Word Emotion Induction for Multiple Languages as a Deep Multi-Task Learning Problem

Sven Buechel & Udo Hahn
{sven.buechel|udo.hahn}@uni-jena.de
Jena University Language & Information Engineering (JULIE) Lab
Friedrich-Schiller-Universität Jena, Jena, Germany
http://www.julielab.de

Abstract

Predicting the emotional value of lexical items is a well-known problem in sentiment analysis. While research has focused on polarity for quite a long time, meanwhile this early focus has been shifted to more expressive emotion representation models (such as Basic Emotions or Valence-Arousal-Dominance). This change resulted in a proliferation of heterogeneous formats and, in parallel, often small-sized, non-interoperable resources (lexicons and corpus annotations). In particular, the limitations in size hampered the application of deep learning methods in this area because they typically require large amounts of input data. We here present a solution to get around this language data bottleneck by rephrasing word emotion induction as a multi-task learning problem. In this approach, the prediction of each independent emotion dimension is considered as an individual task and hidden layers are *shared* between these dimensions. We investigate whether multi-task learning is more advantageous than single-task learning for emotion prediction by comparing our model against a wide range of alternative emotion and polarity induction methods featuring 9 typologically diverse languages and a total of 15 conditions. Our model turns out to outperform each one of them. Against all odds, the proposed deep learning approach yields *the largest gain* on *the smallest* data sets, merely composed of one thousand samples.

1 Introduction

Deep Learning (DL) has radically changed the rules of the game in NLP by dramatically boosting performance figures in almost all applications areas. Yet, one of the major premises of high-performance DL engines is their dependence on huge amounts of training data. As such, DL seems ill-suited for areas where training data are scarce, such as in the field of word emotion induction.

We will use the terms *polarity* and *emotion* here to distinguish between research focusing on "semantic orientation" (Hatzivassiloglou and McKeown, 1997) (the positiveness or negativeness) of affective states, on the one hand, and approaches which provide predictions based on some of the many more elaborated representational systems for affective states, on the other hand.

Originally, research activities focused on polarity alone. In the meantime, a shift towards more expressive representation models for emotion can be observed that heavily draws inspirations from psychological theory, e.g., Basic Emotions (Ekman, 1992) or the Valence-Arousal-Dominance model (Bradley and Lang, 1994).

Though this change turned out to be really beneficial for sentiment analysis in NLP, a large variety of mutually incompatible encodings schemes for emotion and, consequently, annotation formats for emotion metadata in corpora have emerged that hinder the interoperability of these resources and their subsequent reuse, e.g., on the basis of alignments or mergers (Buechel and Hahn, 2017).

As an alternative way of dealing with thus unwarranted heterogeneity, we here examine the potential of multi-task learning (MTL; Caruana (1997)) for word-level emotion prediction. In MTL for neural networks, a single model is fitted to solve multiple, independent tasks (in our case, to predict different emotional dimensions) which typically results in learning more robust and meaningful intermediate representations. MTL has been shown to greatly decrease the risk of overfitting (Baxter, 1997), work well for various NLP tasks (Setiawan et al., 2015; Liu et al., 2015; Søgaard and Goldberg, 2016; Cummins et al., 2016; Liu et al., 2017; Peng et al., 2017), and practically increases sample size, thus making it a natural choice for small-sized data sets typically found in the area of word emotion induction.

Proceedings of NAACL-HLT 2018, pages 1907–1918
New Orleans, Louisiana, June 1 - 6, 2018.

After a discussion of related work in Section 2, we will introduce several reference methods and describe our proposed deep MTL model in Section 3. In our experiments (Section 4), we will first validate our claim that MTL is superior to single-task learning for word emotion induction. After that, we will provide a large-scale evaluation of our model featuring 9 typologically diverse languages and multiple publicly available embedding models for a total of 15 conditions. Our MTL model surpasses the current state-of-the-art for each of them, and even performs competitive relative to human reliability. Most notably however, our approach yields the largest benefit on the smallest data sets, comprising merely one thousand samples. This finding, counterintuitive as it may be, strongly suggests that MTL is particularly beneficial for solving the word emotion induction problem. Our code base as well as the resulting experimental data is freely available.[1]

2 Related Work

This section introduces the emotion representation format underlying our study and describes external resources we will use for evaluation before we discuss previous methodological work.

Emotion Representation and Data Sets. Psychological models of emotion can typically be subdivided into *discrete* (or *categorical*) and *dimensional* ones (Stevenson et al., 2007; Calvo and Mac Kim, 2013). Discrete models are centered around particular sets of emotional categories considered to be fundamental. Ekman (1992), for instance, identifies six *Basic Emotions* (Joy, Anger, Sadness, Fear, Disgust and Surprise).

In contrast, dimensional models consider emotions to be composed of several influencing factors (mainly two or three). These are often referred to as *Valence* (a positive–negative scale), *Arousal* (a calm–excited scale), and *Dominance* (perceived degree of control over a (social) situation)—the VAD model (Bradley and Lang (1994); see Figure 1 for an illustration). Many contributions though omit Dominance (the VA model) (Russell, 1980). For convenience, we will still use the term "VAD" to jointly refer to both variants (with and without Dominance).

VAD is the most common framework to acquire empirical emotion values for words in psychology.

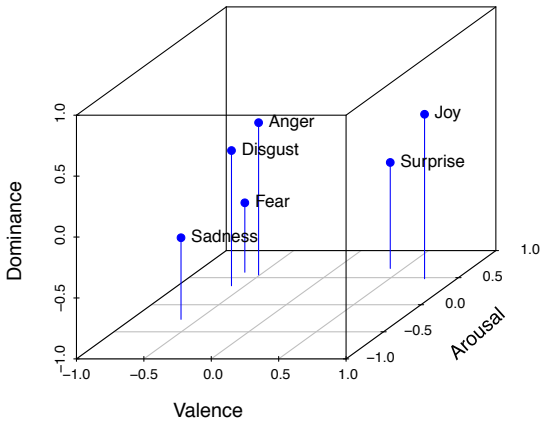

Figure 1: Affective space spanned by the Valence-Arousal-Dominance (VAD) model, together with the position of six Basic Emotions; as determined by Russell and Mehrabian (1977).

Over the years, a considerable number of such resources (also called "emotion lexicons") have emerged from psychological research labs (as well as some NLP labs) for diverse languages. The emotion lexicons we use in our experiments are listed in Table 1. An even more extensive list of such data sets is presented by Buechel and Hahn (2018). For illustration, we also provide three sample entries from one of those lexicons in Table 2. As can be seen, the three affective dimensions behave complementary to each other, e.g., "terrorism" and "orgasm" display similar Arousal but opposing Valence.

The task we address in this paper is to predict the values for Valence, Arousal and Dominance, given a lexical item. As is obvious from these examples, we consider emotion prediction as a regression, not as a classification problem (see arguments discussed in Buechel and Hahn (2016)).

In this paper, we focus on the VAD format for the following reasons: First, note that the Valence dimension exactly corresponds to polarity (Turney and Littman, 2003). Hence, with the VAD model, emotion prediction can be seen as a generalization over classical polarity prediction. Second, to the best of our knowledge, the amount and diversity of available emotion lexicons with VAD encodings is larger than for any other format (see Table 1).

Word Embeddings. Word embeddings are dense, low-dimensional vector representations of words trained on large volumes of raw text in an unsupervised manner. The following are among today's most popular embedding algorithms:

[1] https://github.com/JULIELab/wordEmotions

Source	ID	Language	Format	# Entries
Bradley and Lang (1999)	EN	English	VAD	1,034
Warriner et al. (2013)	EN+	English	VAD	13,915
Redondo et al. (2007)	ES	Spanish	VAD	1,034
Stadthagen-Gonzalez et al. (2017)	ES+	Spanish	VA	14,031
Schmidtke et al. (2014)	DE	German	VAD	1,003
Yu et al. (2016a)	ZH	Chinese	VA	2,802
Imbir (2016)	PL	Polish	VAD	4,905
Montefinese et al. (2014)	IT	Italian	VAD	1,121
Soares et al. (2012)	PT	Portuguese	VAD	1,034
Moors et al. (2013)	NL	Dutch	VAD	4,299
Sianipar et al. (2016)	ID	Indonesian	VAD	1,490

Table 1: Emotion lexicons used in our experiments (with their bibliographic source, identifier, language they refer to, emotion representation format, and number of lexical entries they contain).

Word	Valence	Arousal	Dominance
sunshine	8.1	5.3	5.4
terrorism	1.6	7.4	2.7
orgasm	8.0	7.2	5.8

Table 2: Three sample entries from Warriner et al. (2013). They use 9-point scales ranging from 1 (most negative/calm/submissive) to 9 (most positive/excited/dominant).

WORD2VEC (with its variants SGNS and CBOW) features an extremely trimmed down neural network (Mikolov et al., 2013). FASTTEXT is a derivative of WORD2VEC, also incorporating sub-word character n-grams (Bojanowski et al., 2017). Unlike the former two algorithms which fit word embeddings in a streaming fashion, GLOVE trains word vectors directly on a word co-occurrence matrix under the assumption to make more efficient use of word statistics (Pennington et al., 2014). Somewhat similar, SVD_{PPMI} performs singular value decomposition on top of a point-wise mutual information co-occurrence matrix (Levy et al., 2015).

In order to increase the reproducibility of our experiments, we rely on the following widely used, publicly available embedding models trained on very large corpora (summarized in Table 3): the SGNS model trained on the Google News corpus[2] (GOOGLE), the FASTTEXT model trained on Common Crawl[3] (COMMON), as well as the FASTTEXT models for a wide range of languages trained on the respective Wikipedias[4] (WIKI). Note that WIKI denotes multiple embedding models with different training and vocabulary sizes (see Grave et al. (2018) for further details). Additionally, we were given the opportunity to reuse the English embedding model from Sedoc et al. (2017) (GIGA), a strongly related contribution (see below). Their embeddings were trained on the English Gigaword corpus (Parker et al., 2011).

Word-Level Prediction. One of the early approaches to word *polarity* induction which is still popular today (Köper and Schulte im Walde, 2016) was introduced by Turney and Littman (2003). They compute the polarity of an unseen word based on its point-wise mutual information (PMI) to a set of positive and negative seed words, respectively.

SemEval-2015 Task 10E featured polarity induction on Twitter (Rosenthal et al., 2015). The best system relied on support vector regression (SVR) using a radial base function kernel (Amir et al., 2015). They employ the embedding vector of the target word as features. The results of their SVR-based system were beaten by the DENSIFIER algorithm (Rothe et al., 2016). DENSIFIER learns an orthogonal transformation of an embedding space into a subspace of strongly reduced dimensionality.

Hamilton et al. (2016) developed SENTPROP, a graph-based, semi-supervised learning algorithm which builds up a word graph, where vertices correspond to words (of known as well as unknown polarity) and edge weights correspond to the similarity between them. The polarity information is then propagated through the graph, thus computing scores for unlabeled nodes. According to their evaluation, DENSIFIER seems to be superior overall, yet SENTPROP produces competitive results

[2] https://code.google.com/archive/p/word2vec/

[3] https://fasttext.cc/docs/en/english-vectors.html

[4] https://github.com/facebookresearch/fastText/blob/master/pretrained-vectors.md

ID	Language	Method	Corpus	# Tokens	# Types	# Dimensions
GOOGLE	English	SGNS	Google News	1×10^{11}	3×10^{6}	300
COMMON	English	FASTTEXT	Common Crawl	6×10^{11}	2×10^{6}	300
GIGA	English	CBOW	Gigawords	4×10^{9}	2×10^{6}	300
WIKI	all	FASTTEXT	Wikipeda	—	—	300

Table 3: Embedding models used for our experiments with identifier, language, embedding algorithm, training corpus, its size in the number of tokens, size of the vocabulary (types) of the resulting embedding model and its dimensionality.

only when the seed lexicon or the corpus the word embeddings are trained on is very small.[5]

For word *emotion* induction, a very similar approach to SENTPROP has been proposed by Wang et al. (2016a). They also propagate affective information (Valence and Arousal, in this case) through a word graph with similarity weighted edges.

Sedoc et al. (2017) recently proposed an approach based on signed spectral clustering where a word graph is constructed not only based on word similarity but also on the considered affective information (again, Valence and Arousal). The emotion value of a target word is then computed based on the seed words in its cluster. They report to outperform the results from Wang et al. (2016a).

Contrary to the trend to graph-based methods, the best system of the IALP 2016 Shared Task on Chinese word emotion induction (Yu et al., 2016b) employed a simple feed-forward neural network (FFNN) with one hidden layer in combination with boosting (Du and Zhang, 2016).

Another very recent contribution which advocates a supervised set-up was published by Li et al. (2017). They propose ridge regression, again using word embeddings as features. Even with this simple approach, they report to outperform many of the above methods in the VAD prediction task.[6]

Sentence-Level and Text-Level Prediction. Different from the word-level prediction task (the one we focus on in this contribution), the determination of emotion values for higher-level linguistic units (especially sentences and texts) is also heavily investigated. For this problem, DL approaches are meanwhile fully established as the method of choice (Wang et al., 2016b; Abdul-Mageed and Ungar, 2017; Felbo et al., 2017; Mohammad and Bravo-Marquez, 2017).

[5]Personal correspondence with William L. Hamilton; See also README at https://github.com/williamleif/socialsent

[6]However, they also report extremely weak performance figures for some of their reference methods.

It is important to note, however, that the methods discussed for these higher-level units cannot easily be transferred to solve the word emotion induction problem. Sentence-level and text-level architectures are either adapted to *sequential* input data (typical for RNN, LSTM, GRNN and related architectures) or *spatially arranged* input data (as with CNN architectures). However, for word embeddings (the default input for word emotion induction) there does not seem to be any meaningful order of their components. Therefore, these more sophisticated DL methods are, for the time being, not applicable for the study at hand.

3 Methods

In this section, we will first introduce various reference methods (two originally polarity-based for which we offer adaptations for VAD prediction) before defining our own neural MTL model and discussing its difference from previous work.

Let $V := \{w_1, w_2, ..., w_m\}$ be our word vocabulary and let $E := \{e_1, e_2, ..., e_m\}$ be a set of embedding vectors such that $e_i \in \mathbb{R}^n$ denotes the n-dimensional vector representation of word w_i. Let $D := \{d_1, d_2, ..., d_l\}$ be a set of emotional dimensions. Our task is to predict the empirically determined emotion vector $emo(w) \in \mathbb{R}^l$ given a word w and the embedding space E.

3.1 Reference Methods

Linear Regression Baseline (LinReg). We propose (multi-variate) linear regression as an obvious baseline for the problem:

$$emo_{LR}(w_k) := We_k + b \quad (1)$$

where W is a matrix, W_{i*} contains the regression coefficients for the i-th affective dimension and b is the vector of bias terms. The model parameters are fitted using ordinary least squares. Technically, we use the `scikit-learn.org` implementation with default parameters.

Ridge Regression (RidgReg). Li et al. (2017) propose ridge regression for word emotion induction. Ridge regression works identically to linear regression during prediction, but introduces L_2 regularization during training. Following the authors, for our implementation, we again use the `scikit-learn` implementation with default parameters.

Turney-Littman Algorithm (TL). As one of the earliest contributions in the field, Turney and Littman (2003) defined a simple PMI-based approach to determine the semantic polarity SP_{TL} of a word w:

$$SP_{TL}(w) := \sum_{s \in seeds^+} pmi(w,s) - \sum_{s \in seeds^-} pmi(w,s) \quad (2)$$

where $seeds^+$ and $seeds^-$ are sets of positive and negative seed words, respectively. Since this algorithm is still popular today (Köper and Schulte im Walde, 2016), we here provide a novel modification for adapting this originally polarity-based approach to word emotion induction with vectorial seed and output values.

First, we replace PMI-based association of seed and target word w and s by their similarity sim based on their word embeddings e_w and e_s:

$$sim(w,s) := max(0, \frac{e_w \cdot e_s}{||e_w|| \times ||e_s||}) \quad (3)$$

$$emo(w) := \sum_{s \in seeds^+} sim(w,s) - \sum_{s \in seeds^-} sim(w,s) \quad (4)$$

Although this step is technically not required for the adaptation, it renders the TL algorithm more comparable to the other approaches evaluated in Section 4 besides from most likely increasing performance. Equation (4) can be rewritten as

$$emo(w) := \sum_{s \in seeds} sim(w,s) \times emo(s) \quad (5)$$

where $seeds := seeds^+ \cup seeds^-$ and $emo(s)$ maps to 1, if $s \in seeds^+$, and -1, if $s \in seeds^-$.

Equation (5) can be trivially adapted to an n-dimensional emotion format by redefining $emo(s)$ such that it maps to a vector from $\mathbb{R}^n$ instead of $\{-1, 1\}$. Our last step is to introduce a normalization term such that $emo(w)_{TL}$ lies within the range of the seed lexicon.

$$emo_{TL}(w) := \frac{\sum_{s \in seeds} sim(w,s) \times emo(s)}{\sum_{s \in seeds} sim(w,s)} \quad (6)$$

As can be seen from Equation (6), for the more general case of n-dimensional emotion prediction, the Turney-Littman algorithm naturally translates into a weighted average where the seed emotion values are weighted according to the similarity to the target item.

Densifier. Rothe et al. (2016) train an orthogonal matrix $Q \in \mathbb{R}^{n \times n}$ (n being the dimensionality of the word embeddings) such that applying Q to an embedding vector e_i concentrates all the polarity information in its first dimension such that the polarity of a word w_i can be computed as

$$SP_{\text{DENSIFIER}}(w_i) := pQe_i \quad (7)$$

where $p = (1, 0, 0, ..., 0)^T \in \mathbb{R}^{1 \times n}$.

For fitting Q, the seeds are arranged into pairs of equal polarity (the set $pairs^=$) and those of opposing polarity ($pairs^\neq$). A good fit for Q will minimize the distance within the former and maximize the distance within the latter which can be expressed by the following two training objectives:

$$\underset{Q}{\text{argmin}} \sum_{(w_i,w_j) \in pairs^=} |pQ(e_i - e_j)| \quad (8)$$

$$\underset{Q}{\text{argmax}} \sum_{(w_i,w_j) \in pairs^\neq} |pQ(e_i - e_j)| \quad (9)$$

The objectives described in the expressions (8) and (9) are combined into a single loss function (using a weighting factor $\alpha \in [0, 1]$) which is then minimized using stochastic gradient descent (SGD).

To adapt this algorithm to dimensional emotion formats, we construct a positive seed set, $seeds_v^+$, and a negative seed set, $seeds_v^-$, for each emotion dimension $v \in D$. Let M_v be the mean value of all the entries of the training lexicon for the affective dimension v. Let SD_v be the respective standard deviation and $\beta \in \mathbb{R}$, $\beta \geq 0$. Then all entries greater than $M_v + \beta SD_v$ are assigned to $seeds_v^+$ and those less than $M_v - \beta SD_v$ are assigned to $seeds_v^-$. Q is fitted individually for each emotion dimension v.

Training was performed according to the original paper with the exception that (following Hamilton et al. (2016)) we did not apply the proposed re-orthogonalization after each training

step, since we did not find any evidence that this procedure actually results in improved performance. The hyperparameters α and β were set to .7 and .5 (respectively) for all experiments based on a pilot study. Since the original implementation is not accessible, we devised our own using `tensorflow.org`.

Boosted Neural Networks (ensembleNN). Du and Zhang (2016) propose simple FFNNs in combination with a boosting algorithm. An FFNN consists of an *input* or *embedding layer* with activation $a^{(0)} \in \mathbb{R}^n$ which is equal to the embedding vector e_k when predicting the emotion of a word w_k. The input layer is followed by multiple hidden layers with activation

$$a^{(l+1)} := \sigma(W^{(l+1)}a^{(l)} + b^{(l+1)}) \qquad (10)$$

where $W^{(l+1)}$ and $b^{(l+1)}$ are the weights and biases for layer $l+1$ and σ is a nonlinear activation function. Since we treat emotion prediction as a regression problem, the activation on the output layer a^{out} (where out is the number of non-input layers in the network) is computed as the affine transformation

$$a^{(out)} := W^{(out)}a^{(out-1)} + b^{(out)} \qquad (11)$$

Boosting is a general machine learning technique where several weak estimators are combined to form a strong estimator. The authors used FFNNs with a single hidden layer of 100 units and rectified linear unit (ReLU) activation. The boosting algorithm AdaBoost.R2 (Drucker, 1997) was used to train the ensemble (one per affective dimension). Our re-implementation copies their technical set-up[7] exactly using `scikit-learn`.

3.2 Multi-Task Learning Neural Network

The approaches introduced in Section 3.1 and Section 2 vary largely in their methodological foundations, i.e., they comprise semi-supervised and supervised machine learning techniques—both statistical and neural ones. Yet, they all have in common that they treat the prediction of the different emotional dimensions *as separate* tasks. That is, they fit one individual model per VAD dimension without sharing parameters between them.

In contradistinction, the key feature of our approach is that we fit a single FFNN model to predict *all VAD dimensions jointly*, thus applying multi-task learning to word emotion induction. Hence, we treat the prediction of Valence, Arousal and Dominance as three independent tasks. Our multi-task learning neural network (MTLNN) (depicted in Figure 2) has an output layer of three units such that each output unit represents one of the VAD dimensions. However, the activation in our two hidden layers (of 256 and 128 units, respectively) is *shared* across all VAD dimensions, and so are the associated weights and biases.

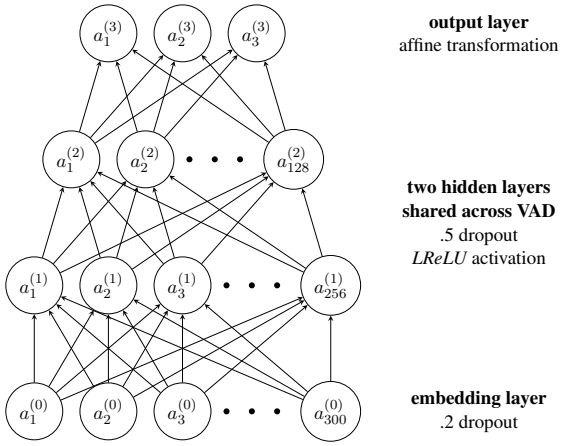

Figure 2: MTL architecture for VAD prediction.

Thus, while we train our MTLNN model it is forced to learn intermediate representations of the input which are generally informative for all VAD dimensions. This serves as a form of regularization, since it becomes less likely for our model to fit the noise in the training set as noise patterns may vary across emotional dimensions. Simultaneously, this has an effect similar to an increase of the training size, since each sample now leads to additional error signals during backpropagation. Intuitively, both properties seem extremely useful for relatively small-sized emotion lexicons (see Section 4 for empirical evidence).

The remaining specifications of our model are as follows. We use *leaky* ReLU activation (LReLU) as nonlinearity (Maas et al., 2013).

$$LReLU(z_i) := max(\gamma z_i, z_i) \qquad (12)$$

with $\gamma := .01$ for our experiments. For regularization, dropout (Srivastava et al., 2014) is applied during training with a probability of .2 on the embedding layer and .5 on the hidden layers. We train for $15,000$ iterations (well beyond convergence on each data set we use) with the ADAM optimizer (Kingma and Ba, 2015) of .001 base learning rate,

[7] Original settings available at `https://github.com/StevenLOL/ialp2016_Shared_Task`

batch size of 128 and Mean-Squared-Error loss. The weights are randomly initialized (drawn from a normal distribution with a standard deviation .001) and biases are uniformly initialized as .01. `Tensorflow` is used for implementation.

4 Results

In this section, we first validate our assumption that MTL is superior to single-task learning for word emotion induction. Next, we compare our proposed MTLNN model in a large-scale evaluation experiment.

Performance figures will be measured as Pearson correlation (r) between our automatically predicted values and human gold ratings. The Pearson correlation between two data series $X = x_1, x_2, ..., x_n$ and $Y = y_1, y_2, ..., y_n$ takes values between $+1$ (perfect positive correlation) and -1 (perfect negative correlation) and is computed as

$$r_{xy} := \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (13)$$

where $\bar{x}$ and $\bar{y}$ denote the mean values for X and Y, respectively.

4.1 Single-Task vs. Multi-Task Learning

The main hypothesis of this contribution is that an MTL set-up is superior to single-task learning for word emotion induction. Before proceeding to the large-scale evaluation of our proposed model, we will first examine this aspect of our work.

For this, we use the following experimental set-up: We will compare the MTLNN model against its single-task learning counterpart (SepNN). SepNN simultaneously trains three separate neural networks where only the input layer, yet no parameters of the intermediate layers are shared across the models. Each of the separate networks is identical to MTLNN (same layers, dropout, initialization, etc.), yet has only one output neuron, thus modeling only one of the three affective VAD dimensions. SepNN is equivalent to fitting our proposed model (but with only one output unit) to the different VAD dimensions individually, one after the other. Yet, training these separate networks simultaneously (not jointly!) makes both approaches, MTLNN and SepNN, easier to compare.

We will run MTLNN against SepNN on the EN and the EN+ data set (the former is very small, the latter relatively large; see Table 1) using the following set-up: for each gold lexicon and model, we randomly split the data $9/1$ and train for $15,000$ iterations on the larger split (the same number of steps is used for the main experiment). After each one-thousand iterations step, model performance is tested on the held-out data. This process will be repeated 20 times and the performance figures at each one-thousand iterations step will be averaged. In a final step, we will average the results for each of the three emotional dimensions and only plot this average value. The results of this experiment are depicted in Figure 3.

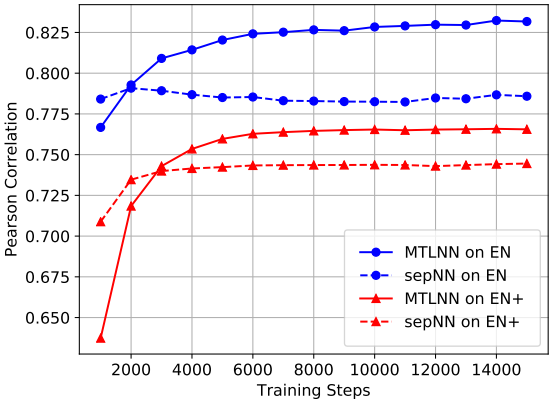

Figure 3: Performance of our proposed MTLNN model vs. its single-task learning counterpart SepNN against training steps.

First of all, each combination of model and data set displays a satisfactory performance of at least $r \approx .75$ after 15,000 steps compared to previous work (see below). Overall, performance is higher for the smaller EN lexicon. Although counterintuitive (since smaller lexicons lead to fewer training samples), this finding is consistent with prior work (Sedoc et al., 2017; Li et al., 2017) and is probably related to the fact that smaller lexicons usually comprise a larger portion of strongly emotion-bearing words. In contrast, larger lexicons add more neutral words which tend to be harder to predict in terms of correlation.

As hypothesized, the MTLNN model does indeed outperform the single task model on both data sets. Our data also suggest that the gain from the MTL approach is larger on smaller data sets (again in concordance with our expectations). Figure 3 reveals that this might be due to the regularizing effect of MTL, since the SepNN model shows signs of overfitting on the EN data set. Yet, even

Language	Data	Embeddings	LinReg	RidgReg	TL	Densifier	ensembleNN	MTLNN
English	EN+	GOOGLE	0.696	0.696	0.631	0.622	0.728	**0.739*****
English	EN+	COMMON	0.719	0.719	0.659	0.652	0.762	**0.767*****
English	EN+	WIKI	0.666	0.666	0.591	0.584	0.706	**0.712*****
English	EN	GOOGLE	0.717	0.732	0.723	0.712	0.688	**0.810*****
English	EN	COMMON	0.731	0.741	0.741	0.726	0.717	**0.824*****
English	EN	WIKI	0.656	0.667	0.674	0.665	0.681	**0.777*****
Spanish	ES	WIKI	0.698	0.709	0.704	0.690	0.700	**0.804*****
Spanish	ES+	WIKI	0.693	0.694	0.603	0.598	0.766	**0.778*****
German	DE	WIKI	0.709	0.719	0.714	0.710	0.700	**0.801*****
Chinese	ZH	WIKI	0.716	0.717	0.586	0.599	0.737	**0.744****
Polish	PL	WIKI	0.650	0.650	0.577	0.553	0.687	**0.712*****
Italian	IT	WIKI	0.656	0.665	0.672	0.659	0.630	**0.751*****
Portuguese	PT	WIKI	0.673	0.684	0.685	0.678	0.672	**0.768*****
Dutch	NL	WIKI	0.651	0.652	0.559	0.532	0.704	**0.730*****
Indonesian	ID	WIKI	0.581	0.586	0.581	0.576	0.575	**0.660*****
Average			0.638	0.659	0.611	0.605	0.676	**0.728*****

Table 4: Results of our main experiment in averaged Pearson correlation; best result per condition (in rows) in bold, second best result underlined; significant difference (paired two-tailed t-test) over the second best system marked with "*", "**", or "***" for $p < .05$, .01, or .001, respectively.

when the separate model does not overfit (as on the EN+ lexicon), MTLNN reveals better results.

Although SepNN needs fewer *training steps* before convergence, the MTLNN model trains much faster, thus still converging faster in terms of *runtime* (about a minute on a middle-class GPU). This is because MTLNN has only about a third as many parameters as the separate model SepNN.

4.2 Comparison against Reference Methods

We combined each of the selected lexicon data sets (Table 1) with each of the applicable publicly available embedding models (Section 2; the embedding model provided by Sedoc et al. (2017) will be used separately) for a total of 15 conditions, i.e, the rows in Table 4.

For each of these conditions, we performed a 10-fold cross-validation (CV) for each of the 6 methods presented in Section 3 such that each method is presented with the identical data splits.[8] For each condition, algorithm, and VA(D) dimension, we compute the Pearson correlation r between gold ratings and predictions. For conciseness, we present only the average correlation over the respective affective dimensions in Table 4 (Valence and Arousal for ES+ and ZH, VAD for the others). Note that the methods we compare ourselves against comprise the current state-of-the art in both polarity and emotion induction (as described in Section 2).

[8]This procedure constitutes a more direct comparison than using different splits for each method and allows using *paired* t-tests.

As can be seen, our proposed MTLNN model outperforms all other approaches in each of the 15 conditions. Regarding the average over all affective dimensions and conditions, it outperforms the second best system, ensembleNN, by more than 5%-points. In line with our results from Section 4.1, those improvements are especially pronounced on smaller data sets containing one up to two thousand entries (EN, ES, IT, PT, ID) with close to 10%-points improvement over the respective second-best system.

Concerning the relative ordering of the affective dimensions, in line with former studies (Sedoc et al., 2017; Li et al., 2017), the performance figures for the Valence dimension are usually much higher than for Arousal and Dominance. Using MTLNN, for many conditions, we see the pattern that Valence is about 10%-points above the VAD average, Arousal being 10%-points below and Dominance being roughly equal to the average over VAD (this applies, e.g., to EN, EN+ and IT). On other data sets (e.g., PL, NL and ID), the ordering between Arousal and Dominance is less clear though Valence still stands out with the best results. We observe the same general pattern for the reference methods, as well.

Concerning the comparison to Sedoc et al. (2017), arguably one of most related contributions, they report a performance of $r = .768$ for Valence and .582 for Arousal on the EN+ data set in a 10-fold CV using their own embeddings. In contrast, MTLNN using the COMMON model achieves $r = .870$ and .674 in the same set-up—about 10%-

	Valence	Arousal	Dominance
MTLNN EN	.918	.730	.825
MTLNN EN+	.870	.674	.758
ISR EN $\sim$ EN+	.953	.759	.795
SHR EN+	.914	.689	.770

Table 5: Comparison of the MTLNN model against inter-study reliability (ISR) between the EN and the EN+ data set and split-half reliability (SHR) of the EN+ data set (in Pearson correlation).

points better on both dimensions. However, the COMMON model was trained on much more data than the embeddings Sedoc et al. (2017) use. For the most direct comparison, we also repeated this experiment using *their* embedding model (GIGA). We find that MTLNN still clearly outperforms their results with $r = .814$ for Valence and .607 for Arousal.[9]

MTLNN achieves also very strong results in direct comparison to human performance (see Table 5). Warriner et al. (2013) (who created EN+) report an inter-study reliability (ISR; i.e., the correlation of the aggregated ratings from two different studies) between the EN and the EN+ lexicon of $r = .953$, .759 and .795 for VAD, respectively. Since EN is a subset of EN+, we can compare these performance figures against our own results on the EN data set where we achieved $r = .918$, .730 and .825, respectively. Thus, our proposed method did actually outperform human reliability for Dominance and is competitive for Valence and Arousal, as well.

This general observation is also backed up by split-half reliability data (SHR; i.e., when randomly splitting all individual ratings in two groups and averaging the ratings within each group, how strong is the correlation between these averaged ratings?). For the EN+ data set, Warriner et al. (2013) report an SHR of $r = .914$, .689 and .770 for VAD, respectively. Again, our MTLNN model performs very competitive with $r = .870$, .674 and .758, respectively using the COMMON embeddings.

5 Conclusion

In this paper, we propose multi-task learning (MTL) as a simple, yet surprisingly efficient method to improve the performance and, at the same time, to deal with existing data limitations in word emotion induction—the task to predict a complex emotion score for an individual word. We validated our claim that MTL is superior to single-task learning by achieving better results with our proposed method in performance as well as training time compared to its single-task counterpart. We performed an extensive evaluation of our model on 9 typologically diverse languages, using different kinds of word embedding models for a total 15 conditions. Comparing our approach to state-of-the-art methods from word polarity and word emotion induction, our model turns out to be superior in each condition, thus setting a novel state-of-the-art performance for both polarity *and* emotion induction. Moreover, our results are even competitive to human annotation reliability in terms of inter-study as well as split-half reliability. Since this contribution was restricted to the VAD format of emotion representation, in future work we will examine whether MTL yields similar gains for other representational schemes, as well.

[9]We also clearly outperform their results for the NL and ES+ data sets. For these cases, our embedding models were similar in training size.

Acknowledgments

We would like to thank the Positive Psychology Center, University of Pennsylvania for providing us with the embedding model used in Sedoc et al. (2017), Johannes Hellrich, JULIE Lab, for insightful discussions, and the reviewers for their valuable comments.

References

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*. Vancouver, British Columbia, Canada, July 30 - August 4, 2017, volume 1: Long Papers, pages 718–728.

Silvio Amir, Ramón F. Astudillo, Wang Ling, Bruno Martins, Mário J. Silva, and Isabel Trancoso. 2015. INESC-ID: A regression model for large scale Twitter sentiment lexicon induction. In *SemEval 2015 — Proceedings of the 9th International Workshop on Semantic Evaluation @ NAACL-HLT 2015*. Denver, Colorado, USA, June 4-5, 2015, pages 613–618.

Jonathan Baxter. 1997. A Bayesian/information theoretic model of learning to learn via multiple task sampling. *Machine Learning* 28(1):7–39.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomáš Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* 5(1):135–146.

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, FL.

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. Including Prestigious Applications of Artificial Intelligence (PAIS 2016)*. The Hague, Netherlands, August 29 - September 2, 2016, volume 285 of *Frontiers in Artificial Intelligence and Applications*, pages 1114–1122.

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*. Valencia, Spain, April 3–7, 2017, volume 2: Short Papers, pages 578–585.

Sven Buechel and Udo Hahn. 2018. Representation mapping: A novel approach to generate high-quality multi-lingual emotion lexicons. In *LREC 2018 — Proceedings of the 11th International Conference on Language Resources and Evaluation*. Miyazaki, Japan, May 7–12, 2018.

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

Rich Caruana. 1997. Multitask learning. *Machine Learning* 28(1):41–75.

Ronan Cummins, Meng Zhang, and Ted Briscoe. 2016. Constrained multi-task learning for automated essay scoring. In *ACL 2016 — Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Berlin, Germany, August 7-12, 2016, volume 2: Short Papers, pages 789–799.

Harris Drucker. 1997. Improving regressors using boosting techniques. In *ICML '97 — Proceedings of the 14th International Conference on Machine Learning*. Nashville, Tennessee, USA, July 8-12, 1997, pages 107–115.

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 [20th] 2016 International Conference on Asian Language Processing*. Tainan, Taiwan, November 21-23, 2016, pages 161–163.

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

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *EMNLP 2017 — Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark, September 9-11, 2017, pages 1615–1625.

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

William L. Hamilton, Kevin Clark, Jure Leskovec, and Daniel 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*. Austin, Texas, USA, November 1-5, 2016, pages 595–605.

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*. Madrid, Spain, July 7-12, 1997, pages 174–181.

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.

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.

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*. Portorož, Slovenia, 23-28 May 2016, pages 2595–2598.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics* 3:211–225.

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.

PengFei Liu, Xipeng Qiu, and Xuanjing Huang. 2017. Adversarial multi-task learning for text classification. In *ACL 2017 — Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. Vancouver, British Columbia, Canada, July 30 - August 4, 2017, volume 1: Long Papers, pages 1–10.

Xiaodong Liu, Jianfeng Gao, Xiaodong He, Li Deng, Kevin Duh, and Ye-Yi Wang. 2015. Representation learning using multi-task deep neural networks for semantic classification and information retrieval. In *NAACL-HLT 2015 — Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Denver, Colorado, USA, May 31 - June 5, 2015, pages 912–921.

Andrew L. Maas, Awni Y. Hannun, and Andrew Y. Ng. 2013. Rectifier nonlinearities improve neural network acoustic models. In *Proceedings of the Workshop on Deep Learning for Audio, Speech and Language Processing @ ICML 2013*. Atlanta, Georgia, USA, 16 June 2013.

Tomáš Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26 — NIPS 2013. Proceedings of the 27th Annual Conference on Neural Information Processing Systems*. Lake Tahoe, Nevada, USA, December 5-10, 2013, pages 3111–3119.

Saif M. 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*. Copenhagen, Denmark, September 8, 2017, pages 34–49.

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.

Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2011. English Gigaword Fifth Edition. Technical Report LDC2011T07, Linguistic Data Consortium, Philadelphia, Pennsylvania, USA. https://catalog.ldc.upenn.edu/LDC2011T07.

Hao Peng, Sam Thomson, and Noah A. Smith. 2017. Deep multitask learning for semantic dependency parsing. In *ACL 2017 — Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. Vancouver, British Columbia, Canada, July 30 - August 4, 2017, volume 1: Long Papers, pages 2037–2048.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GLOVE: Global vectors for word representation. In *EMNLP 2014 — Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. Doha, Qatar, October 25-29, 2014, pages 1532–1543.

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.

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*. Denver, Colorado, USA, June 4-5, 2015, pages 451–463.

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

James A. Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology* 39(6):1161–1178.

James A. Russell and Albert Mehrabian. 1977. Evidence for a three-factor theory of emotions. *Journal of Research in Personality* 11(3):273–294.

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*. Valencia, Spain, April 3-7, 2017, volume 2: Short Papers, pages 564–571.

Hendra Setiawan, Zhongqiang Huang, Jacob Devlin, Thomas Lamar, Rabih Zbib, Richard M. Schwartz, and John Makhoul. 2015. Statistical machine translation features with multitask tensor networks. In *ACL-IJCNLP 2015 — Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics & 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*. Beijing, China, July 26-31, 2015, volume 1: Long Papers, pages 31–41.

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.

Anders Søgaard and Yoav Goldberg. 2016. Deep multi-task learning with low level tasks supervised at lower layers. In *ACL 2016 — Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Berlin, Germany, August 7–12, 2016, volume 2: Short Papers, pages 231–235.

Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan R. Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* 15:1929–1958.

Hans Stadthagen-Gonzalez, 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.

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.

Jin Wang, Liang-Chih Yu, K. Robert Lai, and Xuejie Zhang. 2016a. 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.

Jin Wang, Liang-Chih Yu, K. Robert Lai, and Xuejie Zhang. 2016b. Dimensional sentiment analysis using a regional CNN-LSTM model. In *ACL 2016 — Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Berlin, Germany, August 7-12, 2016, volume 2: Short Papers, pages 225–230.

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.

Liang-Chih Yu, Lung-Hao Lee, Shuai Hao, Jin Wang, Yunchao He, Jun Hu, K. Robert Lai, and Xuejie Zhang. 2016a. 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*. San Diego, California, USA, June 12-17, 2016, pages 540–545.

Liang-Chih Yu, Lung-Hao Lee, and Kam-Fai Wong. 2016b. Overview of the IALP 2016 Shared Task on dimensional sentiment analysis for Chinese words. In *IALP 2016 — Proceedings of the [20th] 2016 International Conference on Asian Language Processing*. Tainan, Taiwan, November 21-23, 2016, pages 156–160.