` rather than `TargetPred` is only an option for entries known at training time.

Table 6 depicts the results of this comparison averaged over all gold lexicons. As hypothesized, the `TargetPred` lexicons agree, on average, more with human judgment than the `TargetMT` lexicons, suggesting that the word emotion model acts as a value-adding post-processor, partly mitigating rating inconsistencies introduced by mere translation of the lexicons. The observation holds for each individual emotion variable with particularly large benefits for Arousal, where the post-processed `TargetPred` lexicons are on average

| | Val | Aro | Dom | Joy | Ang | Sad | Fea | Dis |
|---|---|---|---|---|---|---|---|---|
| Pred | .871 | .652 | .733 | .767 | .734 | .692 | .728 | .650 |
| MT | .796 | .515 | .613 | .699 | .677 | .636 | .654 | .579 |
| Diff | .076 | .137 | .119 | .068 | .057 | .056 | .074 | .071 |

Table 6: Quality of TargetMT vs. TargetPred in terms of average Pearson correlation over all languages and gold standards. Diff $\coloneqq$ Pred $-$ MT.

14%-points better compared to the translation-only TargetMT lexicons. This seems to indicate that lexical Arousal is less consistent between translational equivalents compared to other emotional meaning components like Valence and Sadness, which appear to be more robust against translation.

**Gold vs. Silver Evaluation.** How meaningful is silver evaluation without gold data? We compute the Pearson correlation between gold and silver evaluation results across languages per emotion variable. For languages where we consider multiple datasets during gold evaluation, we first average the gold evaluation results for each emotion variable. As can be seen from Table 7, the correlation values range between $r = .91$ for Joy and $r = .27$ for Disgust. This relatively large dispersion is not surprising when we take into account that we correlate very small data series (for Valence and Arousal there are just 12 languages for which both gold and silver evaluation results are available; for BE5 there are only 5 such languages). However, the mean over all correlation values in Table 7 is .64, indicating that there is a relatively strong correlation between both types of evaluation. This suggests that the silver evaluation may be used as a rather reliable proxy of lexicon quality even in the absence of language-specific gold data.

| | Val | Aro | Dom | Joy | Ang | Sad | Fea | Dis |
|---|---|---|---|---|---|---|---|---|
| #Lg | 12 | 12 | 8 | 5 | 5 | 5 | 5 | 5 |
| $r$ | .54 | .57 | .52 | .91 | .85 | .57 | .87 | .27 |

Table 7: Agreement between gold and silver evaluation across languages in Pearson's $r$ relative to the number of applicable languages ("#Lg").

## 7 Conclusion

Emotion lexicons are at the core of sentiment analysis, a rapidly flourishing field of NLP. Yet, despite large community efforts, the coverage of existing lexicons is still limited in terms of languages, size, and types of emotion variables. While there are techniques to tackle these three forms of sparsity in isolation, we introduced a methodology which allows us to cope with them simultaneously by jointly combining emotion representation mapping, machine translation, and embedding-based lexicon expansion.

Our study is "large-scale" in many respects. We created representationally complex lexicons—comprising 8 distinct emotion variables—for 91 languages with up to 2 million entries each. The evaluation of the generated lexicons featured 26 manually annotated datasets spanning 12 diverse languages. The predicted ratings showed consistently high correlation with human judgment, compared favorably with state-of-the-art monolingual approaches to lexicon expansion and even surpassed human inter-study reliability in some cases.

The sheer number of test sets we used allowed us to validate fundamental methodological assumptions underlying our approach. Firstly, the evaluation procedure, which is integrated into the generation methodology, allows us to reliably estimate the quality of resulting lexicons, *even without target language gold standard*. Secondly, our data suggests that embedding-based word emotion models can be used as a *repair mechanism*, mitigating poor target-language emotion estimates acquired by simple word-to-word translation.

Future work will have to deepen the way we deal with word sense ambiguity by way of exchanging the simplifying type-level approach our current work is based on with a semantically more informed sense-level approach. A promising direction would be to combine a multilingual sense inventory such as BABELNET (Navigli and Ponzetto, 2012) with sense embeddings (Camacho-Collados and Pilehvar, 2018).

## Acknowledgments

We would like to thank the anonymous reviewers for their helpful suggestions and comments, and Tinghui Duan, JULIE LAB, for assisting us with the Chinese gold data. This work was partially funded by the German Federal Ministry for Economic Affairs and Energy (funding line "Big Data in der makroökonomischen Analyse" [Big data in macroeconomic analysis]; Fachlos 2; GZ 23305/003#002).

## References


Mohamed Abdalla and Graeme Hirst. 2017. Cross-lingual sentiment analysis without (good) translation. In *IJCNLP 2017 — Proceedings of the 8th International Joint Conference on Natural Language Processing*, volume 1: Long Papers, pages 506–515, Taipei, Taiwan, November 27 – December 1, 2017.

Muhammad Abdul-Mageed and Lyle H. Ungar. 2017. EmoNet: Fine-grained emotion detection with gated recurrent neural networks. In *ACL 2017 — Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, volume 1: Long Papers, pages 718–728, Vancouver, British Columbia, Canada, July 30 – August 4, 2017.

Soraia M. Alarcão and Manuel J. Fonseca. 2017. Identifying emotions in images from valence and arousal ratings. *Multimedia Tools and Applications*, 77(13):17413–17435.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SENTIWORDNET 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC 2010 — Proceedings of the 7th International Conference on Language Resources and Evaluation*, pages 2200–2204, La Valletta, Malta, May 17–23, 2010.

Jeremy Barnes, Roman Klinger, and Sabine Schulte im Walde. 2018. Bilingual sentiment embeddings: Joint projection of sentiment across languages. In *ACL 2018 — Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, volume 1: Long Papers, pages 2483–2493, Melbourne, Victoria, Australia, July 15–20, 2018.

Jeremy Barnes, Samia Touileb, Lilja Øvrelid, and Erik Velldal. 2019. Lexicon information in neural sentiment analysis: a multi-task learning approach. In *NoDaLiDa 2019 — Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 175–186, Turku, Finland, September 30 – October 2, 2019.

Yves Bestgen. 2008. Building affective lexicons from specific corpora for automatic sentiment analysis. In *LREC 2008 — Proceedings of the 6th International Conference on Language Resources and Evaluation*, pages 496–500, Marrakesh, Morocco, May 28–30, 2008.

Laura-Ana-Maria Bostan and Roman Klinger. 2018. An analysis of annotated corpora for emotion classification in text. In *COLING 2018 — Proceedings of the 27th International Conference on Computational Linguistics*, pages 2104–2119, Santa Fe, New Mexico, USA, August 20–26, 2018.

Margaret M. Bradley and Peter J. Lang. 1994. Measuring emotion: The Self-Assessment Manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1):49–59.

Margaret M. Bradley and Peter J. Lang. 1999. Affective norms for English words (ANEW): Stimuli, instruction manual and affective ratings. Technical Report C-1, The Center for Research in Psychophysiology, University of Florida, Gainesville, Florida, USA.

Margaret M. Bradley and Peter J. Lang. 2010. Affective norms for English words (ANEW): Stimuli, instruction manual and affective ratings. Technical Report C-2, University of Florida, Gainesville, Forida, USA.

Benny B. Briesemeister, Lars Kuchinke, and Arthur M. Jacobs. 2011. Discrete Emotion Norms for Nouns: Berlin Affective Word List (DENN–BAWL). *Behavior Research Methods*, 43(2):#441.

Sven Buechel and Udo Hahn. 2016. Emotion analysis as a regression problem: Dimensional models and their implications on emotion representation and metrical evaluation. In *ECAI 2016 — Proceedings of the 22nd European Conference on Artificial Intelligence*, pages 1114–1122, The Hague, The Netherlands, August 29 – September 2, 2016.

Sven Buechel and Udo Hahn. 2017. EMOBANK: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *EACL 2017 — Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, volume 2: Short Papers, pages 578–585, Valencia, Spain, April 3–7, 2017.

Sven Buechel and Udo Hahn. 2018a. Emotion representation mapping for automatic lexicon construction (mostly) performs on human level. In *COLING 2018 — Proceedings of the 27th International Conference on Computational Linguistics*, pages 2892–2904, Santa Fe, New Mexico, USA, August 20–26, 2018.

Sven Buechel and Udo Hahn. 2018b. Word emotion induction for multiple languages as a deep multi-task learning problem. In *NAACL-HLT 2018 — Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1: Long Papers, pages 1907–1918, New Orleans, Louisiana, USA, June 1–6, 2018.

Sven Buechel, Johannes Hellrich, and Udo Hahn. 2016. Feelings from the past: Adapting affective lexicons for historical emotion analysis. In *LT4DH 2016 — Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities @ COLING 2016*, pages 54–61, Osaka, Japan, December 11, 2016.

Rafael A. Calvo and Sunghwan Mac Kim. 2013. Emotions in text: Dimensional and categorical models. *Computational Intelligence*, 29(3):527–543.

José Camacho-Collados and Mohammad Taher Pilehvar. 2018. From word to sense embeddings: A survey on vector representations of meaning. *Journal of Artificial Intelligence Research*, 63:743–788.

Yanqing Chen and Steven Skiena. 2014. Building sentiment lexicons for all major languages. In *ACL 2014 — Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 2: Short Papers, pages 383–389, Baltimore, Maryland, USA, June 23–25, 2014.

Ta-Chung Chi and Yun-Nung Chen. 2018. CLUSE : Cross-Lingual Unsupervised Sense Embeddings. In *EMNLP 2018 — Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 271–281, Brussels, Belgium, October 31 – November 4, 2018.

Paul Cook and Suzanne Stevenson. 2010. Automatically identifying changes in the semantic orientation of words. In *LREC 2010 — Proceedings of the 7th International Conference on Language Resources and Evaluation*, pages 28–34, La Valletta, Malta, May 17–23, 2010.

Bojana Ćoso, Marc Guasch, Pilar Ferré, and José Antonio Hinojosa. 2019. Affective and concreteness norms for 3,022 Croatian words. *Quarterly Journal of Experimental Psychology*, 72(9):2302–2312.

Per Davidson and Åse Innes-Ker. 2014. Valence and arousal norms for Swedish affective words. *Lund Psychological Reports*, 14:#2.

Luna De Bruyne, Pepa Atanasova, and Isabelle Augenstein. 2019. Joint emotion label space modelling for affect lexica. *arXiv:1911.08782 [cs]*.

Steven Du and Xi Zhang. 2016. Aicyber's system for IALP 2016 Shared Task: Character-enhanced word vectors and boosted neural networks. In *IALP 2016 — Proceedings of the 2016 International Conference on Asian Language Processing*, pages 161–163, Tainan, Taiwan, November 21–23, 2016.

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

Pilar Ferré, Marc Guasch, Natalia Martínez-García, Isabel Fraga, and José Antonio Hinojosa. 2017. Moved by words: Affective ratings for a set of 2,266 Spanish words in five discrete emotion categories. *Behavior Research Methods*, 49(3):1082–1094.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *LREC 2018 — Proceedings of the 11th International Conference on Language Resources and Evaluation*, pages 3483–3487, Miyazaki, Japan, May 7–12, 2018.

William L. Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky. 2016. Inducing domain-specific sentiment lexicons from unlabeled corpora. In *EMNLP 2016 — Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 595–605, Austin, Texas, USA, November 1–5, 2016.

Vasileios Hatzivassiloglou and Kathleen R. McKeown. 1997. Predicting the semantic orientation of adjectives. In *ACL-EACL 1997 — Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics & 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 174–181, Madrid, Spain, July 7–12, 1997.

Johannes Hellrich, Sven Buechel, and Udo Hahn. 2018. JESEMEE: Interleaving semantics and emotions in a Web service for the exploration of language change phenomena. In *COLING 2018 — Proceedings of the 27th International Conference on Computational Linguistics*, volume System Demonstrations, pages 10–14, Santa Fe, New Mexico, USA, August 20–26, 2018.

José Antonio Hinojosa, Natalia Martínez-García, Cristina Villalba-García, Uxia Fernández-Folgueiras, Alberto Sánchez-Carmona, Miguel Angel Pozo, and Pedro R. Montoro. 2016. Affective norms of 875 Spanish words for five discrete emotional categories and two emotional dimensions. *Behavior Research Methods*, 48(1):272–284.

Holger Hoffmann, Andreas Scheck, Timo Schuster, Steffen Walter, Kerstin Limbrecht-Ecklundt, Harald C. Traue, and Henrik Kessler. 2012. Mapping discrete emotions into the dimensional space: An empirical approach. In *SMC 2012 — Proceedings of the 2012 IEEE International Conference on Systems, Man, and Cybernetics*, pages 3316–3320, Seoul, Korea, October 14–17, 2012.

Markus J. Hofmann, Lars Kuchinke, Sascha Tamm, Melissa L.-H. Võ, and Arthur M. Jacobs. 2009. Affective processing within 1/10th of a second: High arousal is necessary for early facilitative processing of negative but not positive words. *Cognitive, Affective, & Behavioral Neuroscience*, 9(4):389–397.

Clayton J. Hutto and Eric Gilbert. 2014. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *ICWSM 2014 — Proceedings of the 8th International AAAI Conference on Weblogs and Social Media*, pages 216–225, Ann Arbor, Michigan, USA, June 1–4, 2014.

Kamil K. Imbir. 2016. Affective Norms for 4900 Polish Words Reload (ANPW_R): Assessments for valence, arousal, dominance, origin, significance, concreteness, imageability and, age of acquisition. *Frontiers in Psychology*, 7:#1081.

Philipp Kanske and Sonja A. Kotz. 2010. Leipzig Affective Norms for German: A reliability study. *Behavior Research Methods*, 42(4):987–991.

Aycan Kapucu, Aslı Kılıç, Yıldız Özkılıç, and Bengisu Sarıbaz. 2018. Turkish emotional word norms for arousal, valence, and discrete emotion categories. *Psychological Reports*, pages 1–22. [Available online Dec 4, 2018].

Sopan Khosla, Niyati Chhaya, and Kushal Chawla. 2018. AFF2VEC : Affect–enriched distributional word representations. In *COLING 2018 — Proceedings of the 27th International Conference on Computational Linguistics*, pages 2204–2218, Santa Fe, New Mexico, USA, August 20–26, 2018.

Diederik Kingma and Jimmy Ba. 2015. ADAM: A method for stochastic optimization. In *ICLR 2015 — Proceedings of the 3rd International Conference on Learning Representations*, San Diego, California, USA, May 7–9, 2015.

Svetlana Kiritchenko and Saif M. Mohammad. 2016. Capturing reliable fine-grained sentiment associations by crowdsourcing and best-worst scaling. In *NAACL-HLT 2016 — Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 811–817, San Diego, California, USA, June 12–17, 2016.

Maximilian Köper and Sabine Schulte im Walde. 2016. Automatically generated affective norms of abstractness, arousal, imageability and valence for 350,000 German lemmas. In *LREC 2016 — Proceedings of the 10th International Conference on Language Resources and Evaluation*, pages 2595–2598, Portorož, Slovenia, May 23–28, 2016.

Agnieszka Landowska. 2018. Towards new mappings between emotion representation models. *Applied Sciences*, 8(2):#274.

Nicolas Leveau, Sandra Jhean-Larose, Guy Denhière, and Ba-Linh Nguyen. 2012. Validating an interlingual metanorm for emotional analysis of texts. *Behavior Research Methods*, 44(4):1007–1014.

Minglei Li, Qin Lu, Yunfei Long, and Lin Gui. 2017. Inferring affective meanings of words from word embedding. *IEEE Transactions on Affective Computing*, 8(4):443–456.

Ying Li, Tomas Engelthaler, Cynthia S. Q. Siew, and Thomas T. Hills. 2019. The MACROSCOPE: A tool for examining the historical structure of language. *Behavior Research Methods*, 51(4):1864–1877.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *ACL 2018 — Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, volume 1: Long Papers, pages 174–184, Melbourne, Victoria, Australia, July 15–20, 2018.

Saif Mohammad and Felipe Bravo-Marquez. 2017. WASSA-2017 Shared Task on Emotion Intensity. In *WASSA 2017 — Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis @ EMNLP 2017*, pages 34–49, Copenhagen, Denmark, September 8, 2017.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SEMEVAL-2018 Task 1: Affect in Tweets. In *SemEval 2018 — Proceedings of the 12th International Workshop on Semantic Evaluation @ NAACL-HLT 2018*, pages 1–17, New Orleans, Louisiana, USA, June 5–6, 2018.

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

Catherine Monnier and Arielle Syssau. 2008. Semantic contribution to verbal short-term memory: Are pleasant words easier to remember than neutral words in serial recall and serial recognition? *Memory & Cognition*, 36(1):35–42.

Maria Montefinese, Ettore Ambrosini, Beth Fairfield, and Nicola Mammarella. 2014. The adaptation of the Affective Norms for English Words (ANEW) for Italian. *Behavior Research Methods*, 46(3):887–903.

Agnes Moors, Jan De Houwer, Dirk Hermans, Sabine Wanmaker, Kevin van Schie, Anne-Laura Van Harmelen, Maarten De Schryver, Jeffrey De Winne, and Marc Brysbært. 2013. Norms of valence, arousal, dominance, and age of acquisition for 4,300 Dutch words. *Behavior Research Methods*, 45(1):169–177.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BABELNET: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Elisavet Palogiannidi, Polychronis Koutsakis, Elias Iosif, and Alexandros Potamianos. 2016. Affective lexicon creation for the Greek language. In *LREC 2016 — Proceedings of the 10th International Conference on Language Resources and Evaluation*, pages 2867–2872, Portorož, Slovenia, May 23–28, 2016.

Sungjoon Park, Jiseon Kim, Jaeyeol Jeon, Heeyoung Park, and Alice Oh. 2019. Toward dimensional emotion detection from categorical emotion annotations. *arXiv:1911.02499 [cs, eess]*.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. SCIKIT-LEARN: Machine learning in PYTHON. *Journal of Machine Learning Research*, 12(85):2825–2830.

Ana P. Pinheiro, Marcelo Dias, João Pedrosa, and Ana P. Soares. 2017. Minho Affective Sentences (MAS): Probing the roles of sex, mood, and empathy in affective ratings of verbal stimuli. *Behavior Research Methods*, 49(2):698–716.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. In Robert Plutchik and Henry Kellerman, editors, *Emotion: Theory, Research and Experience*, volume 1: Theories of Emotion, pages 3–33. Academic Press, New York, NY, USA.

Jaime Redondo, Isabel Fraga, Isabel Padrón, and Montserrat Comesaña. 2007. The Spanish adaptation of ANEW (Affective Norms for English Words). *Behavior Research Methods*, 39(3):600–605.

Nils Reimers and Iryna Gurevych. 2017. Reporting score distributions makes a difference: Performance study of LSTM-networks for sequence tagging. In *EMNLP 2017 — Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 338–348, Copenhagen, Denmark, September 9–11, 2017.

Monika Riegel, Małgorzata Wierzba, Marek Wypych, Łukasz Żurawski, Katarzyna Jednoróg, Anna Grabowska, and Artur Marchewka. 2015. Nencki Affective Word List (NAWL): The cultural adaptation of the Berlin Affective Word List–Reloaded (BAWL–R) for Polish. *Behavior Research Methods*, 47(4):1222–1236.

Sara Rosenthal, Preslav I. Nakov, Svetlana Kiritchenko, Saif M. Mohammad, Alan Ritter, and Veselin Stoyanov. 2015. SEMEVAL 2015 Task 10: Sentiment Analysis in Twitter. In *SemEval 2015 — Proceedings of the 9th International Workshop on Semantic Evaluation @ NAACL-HLT 2015*, pages 451–463, Denver, Colorado, USA, June 4–5, 2015.

Sascha Rothe, Sebastian Ebert, and Hinrich Schütze. 2016. Ultradense word embeddings by orthogonal transformation. In *NAACL 2016 — Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 767–777, San Diego, California, USA, June 12–17, 2016.

David S. Schmidtke, Tobias Schröder, Arthur M. Jacobs, and Markus Conrad. 2014. ANGST: Affective Norms for German Sentiment Terms, derived from the Affective Norms for English Words. *Behavior Research Methods*, 46(4):1108–1118.

João Sedoc, Daniel Preoţiuc-Pietro, and Lyle H. Ungar. 2017. Predicting emotional word ratings using distributional representations and signed clustering. In *EACL 2017 — Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, volume 2: Short Papers, pages 564–571, Valencia, Spain, April 3–7, 2017.

Samira Shaikh, Kit Cho, Tomek Strzalkowski, Laurie Feldman, John Lien, Ting Liu, and George Aaron Broadwell. 2016. ANEW+ : Automatic expansion and validation of Affective Norms of Words lexicons in multiple languages. In *LREC 2016 — Proceedings of the 10th International Conference on Language Resources and Evaluation*, pages 1127–1132, Portorož, Slovenia, May 23–28, 2016.

Agnes Sianipar, Pieter van Groenestijn, and Ton Dijkstra. 2016. Affective meaning, concreteness, and subjective frequency norms for Indonesian words. *Frontiers in Psychology*, 7:#1907.

Ana Paula Soares, Montserrat Comesaña, Ana P. Pinheiro, Alberto Simões, and Carla Sofia Frade. 2012. The adaptation of the Affective Norms for English Words (ANEW) for European Portuguese. *Behavior Research Methods*, 44(1):256–269.

Hans Stadthagen-González, Pilar Ferré, Miguel A. Pérez-Sánchez, Constance Imbault, and José Antonio Hinojosa. 2018. Norms for 10,491 Spanish words for five discrete emotions: Happiness, disgust, anger, fear, and sadness. *Behavior Research Methods*, 50(5):1943–1952.

Hans Stadthagen-González, Constance Imbault, Miguel A. Pérez-Sánchez, and Marc Brysbært. 2017. Norms of valence and arousal for 14,031 Spanish words. *Behavior Research Methods*, 49(1):111–123.

Ryan A. Stevenson, Joseph A. Mikels, and Thomas W. James. 2007. Characterization of the Affective Norms for English Words by discrete emotional categories. *Behavior Research Methods*, 39(4):1020–1024.

Carlo Strapparava and Alessandro Valitutti. 2004. WORDNET-AFFECT: An affective extension of WORDNET. In *LREC 2004 — Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 1083–1086, Lisbon, Portugal, May 24–30, 2004.

Enrica Troiano, Sebastian Padó, and Roman Klinger. 2019. Crowdsourcing and validating event-focused emotion corpora for German and English. In *ACL 2019 — Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4005–4011, Florence, Italy, July 28 – August 2, 2019.

Peter D. Turney and Michael L. Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, 21(4):315–346.

Melissa L.-H. Võ, Markus Conrad, Lars Kuchinke, Karolina Urton, Markus J. Hofmann, and Arthur M. Jacobs. 2009. The Berlin Affective Word List Reloaded (BAWL–R). *Behavior Research Methods*, 41(2):534–538.

Jin Wang, Liang-Chih Yu, K. Robert Lai, and Xuejie Zhang. 2016. Community-based weighted graph model for valence-arousal prediction of affective

words. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(11):1957–1968.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbært. 2013. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, 45(4):1191–1207.

Małgorzata Wierzba, Monika Riegel, Marek Wypych, Katarzyna Jednoróg, Paweł Turnau, Anna Grabowska, and Artur Marchewka. 2015. Basic emotions in the Nencki Affective Word List (NAWL BE): New method of classifying emotional stimuli. *PLoS ONE*, 10(7):#e0132305.

Zhao Yao, Jia Wu, Yanyan Zhang, and Zhenhong Wang. 2017. Norms of valence, arousal, concreteness, familiarity, imageability, and context availability for 1,100 Chinese words. *Behavior Research Methods*, 49(4):1374–1385.

Liang-Chih Yu, Lung-Hao Lee, Shuai Hao, Jin Wang, Yunchao He, Jun Hu, K. Robert Lai, and Xuejie Zhang. 2016. Building Chinese affective resources in valence-arousal dimensions. In *NAACL-HLT 2016 — Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 540–545, San Diego, California, USA, June 12–17, 2016.

Liang-Chih Yu, Jin Wang, K. Robert Lai, and Xuejie Zhang. 2017. Refining word embeddings for sentiment analysis. In *EMNLP 2017 — Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 534–539, Copenhagen, Denmark, September 9–11, 2017.

Feng Zhou, Shu Kong, Charless C. Fowlkes, Tao Chen, and Baiying Lei. 2020. Fine-grained facial expression analysis using dimensional emotion model. *Neurocomputing*. [Available online Jan 23, 2020].

# A Appendices

## A.1 Data Preparation

The exact design of the Source train-dev-test split is as follows: All entries (words plus ratings) from all splits are taken from Warriner et al. (2013). The data was then partitioned based on the overlap with the two precursory versions by Bradley and Lang (1999) (the original ANEW) and Bradley and Lang (2010) (an early extended version of ANEW roughly twice as large). Source-test was built by intersecting the lexicon from Warriner et al. (2013) with the original ANEW. A similar process was applied for Source-dev: we intersected the words from Warriner et al. (2013) and Bradley and Lang (2010) and removed the ones present in Source-test. Lastly, Source-train is made up by all words from Warriner et al. (2013) which are neither in Source-test nor in Source-dev. The reason why the ratings in Source are taken exclusively from Warriner et al. (2013) is that these are distributed under a more permissive license compared to their precursors.

We removed multi-token entries (e.g., *boa constrictor*) and entries with upper case characters (e.g., *Budweiser*) from all data splits of Source, thus restricting the lexicon to single-token, non-proper noun entries to make it more suitable for word embedding-based research. All splits combined have 13,791 entries (train: 11,463, dev: 1,296, test: 1,032), thus removing less than 1% from the original lexicon.[5]

Regarding the remaining gold standards, the only cases which needed additional preparation or cleansing steps were zh1 (Yu et al., 2016) and zh2 (Yao et al., 2017). zh1 was created and is distributed using traditional Chinese characters, whereas the embedding model by Grave et al. (2018) employs simplified ones. Therefore, we converted zh1 into simplified characters using GOOGLE TRANSLATE[6] prior to evaluation.

While manually examining the zh2 lexicon, we noticed several cases where the ratings seemed rather counter-intuitive (e.g., seemingly positive words which received very negative ratings). We contacted the authors who confirmed the problem and sent us a corrected version. We did not find any such problems in the second version. We consulted with a Chinese native speaker for both of these procedures regarding the zh1 and zh2 lexicons.

## A.2 Model Training and Implementation

Training of the MTLFFN model closely followed the procedure specified by Buechel and Hahn (2018b): For each language, the model was trained for roughly 15k iterations (exactly 168 epochs) with a batch size of 128 using the Adam optimizer (Kingma and Ba, 2015) with learning rate $10^{-3}$, and .5 dropout on the hidden layers and .2 on the input layer. As nonlinear activation function we used leaky ReLU with "leakage" of 0.01.

Embedding vectors are the only model input. They have 300 dimensions for every language, independent of their respective training data size (Grave et al., 2018). Since the automatic translation of Source is not guaranteed to result in single-word translations, we use the following workaround to derive embedding vectors for multi-token translations: If the translation as a whole cannot be found in the embedding model, the multi-token term gets split up into its constituent parts, using spaces, apostrophes or hyphens as separators. Each substring is looked up in the embedding model, the averaged vector is taken as input. If no substring is recognized, we use the zero vector instead. We also use the zero vector for single-token entries in TargetMT that are missing in the embeddings.

Since Buechel and Hahn (2018b) considered only VAD but not BE5 datasets, we conducted a development experiment on the TargetMT-dev sets for all 91 languages where we assessed whether MTL is advantageous for BE5 variables as well, or for a combination of VAD and BE5 variables. We found that MTL improved performance when applied separately among all VAD and BE5 variables. Yet, when jointly learning all eight emotion variables, the results were somewhat inconclusive. Performance *increased* for BE5, but *decreased* for VAD. Hence, for lexicon creation, we took a cautious approach and trained *two separate models per language*, one for VAD, the other for BE5. An analysis of MTL across VAD and BE5 is left for future work.

The MTLFFN model is implemented in PYTORCH, adapting part of the TENSORFLOW code from Buechel and Hahn (2018b). The ridge regression baseline model is implemented with SCIKIT-LEARN (Pedregosa et al., 2011) using default parameters.

[5] The data split is available at: https://github.com/JULIELab/XANEW

[6] In this case the regular Web application, not the API, was used: https://translate.google.com/

| No. | ISO | Full Name | Size | Val | Aro | Dom | Joy | Ang | Sad | Fea | Dis | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | en | English | 2,000,004 | .94 | .76 | .88 | .90 | .91 | .90 | .89 | .89 | .88 |
| 2 | es | Spanish | 2,001,183 | .89 | .70 | .80 | .83 | .86 | .85 | .82 | .81 | .82 |
| 3 | it | Italian | 2,001,137 | .88 | .69 | .81 | .82 | .85 | .84 | .82 | .81 | .81 |
| 4 | de | German | 2,000,507 | .89 | .66 | .81 | .82 | .84 | .82 | .80 | .81 | .81 |
| 5 | sv | Swedish | 2,000,980 | .87 | .64 | .80 | .82 | .84 | .82 | .81 | .80 | .80 |
| 6 | pt | Portuguese | 2,001,078 | .86 | .70 | .78 | .78 | .83 | .81 | .78 | .82 | .79 |
| 7 | id | Indonesian | 2,002,221 | .85 | .67 | .79 | .78 | .82 | .80 | .79 | .77 | .79 |
| 8 | hu | Hungarian | 2,000,975 | .86 | .67 | .79 | .80 | .82 | .79 | .77 | .79 | .79 |
| 9 | fr | French | 2,001,517 | .85 | .65 | .79 | .78 | .82 | .81 | .78 | .81 | .78 |
| 10 | fi | Finnish | 2,000,841 | .86 | .64 | .79 | .81 | .82 | .78 | .77 | .80 | .78 |
| 11 | ro | Romanian | 2,001,501 | .85 | .65 | .78 | .78 | .82 | .81 | .79 | .78 | .78 |
| 12 | cs | Czech | 2,001,203 | .84 | .64 | .77 | .78 | .82 | .80 | .79 | .79 | .78 |
| 13 | pl | Polish | 2,001,460 | .85 | .63 | .78 | .80 | .82 | .80 | .78 | .78 | .78 |
| 14 | nl | Dutch | 2,000,721 | .85 | .64 | .78 | .77 | .80 | .79 | .77 | .78 | .77 |
| 15 | no | Norwegian (Bokmål) | 2,000,876 | .84 | .63 | .77 | .78 | .82 | .78 | .78 | .78 | .77 |
| 16 | tr | Turkish | 2,002,489 | .84 | .62 | .78 | .78 | .80 | .80 | .75 | .77 | .77 |
| 17 | ru | Russian | 2,001,317 | .82 | .64 | .75 | .80 | .81 | .77 | .77 | .77 | .77 |
| 18 | el | Greek | 2,001,704 | .82 | .63 | .76 | .78 | .80 | .78 | .77 | .78 | .77 |
| 19 | uk | Ukrainian | 2,001,261 | .83 | .63 | .77 | .78 | .80 | .77 | .76 | .77 | .76 |
| 20 | et | Estonian | 2,001,125 | .83 | .59 | .75 | .77 | .81 | .78 | .77 | .78 | .76 |
| 21 | ca | Catalan | 2,001,538 | .84 | .60 | .80 | .77 | .79 | .78 | .76 | .74 | .76 |
| 22 | da | Danish | 2,000,654 | .84 | .61 | .77 | .78 | .79 | .77 | .73 | .79 | .76 |
| 23 | lv | Latvian | 1,642,923 | .82 | .63 | .75 | .76 | .79 | .78 | .76 | .77 | .76 |
| 24 | lt | Lithuanian | 2,001,306 | .83 | .63 | .77 | .75 | .79 | .77 | .75 | .76 | .76 |
| 25 | bg | Bulgarian | 2,001,391 | .82 | .60 | .76 | .75 | .77 | .77 | .73 | .76 | .74 |
| 26 | he | Hebrew | 2,001,984 | .80 | .62 | .72 | .76 | .78 | .76 | .74 | .75 | .74 |
| 27 | zh | Chinese | 2,001,799 | .79 | .60 | .75 | .72 | .77 | .75 | .75 | .73 | .73 |
| 28 | mk | Macedonian | 1,356,402 | .82 | .54 | .75 | .77 | .76 | .73 | .72 | .74 | .73 |
| 29 | af | Afrikaans | 883,464 | .80 | .58 | .74 | .76 | .75 | .74 | .71 | .74 | .73 |
| 30 | tl | Tagalog | 716,272 | .80 | .56 | .76 | .70 | .77 | .76 | .74 | .72 | .73 |
| 31 | sk | Slovak | 2,001,221 | .80 | .60 | .75 | .74 | .74 | .73 | .71 | .73 | .72 |
| 32 | sq | Albanian | 1,169,697 | .80 | .57 | .73 | .75 | .75 | .75 | .72 | .72 | .72 |
| 33 | az | Azerbaijani | 2,002,146 | .81 | .60 | .73 | .74 | .75 | .73 | .70 | .71 | .72 |
| 34 | mn | Mongolian | 608,598 | .78 | .57 | .73 | .71 | .78 | .72 | .74 | .74 | .72 |
| 35 | hy | Armenian | 2,001,329 | .80 | .52 | .72 | .75 | .77 | .73 | .71 | .73 | .72 |
| 36 | eo | Esperanto | 2,001,575 | .77 | .55 | .71 | .72 | .76 | .74 | .73 | .73 | .71 |
| 37 | sl | Slovenian | 1,992,272 | .81 | .54 | .75 | .74 | .74 | .70 | .70 | .72 | .71 |
| 38 | hr | Croatian | 2,001,570 | .78 | .56 | .71 | .72 | .74 | .71 | .71 | .73 | .71 |
| 39 | gl | Galician | 1,336,256 | .78 | .53 | .72 | .72 | .76 | .74 | .71 | .71 | .71 |
| 40 | sr | Serbian | 2,002,395 | .76 | .57 | .71 | .72 | .74 | .70 | .70 | .73 | .70 |
| 41 | ar | Arabic | 2,003,155 | .78 | .53 | .70 | .70 | .75 | .72 | .71 | .74 | .70 |
| 42 | fa | Persian | 2,003,533 | .77 | .58 | .70 | .70 | .74 | .73 | .70 | .70 | .70 |
| 43 | ms | Malay | 1,213,397 | .75 | .58 | .69 | .69 | .72 | .70 | .65 | .73 | .69 |
| 44 | mr | Marathi | 848,549 | .74 | .54 | .68 | .70 | .74 | .70 | .69 | .71 | .69 |
| 45 | ka | Georgian | 1,567,232 | .76 | .52 | .72 | .70 | .72 | .71 | .70 | .66 | .69 |
| 46 | ja | Japanese | 2,003,306 | .72 | .58 | .67 | .68 | .71 | .70 | .70 | .68 | .68 |
| 47 | hi | Hindi | 1,879,196 | .76 | .56 | .68 | .69 | .73 | .64 | .65 | .72 | .68 |
| 48 | is | Icelandic | 945,214 | .76 | .55 | .70 | .68 | .70 | .69 | .68 | .64 | .67 |
| 49 | kk | Kazakh | 1,981,562 | .72 | .53 | .65 | .67 | .73 | .69 | .67 | .70 | .67 |
| 50 | ko | Korean | 2,002,600 | .74 | .57 | .69 | .67 | .67 | .66 | .65 | .69 | .67 |
| 51 | be | Belarusian | 1,715,582 | .73 | .49 | .66 | .68 | .71 | .67 | .67 | .70 | .66 |
| 52 | bn | Bengali | 1,471,709 | .74 | .50 | .67 | .67 | .70 | .67 | .67 | .66 | .66 |
| 53 | kn | Kannada | 1,747,421 | .70 | .47 | .65 | .67 | .71 | .68 | .67 | .68 | .65 |
| 54 | cy | Welsh | 502,006 | .72 | .51 | .67 | .64 | .69 | .65 | .64 | .66 | .65 |
| 55 | ur | Urdu | 1,157,969 | .69 | .52 | .61 | .63 | .70 | .65 | .64 | .68 | .64 |
| 56 | ta | Tamil | 2,002,514 | .70 | .51 | .66 | .64 | .66 | .66 | .63 | .64 | .64 |
| 57 | eu | Basque | 1,828,013 | .70 | .46 | .66 | .64 | .68 | .67 | .64 | .64 | .64 |
| 58 | ml | Malayalam | 2,002,920 | .67 | .51 | .62 | .63 | .67 | .67 | .62 | .61 | .63 |
| 59 | gu | Gujarati | 557,270 | .69 | .46 | .62 | .61 | .67 | .65 | .63 | .64 | .62 |
| 60 | si | Sinhalese | 812,356 | .66 | .48 | .59 | .65 | .67 | .62 | .63 | .65 | .62 |
| 61 | te | Telugu | 1,880,585 | .69 | .46 | .62 | .60 | .65 | .63 | .61 | .65 | .61 |
| 62 | ne | Nepali | 580,582 | .68 | .44 | .62 | .63 | .65 | .63 | .61 | .62 | .61 |
| 63 | tg | Tajik | 508,617 | .67 | .38 | .64 | .57 | .65 | .65 | .60 | .60 | .60 |
| 64 | vi | Vietnamese | 2,008,605 | .65 | .47 | .58 | .59 | .65 | .59 | .58 | .62 | .59 |
| 65 | pa | Eastern Punjabi | 403,997 | .67 | .37 | .61 | .59 | .64 | .61 | .58 | .62 | .59 |
| 66 | bs | Bosnian | 1,124,938 | .63 | .43 | .60 | .57 | .64 | .61 | .61 | .60 | .58 |
| 67 | ky | Kirghiz | 751,902 | .65 | .37 | .61 | .56 | .64 | .62 | .59 | .60 | .58 |
| 68 | ga | Irish | 321,249 | .64 | .47 | .59 | .58 | .61 | .61 | .59 | .55 | .58 |
| 69 | fy | West Frisian | 530,054 | .61 | .43 | .54 | .53 | .60 | .59 | .55 | .58 | .56 |
| 70 | uz | Uzbek | 833,860 | .60 | .38 | .55 | .56 | .57 | .56 | .54 | .53 | .53 |
| 71 | sw | Swahili | 391,312 | .59 | .34 | .57 | .52 | .59 | .58 | .57 | .51 | .53 |
| 72 | jv | Javanese | 518,634 | .58 | .45 | .53 | .53 | .56 | .58 | .54 | .49 | .53 |
| 73 | ps | Pashto | 300,927 | .58 | .40 | .56 | .52 | .55 | .54 | .55 | .49 | .53 |
| 74 | am | Amharic | 308,109 | .56 | .31 | .52 | .48 | .53 | .54 | .52 | .47 | .49 |
| 75 | lb | Luxembourgish | 642,504 | .53 | .37 | .47 | .45 | .55 | .52 | .50 | .51 | .49 |
| 76 | su | Sundanese | 327,533 | .54 | .36 | .47 | .45 | .53 | .52 | .48 | .52 | .48 |
| 77 | th | Thai | 2,006,540 | .51 | .38 | .45 | .50 | .49 | .46 | .45 | .49 | .47 |
| 78 | km | Khmer | 247,498 | .51 | .39 | .44 | .49 | .51 | .44 | .45 | .48 | .46 |
| 79 | sd | Sindhi | 139,063 | .47 | .35 | .39 | .41 | .50 | .49 | .50 | .46 | .45 |
| 80 | yi | Yiddish | 205,727 | .49 | .34 | .40 | .43 | .50 | .47 | .45 | .44 | .44 |
| 81 | my | Burmese | 339,628 | .49 | .36 | .42 | .43 | .49 | .45 | .46 | .43 | .44 |
| 82 | la | Latin | 1,088,139 | .47 | .33 | .40 | .39 | .47 | .46 | .43 | .44 | .42 |
| 83 | mt | Maltese | 204,630 | .47 | .32 | .44 | .38 | .43 | .40 | .39 | .38 | .40 |
| 84 | gd | Scottish Gaelic | 150,694 | .45 | .36 | .39 | .40 | .36 | .36 | .35 | .33 | .38 |
| 85 | so | Somali | 177,405 | .40 | .22 | .35 | .36 | .44 | .41 | .41 | .38 | .37 |
| 86 | mg | Malagasy | 415,050 | .40 | .32 | .36 | .34 | .41 | .37 | .36 | .36 | .37 |
| 87 | ht | Haitian | 118,302 | .39 | .22 | .33 | .30 | .42 | .42 | .37 | .38 | .35 |
| 88 | ku | Kurdish (Kurmanji) | 395,645 | .37 | .22 | .33 | .33 | .34 | .33 | .31 | .35 | .32 |
| 89 | ceb | Cebuano | 2,006,001 | .34 | .22 | .29 | .34 | .36 | .32 | .33 | .34 | .32 |
| 90 | co | Corsican | 108,035 | .29 | .24 | .27 | .27 | .32 | .30 | .29 | .30 | .29 |
| 91 | yo | Yoruba | 156,764 | .24 | .08 | .19 | .18 | .24 | .21 | .21 | .26 | .20 |

Table 8: Overview of generated emotion lexicons with silver evaluation results; sorted by **Mean** performance over the eight emotional variables.