Joint Emotion Label Space Modelling for Affect Lexica

Luna De Bruyne[a,*], Pepa Atanasova[b], Isabelle Augenstein[b]

[a] *Ghent University, Ghent, Belgium*
[b] *University of Copenhagen, Copenhagen, Denmark*

Abstract

Emotion lexica are commonly used resources to combat data poverty in automatic emotion detection. However, vocabulary coverage issues, differences in construction method and discrepancies in emotion framework and representation result in a heterogeneous landscape of emotion detection resources, calling for a unified approach to utilising them. To combat this, we present an extended emotion lexicon of 30,273 unique entries, which is a result of merging eight existing emotion lexica by means of a multi-view variational autoencoder (VAE). We showed that a VAE is a valid approach for combining lexica with different label spaces into a joint emotion label space with a chosen number of dimensions, and that these dimensions are still interpretable. We tested the utility of the unified VAE lexicon by employing the lexicon values as features in an emotion detection model. We found that the VAE lexicon outperformed individual lexica, but contrary to our expectations, it did not outperform a naive concatenation of lexica, although it did contribute to the naive concatenation when added as an extra lexicon. Furthermore, using lexicon information as additional features on top of state-of-the-art language models usually resulted in a better performance than when no lexicon information was used.

Keywords: NLP, Emotion detection, Emotion lexica, VAE

*Corresponding author.
Email address: luna.debruyne@ugent.be (Luna De Bruyne)

1. Introduction

Affect lexica are valuable resources in the fields of experimental psychology and natural language processing (NLP). An affect lexicon is a list of words or a database that contains lexical entries with their associated affective value. This value can be conveyed as a polarity association (negative - neutral - positive), in which case we call it a sentiment lexicon, or, following emotion frameworks provided in the field of psychology, it can be denoted as a value associated with an emotion category or emotional dimension. In this case, we use the term emotion lexicon.

In the field of psychology, affect lexica are mostly known as affective norms. These norms can be used as stimuli in emotion research or for designing experiments on word memory and processing (Warriner et al., 2013). Also in NLP, emotion lexica are in high demand, because they can be used to combat data poverty in automatic emotion detection. The lexica can be employed as a straight-forward way to automatically label texts with emotional information, or they can be used as features in supervised machine learning approaches (Ma et al., 2018). Even in state-of-the-art systems for emotion detection (e.g., the winning teams of the SemEval-2018 shared task on multi-label emotion classification), word embeddings in Bi-LSTM architectures are complemented with features from affect lexica (Baziotis et al., 2018; Meisheri & Dey, 2018).

However, methodological issues emerge when employing lexica for emotion detection. Firstly, lexica often cover only a small portion of a dataset's vocabulary. Furthermore, the way they are constructed can vary widely: from lab conditions in the field of psychology (Bradley & Lang, 1999), over crowdsourced annotations (Mohammad & Turney, 2013) to distant supervision (Mohammad & Kiritchenko, 2015). All these construction methods cause a certain amount of noise, either because of divergence in emotion assessment within or between annotators, or because of imperfections in automatic lexicon creation. Finally, there is currently no consensus on a standard emotion framework: categorical frameworks and dimensional frameworks coexist, in which theorists provide

many different sets of categorical labels (Ekman, 1992; Plutchik, 1980) or dimensional axes (Mehrabian & Russell, 1974; Fontaine et al., 2007). This versatility is also reflected in the existing emotion lexica, which show a myriad of different categorical labels or numerical scales. The inconsistency in labels impedes the exchange of data and knowledge resources and calls for a unified emotion lexicon with a high word coverage that consolidates the annotations of the different emotion frameworks.

Although the problem of vocabulary coverage could be tackled by naively concatenating existing emotion lexica and thus having more entries, this approach does not address the problem of disparate label spaces, nor does it deal with the noise introduced during the construction of the lexica. Naively concatenating different lexica results in conflicting information, which makes the lexicon unsuitable for either research in psychology or keyword-based emotion detection, and could hamper learning in supervised machine learning. By contrast, research showed that merging sentiment lexica by using a multi-view variational autoencoder (VAE) outperforms a naive concatenation technique when the lexicon values are used as features in a supervised learning approach for sentiment analysis (Hoyle et al., 2019). The intuition of using a VAE to combine lexica, is that the VAE maps the lexica in a shared latent space, making the information in the different lexica less heterogeneous. Moreover, variational autoencoders are commonly used for their noise filtering ability (Aggarwal et al., 2018) and can thus remove the noise introduced in the lexicon construction process.

We believe there is an additional advantage in using a VAE for combining emotion lexica, namely that the dimension of the VAE's latent space can be chosen. While a dimension of three is preferred for sentiment (corresponding to positivity, neutrality and negativity), multiple sizes are possible when dealing with emotions (corresponding to different emotion frameworks). If it shows that the latent dimensions indeed correspond to interpretable emotion dimensions, it opens possibilities for creating large lexica tailored to specific emotion frameworks that are usable in psychology and straightforward keyword-based emotion labeling as well.

However, in comparison to sentiment lexica, joining emotion lexica is more complex: where sentiment lexica contain information about the polarity of words (negative-neutral-positive), emotion lexica contain more fine-grained affective states. This complexity is also reflected in the dimensionality of the lexica and the corresponding emission distributions used in the VAE: where most sentiment lexica are unidimensional, all emotion lexica used in this study are multidimensional. Moreover, different concepts or dimensions are quantified in emotion lexica, e.g. *anger*, *sadness*, *disgust*, *dominance*, etc. This contrasts with sentiment lexica, where the only concept is polarity. The question is thus whether we can still find a meaningful latent space into which the emotion lexica can be mapped.

In this paper, we examine the use of a multi-view variational autoencoder to combine eight existing English emotion lexica in a bigger, joint emotion lexicon and find indications that the chosen dimension of the latent space can be correlated to emotional dimensions present in the source lexica. We then evaluate the joint lexicon on the downstream task of emotion detection on thirteen emotion datasets, by using the lexicon values as features in a logistic/linear regression classifier. We also combine the lexicon features with word embeddings in a Bi-LSTM architecture and show that adding lexicon features improves the performance of plain word embedding models. Contrary to our expectations, the VAE lexicon does not outperform a naive concatenation of lexica, although it does outperform all individual lexica on the task of emotion detection.

Contributions: This paper contributes to the field of emotion analysis in NLP by a) presenting a unified emotion lexicon of 30,273 unique entries, automatically combined by a multi-view variational autoencoder, and show that this 8-dimensional lexicon is still interpretable b) bringing together a large set of existing emotion detection resources and thus learn more about the relationships between them; c) exploring the use of existing lexica and the joint VAE lexicon for the task of emotion detection.

Section 2 describes background on emotion frameworks and related studies that utilise lexica for emotion detection, or that combine lexica and datasets

with disparate label spaces. In Section 3, we describe the VAE model (Section 3.1) and show how to interpret the dimensions of the resulting joint emotion label space (Section 3.2). In Section 4, we describe our methodology to evaluate the VAE lexicon on the downstream task of emotion detection (Section 4.1) and report the results (Section 4.2), which we further discuss in Section 5. We end this paper with a conclusion in Section 6.

2. Related Work

In this section, we will focus on briefly discussing the different frameworks in emotion theory (Section 2.1), illustrating the use of lexica for emotion detection (Section 2.2) and describing related studies dealing with different emotion frameworks in NLP (Section 2.3).

2.1. Exploring emotion frameworks

Two main approaches of emotion representation exist, namely categorical approaches and dimensional approaches.

In the categorical approach, emotions are represented as specific discrete categories, often with some emotions considered more basic than others. Ekman's (1992) theory of six basic emotions (*joy*, *sadness*, *anger*, *fear*, *disgust*, and *surprise*) is the most well-known, but also Plutchik's (1980) wheel of emotions — in which *joy*, *sadness*, *anger*, *fear*, *disgust*, *surprise*, *trust*, and *anticipation* are considered most basic — is a common framework in emotion studies. However, many other theorists provide basic emotion frameworks, which can count up to fourteen emotion categories (Izard, 1971; Roseman, 1984).

In the dimensional emotion model, emotions are not seen as discrete emotion categories, but as dimensional concepts. That way, emotions are represented as a point in a multidimensional space. According to Mehrabian & Russell (1974), every emotional state can be described by scores on the dimensions *valence* (negativity – positivity), *arousal* (inactivity – activity) and *dominance* (submission – dominance), known as the VAD-model. However, in later work,

Russell (1980) argued that the two dimensions *valence* and *arousal* suffice for describing emotional states, whereas Fontaine et al. (2007) suggest adding a fourth dimension: *unpredictability*.

In general, categorical representations are easier to interpret compared to dimensional representations. However, categorical frameworks are more restricted, as they are mostly limited to basic emotions, while dimensional representations can describe any affective state.

Various resources have been created based on these different frameworks, including emotion lexica. Most emotion lexica provide numerical values per word, either for the dimensions *valence*, *arousal* and *dominance* (Bradley & Lang, 1999; Warriner et al., 2013; Mohammad, 2018a), or for basic emotions (Stevenson et al., 2007; Mohammad & Kiritchenko, 2015; Mohammad, 2018b). Other lexica just annotate each word with one or more emotion categories (Strapparava & Valitutti, 2004; Mohammad & Turney, 2013), corresponding to binary tags. The heterogeneity of emotion frameworks is thus reflected in the landscape of emotion lexica (see Appendix A for a detailed description of existing emotion lexica).

Also emotion datasets have been produced, in which sentences from different textual genres are annotated with emotional information with the goal of training and testing emotion detection models. Here, the categorical framework clearly dominates, mostly with label sets following Ekman's six (Strapparava & Mihalcea, 2007; Mohammad, 2012a; Li et al., 2017), Plutchik's eight (Mohammad et al., 2015; Schuff et al., 2017) or variations thereof (Alm et al., 2005; Mohammad et al., 2018). Although Strapparava & Mihalcea (2007) approach the Ekman's emotions in a dimensional way (by predicting intensities of emotion categories), the only datasets which truly employ the dimensional emotion model (for English data) are the ones of Preoţiuc-Pietro et al. (2016) and Buechel & Hahn (2017a).

2.2. Using lexica for emotion detection

Lexica have been the main approach for tackling the task of sentiment analysis (and by extension emotion detection) for a long time (Cambria et al., 2017). On their own, they can be used to score sentences in a straight-forward way (e.g. by summing scores of sentiment-bearing words in sentences and averaging them), which is the so-called keyword-based approach (Ohana & Tierney, 2009). Moreover, they can serve as features in a supervised learning setting (Bravo-Marquez et al., 2014).

In 2007, the first shared task on emotion detection was organised by Strapparava & Mihalcea (2007) as the Affective Text task in the SemEval series. The task was to identify Ekman's emotion categories and valence in news headlines. UPAR (Chaumartin, 2007) ended first in the subtask of identifying emotion categories and opted for a keyword-based approach with the sentiment lexicon `SentiWordNet` (Esuli & Sebastiani, 2006) and the Ekman emotion lexicon `WordNet Affect` (Strapparava & Valitutti, 2004). The task organizers themselves experimented with two approaches: one based on the `WordNet Affect` lexicon and one corpus-based approach, in which they made use of mood-annotated blog posts as training data (Strapparava & Mihalcea, 2008). Overall, the organizers' lexicon-based approach gave the best performance.

Chaffar & Inkpen (2011) used `WordNet Affect` scores as features (together with bag of word and n-gram features) on the Affective Text (Strapparava & Mihalcea, 2007), Tales (Alm et al., 2005) and Blogs (Aman & Szpakowicz, 2007) datasets with Decision Trees, Naive Bayes and SVM as classifiers. Also Kirange & Deshmukh (2012) performed experiments on the Affective Text dataset and used `WordNet Affect` lexicon features with an SVM. Indeed, Mohammad (2012b) shows that using affect lexica performs better in sentence-level emotion classification than uni- or bigrams alone, using `WordNet Affect` and the `NRC Emotion Lexicon` on the Affective Text and Blogs datasets to support this.

Even in recent studies, lexica are still used, for example as features in more

sophisticated machine learning systems as deep neural networks. In SemEval-2018 Task 1: Affect in Tweets (Mohammad et al., 2018), one of the subtasks was a multi-label emotion classification task (with labels *anger*, *anticipation*, *disgust*, *fear*, *joy*, *love*, *optimism*, *pessimism*, *sadness*, *surprise* and *trust*). Apart from word embeddings, emotion and sentiment lexica were the most used features. Even the two best teams (Baziotis et al., 2018; Meisheri & Dey, 2018) used a Bi-LSTM architecture where word embedding features were complemented with features from affect lexica.

However, relying on lexica to tackle the task of emotion detection has its limitations. The biggest problem is coverage: lexica are often not very extensive. Several studies have tried to expand emotion lexica and used different approaches thereto. Giulianelli & de Kok (2018) for example used a label propagation method (Zhu & Ghahramani, 2002) to expand existing emotion lexica. However, they only work with one original lexicon, namely the `NRC Emotion Lexicon` (Mohammad & Turney, 2013). The choice of lexicon was based on the label set: they used the Hashtag Emotion Corpus (Mohammad & Kiritchenko, 2015) which is labeled with the Plutchik emotion categories, and chose their lexicon accordingly. They thus did not have to manage the combination of different label sets, which is another difficulty of using emotion lexica. In the next section, we will discuss some studies that do take into account different emotion frameworks.

2.3. Dealing with different frameworks

Due to the sometimes restricted nature of lexica, a unified, expanded emotion lexicon is desirable. This need, however, is complicated by the miscellany of emotion frameworks. To give an example, the word *alien* appears in seven emotion lexica and is thus labeled in seven different ways (see Table 1).

Stevenson et al. (2007) and Buechel & Hahn (2017b, 2018) investigated mapping methods to shift between categorical and dimensional word representations. This is not only beneficial for lexicon construction, but also for making annotated corpora and tools comparable. In the first study (Stevenson et al., 2007),

Lexicon	Representations							
`Affective Norms`	V	A	D					
(1-9 interval)	4.45	4.86	3.56					
`ANEW`	V	A	D					
(1-9 interval)	5.6	5.45	4.64					
`NRC VAD`	V	A	D					
(0-1 interval)	0.41	0.615	0.491					
`NRC Emotion`	Ang	Ant	Di	F	J	Sa	Su	T
(binary)	0	0	1	1	0	0	0	0
`NRC Affect Intensity`	Ang			F	J	Sa		
(0-1 interval)*	-			0.422	-	-		
`NRC Hashtag`	Ang	Ant	Di	F	J	Sa	Su	T
(real-valued)*	-	0.657	-	0.623	-	0.640	-	-
`Stevenson`	Ang		Di	F	J	Sa		
(1-5 interval)	1.47		1.69	2.42	1.29	1.28		

Table 1: Representation of the word *alien* in different lexica. Abbreviations: A = Arousal, Ang = Anger, Ant = Anticipation, D = Dominance, Di = Disgust, F = Fear, J = Joy, Sa = Sadness, Su = Surprise, T = Trust, V = Valence.

* In some datasets, not all words get a score for each emotion category. In this example, this is indicated with -.

linear regression was used to predict dimensional (VAD) values from intensity ratings for the categories *happiness*, *anger*, *sadness*, *fear* and *disgust*, and vice versa. They found that no straightforward mapping was possible between intensities of emotion categories and VAD dimensions, but that each emotional category has a different impact on the separate dimensions.

Buechel & Hahn (2017b) trained a kNN model to learn an emotion representation mapping. They used either the intensity ratings of all categories (same as the ones from Stevenson et al. (2007)) to predict one dimension value, or the information of all dimensions to predict the rating of one category. They obtained promising results, with an average Pearson correlation of 0.872 for mapping VAD to an emotion category and 0.844 for mapping categories to dimensions. In subsequent work, a multi-task feed-forward neural network was used to perform the same task and a Pearson correlation of 0.877 was obtained for mapping dimensions to categories and 0.853 for the other direction (Buechel & Hahn, 2018). The downside of this approach is that it cannot increase the coverage of a lexicon.

Recently, Buechel et al. (2020) extended their approach so that they could create almost arbitrarily large emotion lexica in any language. The approach combines embedding-based lexicon expansion with emotion representation mapping. Moreover, it exceeds languages by employing a bilingual word translation model. Although this is not exactly a lexicon combination technique, it does allow the creation of a large emotion lexicon with the emotion labels of a desired emotion framework, on the condition that an emotion lexicon with the desired labels exists (even if this lexicon is not in the target language).

Studies that go beyond mappings and actually employ combination techniques are rare on the level of emotion lexicon construction. However, such techniques do exist for sentiment (polarity) lexica. Emerson & Declerck (2014) merged four German sentiment lexica by rescaling them linearly (multiplying all the scores by a constant factor per lexicon) and then combining the normalised scores by a Bayesian probabilistic model to calculate latent polarity values, which are assumed to be the 'true' values. The original lexica and the

merged lexicon all had polarity values on the [-1, 1] interval. The Bayesian model thus just takes care of the noise coming from different sources.

Hoyle et al. (2019) go one step further and combine six lexica with disparate scales, ranging from binary annotations over two-dimensional ratings to 9-point scales. They use a multi-view variational autoencoder to merge the lexica in a latent space of three dimensions. They evaluate these latent scores on nine sentiment analysis datasets and find that they outperform both the individual lexica as well as a naive combination of the lexica.

The problem of different emotion frameworks also emerges when dealing with datasets. Bostan & Klinger (2018) combine twelve different datasets by means of a rule-based mapping between categorical label sets (e.g. labels like *angry*, *annoyance* and *hate* map to *anger*; *acceptance*, *admiration* and *like* map to *trust*). This results in a final set of eleven emotion categories, in a multi-label approach with continuous values. However, dimensional representations (like the VAD model) are not taken into consideration.

3. Joint Emotion Space Modelling

3.1. Method

There is already a fair number of emotion lexica available for English, however, they all have their own specifics, assets and shortcomings (e.g. regarding emotion framework and vocabulary coverage). Table 2 shows an overview of eight emotion lexica with information about their labels and size. More extensive descriptions can be found in Appendix A.

For maximum vocabulary coverage, it is appropriate to combine multiple lexica when using lexicon information in an emotion detection task. However, seeing the variety of frameworks and perspectives by which lexica are annotated, this is not self-evident.

One could say that, when annotating words to create an emotion lexicon d, noise is added to the real emotion value z^w of a word $w \in W$, resulting in the observed emotion value x_d^w. This noise comes from subjective interpretations,

Name	Labels	Values	Size	Reference
`Affective Norms`	VAD	$[1\text{-}9]^3$	13,915	Warriner et al. (2013)
`ANEW`	VAD	$[1\text{-}9]^3$	1,034	Bradley & Lang (1999)
`NRC Emotion`	Plutchik's 8	$\{0, 1\}^8$	14,182	Mohammad & Turney (2013)
`NRC Affect Intensity`	Ang, F, S, J	$[0\text{-}1]^4$	4,192	Mohammad (2018b)
`NRC VAD`	VAD	$[0\text{-}1]^3$	20,007	Mohammad (2018a)
`NRC Hashtag Emotion`	Plutchik's 8	$[0\text{-}\infty]^8$	16,862	Mohammad & Kiritchenko (2015)
`Stevenson`	Ang, F, S, J, Di	$[1\text{-}5]^5$	1,034	Stevenson et al. (2007)
`WordNet Affect`	Ekman's 6	$\{0, 1\}^6$	1,113	Strapparava & Valitutti (2004)

Table 2: Overview of the used emotion lexica. Abbreviations: A = Arousal, Ang = Anger, D = Dominance, Di = Disgust, F = Fear, J = Joy, S = Sadness, V = Valence.
[ ] = continuous values, { } = discrete values. Exponents refer to the number of dimensions.

construction method, lab conditions, etc. All emotion values that are observed in a lexicon, are thus distorted. Variational autoencoders (VAEs) are commonly used as noise filters (Aggarwal et al., 2018), so the idea is that using a VAE, the latent emotion values of each word can be inferred. The noise added by annotation following different frameworks and perspectives, could thus be eliminated using this approach.

A traditional VAE consists of an encoder g that takes observed values X as input and outputs parameters for the probability distribution $P(Z|X)$ (which is approximated by a family of distributions $Q_\lambda(Z|X)$), from which we can sample to get a latent representation Z. This latent representation is in its turn used as input for a decoder d that outputs the parameters of the probability distribution of the data, in order to reconstruct the original input X (see Figure 1).

Following Hoyle et al. (2019), and as detailed hereafter, we extend the VAE to a multi-view model, in which each view corresponds to a different lexicon. While a traditional VAE can be employed to build latent representations for words coming from one lexicon and one type of annotation scheme only, a multi-view VAE can learn latent representations of words from different lexica even if they are annotated following different annotation schemes. The multiple possible schemes here represent the multiple views in the VAE. This allows us to join lexica with disparate label spaces, mapping the different labels to a common latent space and resulting in a larger, unified emotion lexicon, which we will call the VAE lexicon.

We use the development sets of thirteen datasets (see Section 4) to determine the best hyperparameters, i.e. the dimension of the latent variable (search space: 3,6,8,10, 20, 30, 40), the number of nodes in the fully-connected layer of the encoder and decoder network (search space: 82, 128, 256) and the value of the diagonal in the covariance matrices of the emission distributions (see paragraph 'Generative network', search space: 0.01, 0.05, 0.1, 0.5). We choose a VAE dimensionality of 8, 82 nodes and a value of 0.05 for the covariance matrices.

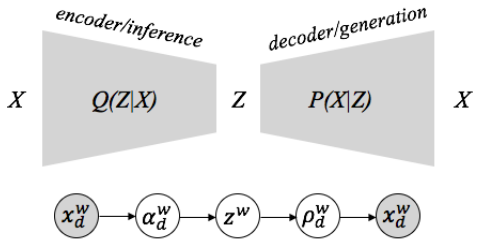


Figure 1: Variational autoencoder model.

Inference network In the first step, the latent values of z^w are drawn from the prior distribution $P(Z)$, parameterized by $\alpha^w = (\alpha_k^w = 1|k \in \{0, ..., N\})$, where N is the dimension of the latent variable, and with $\alpha_k^w = 1$ we assume a uniform prior for the latent dimensions of each word. The goal of the encoder network or inference network is to find parameters for the posterior distribution $P(Z|X)$ given the prior $P(Z)$ and $P(X|Z)$, where:

$$P(Z|X) = \frac{P(X|Z)P(Z)}{P(X)}.$$

Calculating $P(X)$ is solved by computing the integral sum:

$$P(X) = \int_Z P(X \mid Z)P(Z)dZ \tag{1}$$

However, computing the integral sum over all possible configurations of latent variables is computationally intractable, especially for higher dimensionalities of the latent space, e.g., in this work we experiment with dimensions of up to 8. Therefore, we approximate the posterior distribution with a family of distributions $Q_\lambda(Z)$:

$$Q_{\boldsymbol{\lambda}}(\mathrm{Z}) = \prod_{w \in \mathrm{W}} Q_{\boldsymbol{\beta}^w}(\boldsymbol{z}^w) = \prod_{w \in \mathrm{W}} \mathrm{Dir}(\boldsymbol{\beta}^w) \tag{2}$$

where λ are variational parameters.

Whereas the latent variables in regular VAE models are Gaussian, we use a Dirichlet latent variable. A Dirichlet latent variable is preferred as the Dirichlet distribution is a conjugate prior to the multinomial and the categorical distri-

butions. Its posterior being also a Dirichlet distribution, allows for updates of the latent variable from new observations using a closed-form expression.

We calculate $\beta^w \in R^{|Z|}$ as follows: we firstly produce lexicon-specific representations $\omega_d^w \in R^{|Z|}$ by using an encoder e with lexicon-specific parameters ϕ_d (each with a dimension of 82) over the emotion values x_d^w (see Equation 3), and then accumulate these lexicon-specific encodings across lexica according to Equation 4. The Dirichlets are thus more concentrated for words appearing in more lexica.

$$\omega_d^w = softmax\left(e\left(\boldsymbol{x}_d^w; \phi_d\right)\right) \tag{3}$$

$$\boldsymbol{\beta}^w = 1 + \sum_{d \in \mathcal{D}} \omega_d^w \tag{4}$$

Note that lexical ambiguities like homographs (words with same spelling but different meanings) are not taken into account while merging the lexica, as there is no information about context or part of speech in the lexica (except for `WordNet Affect`). Such ambiguous cases may thus have an undesirable effect on the Dirichlets.

Generative network In the decoder or generative network, X is reconstructed by outputting the likelihood of X given the latent representation Z. The joint probability distribution of the data and likelihood is defined as $P(X, Z) = P(X|Z)P(Z)$, where the distribution of the likelihood depends on the lexicon d. The decoder outputs parameters for the emission distribution $P(X|Z)$ of the data, from which X is reconstructed. This distribution is lexicon-dependent and is chosen to correspond to the label type in the individual lexicon as explained next. For `ANEW`, `Affective Norms` and `NRC VAD`, which all have three continuous labels (see Table 2), we model the emission distributions with three-dimensional Gaussians with means ρ_d^w and diagonal covariance matrices equal to $0.05\boldsymbol{I}$. `NRC Affect Intensity`, `NRC Hashtag` and `Stevenson`, which provide four, eight and five continuous labels respectively, we choose four, eight and five-dimensional Gaus-

sians as emission distributions respectively, with means ρ_d^w and diagonal covariance matrices equal to $0.05\boldsymbol{I}$. Finally, `WordNet Affect` and `NRC Emotion`, where the emotion labels are provided as six and eight binary labels respectively, we choose emission distributions of six and eight Bernoulli distributions parameterized collectively by ρ_d^w.

First, for each word w, a latent vector z^w is generated by sampling from the distribution described by the Dirichlet parameters (outputted by the inference network) (Eq. 5). For this sampling process, the generalized reparamaterization trick of Ruiz et al. (2016) is used. Then, the decoder network g (again a 82-dimensional fully-connected layer) with lexicon-specific weights θ_d transforms the generated latent emotion value z^w into a lexicon-specific representation ρ_d^w (Eq. 6). The dimension of ρ_d^w is lexicon-specific. Finally, the lexicon-specific emotion value is reconstructed from ρ_d^w with the lexicon-specific emission distribution P_d (Eq. 7).

$$z^w \sim Dir(\alpha^w) \tag{5}$$

$$\rho_d^w = g(z^w, \theta_d) \tag{6}$$

$$x_d^w = P_d(x_d^w, \rho_d^w) \tag{7}$$

3.2. Interpretation

A VAE allows us to create an extended, unified lexicon by mapping lexica with different label spaces into a common latent space and reducing noise that was introduced during the construction of source lexica. Moreover, another advantage is that the dimensionality of the latent space can be chosen. If we can correlate the dimensions of the latent space to an emotion framework, this opens possibilities to flexibly map different lexica to a target emotion framework and use the resulting lexicon in psychological research and keyword-based emotion detection.

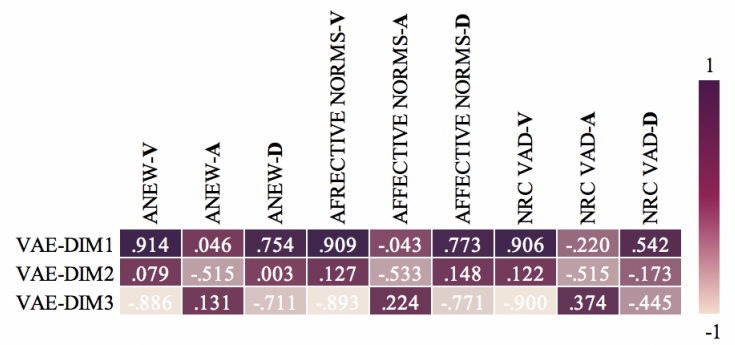
The three most popular emotion frameworks are the basic emotion models by Ekman (1992) and Plutchik (1980), consisting of respectively six and eight

emotion categories, and the VAD model by Mehrabian & Russell (1974), consisting of the three dimensions *valence*, *arousal* and *dominance*. Motivated by these frameworks, we included these sizes ($N = 3; N = 6; N = 8$) when testing different dimensionalities for the VAE lexicon.

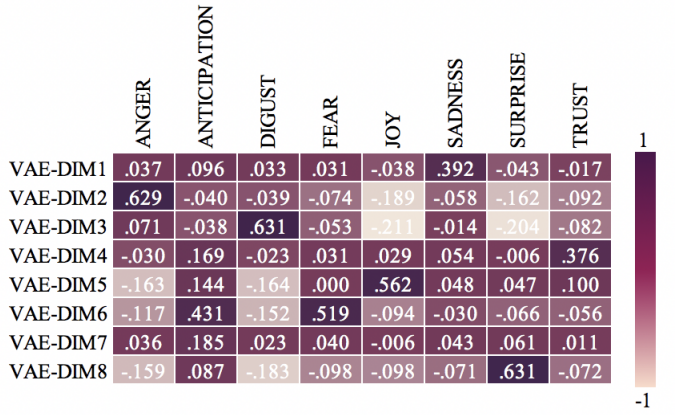
To test whether the VAE dimensions can be linked to these frameworks, we calculate Spearman's correlation between the dimensions of the VAE lexicon and the emotional dimensions in the source lexica. For measuring correlation with the VAD model, we use `ANEW`, `Affective Norms` and `NRC VAD`. For measuring the correlation with Plutchik, we use `NRC Hashtag Emotion Lexicon`. Although `NRC Emotion` is annotated with Plutchik emotions as well, we do not use it in the correlation analysis, because the labels in this lexicon are binary instead of real-valued. For the same reason, we do not correlate the VAE dimensions to the Ekman emotions (as only `WordNet Affect` is annotated with Ekman emotions and the labels are binary as well).

For VAD, we take all words that are shared between the VAE lexicon (version with $N = 3$) and `ANEW`, `Affective Norms` and `NRC VAD`, resulting in 1,034 words (corresponding to the size of the smallest lexicon included in this analysis: `ANEW`). We determine Spearman's r between the VAE values of all these words and the values in `ANEW`, `Affective Norms` and `NRC VAD` respectively. We do the same for the words in `NRC Hashtag Emotion Lexicon` (16,862 words) to determine the correlation between the VAE dimensions (version with $N = 8$) with the Plutchik categories.

Correlation coefficients for the VAE version with $N = 3$ are shown in Figure 2a. We find that there is high correlation ($r > 0.9$) between the first dimension of the VAE lexicon and *valence* and a lower but still strong correlation between this dimension and *dominance* ($0.5 < r < 0.8$). This is in line with the observation of Warriner et al. (2013) that there is a high positive correlation between *valence* and *dominance*. The second dimension has the strongest correlation with *arousal*, although this correlation is negative ($r \approx -0.5$). This means that a high value for this dimension could correspond with a low value for *arousal*. Dimension 3 shows a similar pattern to the first dimension (high



(a) 3-dimensional VAE vs. *valance* (**V**), *arousal* (**A**) and *dominance* (**D**) from `ANEW`, `Affective Norms` and `NRC VAD`.



(b) 8-dimensional VAE vs. *anger*, *anticipation*, *disgust*, *fear*, *joy*, *sadness*, *surprise* and *trust* from `NRC Hashtag Emotion`.

Figure 2: Correlation (Spearman's r) between the VAE dimensions and lexicon labels for their shared words.

	3-dimensional VAE			8-dimensional VAE							
Word	D1	D2	D3	D1	D2	D3	D4	D5	D6	D7	D8
excited	5.72	1.27	1.01	1.02	1.03	1.00	1.21	4.36	2.08	1.12	1.18
depressed	1.06	4.56	5.38	4.73	1.17	4.37	1.20	1.01	1.19	1.22	1.13
furious	1.16	1.12	6.72	1.09	4.60	2.87	1.12	1.00	1.10	1.14	1.09
gloomy	1.04	4.91	3.05	4.95	1.24	1.92	1.22	1.01	1.20	1.29	1.17
relaxed	6.76	1.96	1.28	1.50	1.22	1.08	1.56	5.44	1.23	1.56	1.41
snake	1.73	1.27	7.00	1.23	1.22	2.71	1.29	1.02	4.99	1.34	1.20
tired	1.40	2.10	3.50	1.78	2.36	1.53	1.32	1.05	1.30	1.36	1.30

Table 3: Examples of some words in the VAE lexicon (version with $N = 3$ and $N = 8$) with their values per dimension.

correlation with *valence* and *dominance*), although the correlation is negative here. It thus seems that dimension 1 and dimension 3 (when inverted) correspond to valence and dominance (or the other way around) and that arousal is encoded in dimension 2 by means of a negative correlation.

Table 3 shows a collection of words from the VAE lexicon, including 'excited', 'depressed' and 'snake'. The values are in line with the correlation insights. *Valence* and *dominance* could be interpreted as the first and third dimension (where the third dimension should be inverted), although we can not determine which one is which, and the second dimension seems to correspond to inverted arousal.

When comparing the VAE version with $N = 8$ with the Plutchik emotions, we take the `NRC Hashtag Emotion Lexicon` as reference. We first have a look at the values of some words (see Table 3). The word 'snake', for example, gets all zero values in the `NRC Hashtag Emotion Lexicon` except 0.49 for *disgust* and 0.67 for *fear*, and has the following scores in the VAE(8) lexicon: 1.23, 1.22, 2.71, 1.29, 1.02, 4.99, 1.34, 1.20. There are two values standing out (the third and sixth), possibly matching with *disgust* and *fear*.

Indeed, correlation analysis between the VAE dimensions and the categories from NRC Hashtag shows that each of the VAE dimensions has a category

which it (highly) correlates with (see Figure 2b). The first dimension can be linked to *sadness* ($r \approx 0.4$), the second to *anger* ($r \approx 0.6$), the third to *disgust* ($r \approx 0.6$), the fourth to *trust* ($r \approx 0.4$) and the fifth to *joy* ($r \approx 0.6$). The sixth dimension has the highest correlation with *fear* ($r \approx 0.5$), although it is correlated with *anticipation* as well ($r \approx 0.4$). Also the seventh dimension can be linked to *anticipation*, although the correlation is rather low ($r \approx 0.2$). The eighth dimension is correlated with *surprise* ($r \approx 0.6$). This also corresponds to the 'snake' example, which suggested that the third and sixth dimension could be interpreted as *disgust* and *fear*.

The link between VAE dimensions and Plutchik categories is thus rather clear, with only for the seventh dimension a notable ambiguity. Although *anticipation* is the Plutchik category which this dimension correlates with most highly, the correlation is rather low. Moreover, the opposite direction shows something else: when looking at the VAE dimensions which *anticipation* correlates most highly with, the sixth dimension (which we linked to *fear*) comes out as most related. It is thus not evident to say that dimension 7 needs to be interpreted as anticipation. Furthermore, if anticipation is encoded in the VAE dimensions, it is probably not only encoded through dimension 7, but through dimension 6 as well (see also the example of 'excited' in Table 3, which has a high score for dimension 6).

Overall, this correlation analysis suggests that the learned VAE dimensions can be linked to emotional dimensions present in the source lexica and that a joint VAE lexicon can still be interpretable when the dimension is chosen carefully.

4. Evaluation

4.1. Method

Datasets We evaluate the VAE joint emotion lexicon by using it as features on the downstream task of emotion detection and compare it with the use of the individual lexica and a naive concatenation thereof. The evaluation is done on

eleven commonly used emotion datasets (Blogs, Emotion in Text, ElectoralTweets, ISEAR, Tales, TEC, Affect in Tweets, SSEC, Affective Text, EmoBank and Facebook-VA) and two additional datasets also suited for emotion detection (DailyDialog and Emotion-Stimulus). Information about the labels and size of the datasets is given in Table 4 and extra information is given in Appendix B.

Eight out of thirteen datasets are annotated in a single-label multi-class categorical approach (each instance has one out of multiple classes as label) and are used for emotion classification. For these datasets, ordinary accuracy (percentage of correct predictions) is our evaluation metric of interest. Two datasets have a multi-label multi-class setup, meaning that each instance can have multiple classes as labels. One instance can thus both have *joy* and *surprise* as labels, which is not possible in the single-label datasets (there it would be *joy* or *surprise* instead of both). Multi-label classification can therefore be seen as the aggregation of multiple binary classification subtasks. We report Jaccard accuracy, a metric specifically used in multi-label tasks and defined as the size of the intersection of the predicted labels and gold labels, divided by the size of their union. Lastly, three datasets have dimensional annotations and are used in a regression task where we calculate Pearson correlations to measure the agreement between gold standard and predicted scores. The Pearson correlations are averaged over the number of dimensions in the particular dataset (6 dimensions in Affective Text, 3 in EmoBank and 2 in Facebook-VA). If a train-test split is provided in the original dataset, we use this. Otherwise, we create an 80:20 train-test split. 10% from all data in the training set is reserved for development.

Machine learning models Traditional machine learning and neural methods are used to assess the use of lexica as features for emotion detection. As traditional machine learning approach, we use a logistic regression classifier for the categorical datasets (for the multi-label datasets, we build separate binary classifiers for each of the categories and join the predictions afterwards) and linear

Type	Name	Labels	Size	Reference
SL	Blogs	Ekman, neutral	4,090	Aman & Szpakowicz (2007)
	Emotion in Text	13 categories	40,000	CrowdFlower
	DailyDialog	Ekman, neutral	13,118	Li et al. (2017)
	Electoral Tweets	19 categories	4,058	Mohammad et al. (2015)
	Emotion-Stimulus	Ekman + Sh	2,414	Ghazi et al. (2015)
	ISEAR	Ang, D, F, G, J, Sa, Sh	7,665	Scherer & Wallbott (1994)
	Tales	Ang, F, J, Sa, Su	1,207	Alm et al. (2005)
	TEC	Ekman	21,051	Mohammad (2012a)
ML	Affect in Tweets	Plutchik + L, O, P	10,983	Mohammad et al. (2018)
	SSEC	Plutchik	4,868	Schuff et al. (2017)
Reg	Affective Text	Ekman, V ({0, 1, ... 100})	1,250	Strapparava & Mihalcea (2007)
	EmoBank	VAD ([1-5])	10,548	Buechel & Hahn (2017a)
	Facebook-VA	V, A ({1, 2, ... 9})	2,895	Preoţiuc-Pietro et al. (2016)

Table 4: Overview of the used emotion datasets.

Abbreviations: A = Arousal, Ang = Anger, D = Disgust, F = Fear, J = Joy, G = Guilt, L = Love, ML = Multi-Label, O = Optimism, P = Pessimism, Reg = Regression, Sa = Sadness, Sh = Shame, SL = Single-Label, Su = Surprise, V = Valence.

regression for the continuous datasets (again, we build separate models for each of the dimensions). The logistic regression classifier uses a liblinear solver with L2 regularization and C=1.0. For the neural method, we use a bi-directional LSTM with three layers of size 900 and a dot attention layer. Loss functions are negative log likelihood, binary cross entropy and mean squared error loss for single-label, multi-label and regression datasets respectively.

Each word in the utterance of interest is represented as a vector with its lexicon scores as values. The dimension of the vector is thus equal to the number of labels in the lexicon. These scores are then averaged over the words to get an average vector for the complete data instance. In some lexica, not all words get scores for every label (see e.g. *Anger*, *Joy* and *Sadness* in `NRC Affect Intensity` in Table 1). In that case, we treat the label as 0.

VAE dimensionality We determine which dimensionality of the VAE latent space is most appropriate in the emotion detection experiments on the described datasets. Different sizes for the hidden variable are used, motivated by their correspondence to psychological emotion frameworks. We try dimensionalities $N = 3, N = 6, N = 8$ and $N = 40$, corresponding to the VAD model, the models of Ekman and Plutchik and the dimension of the naively concatenated lexicon feature vector respectively. We further try dimensions $N = 10, N = 20$ and $N = 30$ to assess whether just adding more features is more predictive than learning a valuable representation of the data. We use a logistic regression classifier for the categorical datasets and linear regression for the continuous datasets.

Experiments In a first set of experiments, we use the information from each lexicon separately as features in a simple machine learning model to predict labels/scores for each of the thirteen datasets. We use a logistic regression classifier for the categorical datasets and linear regression for the continuous datasets.

Using the same algorithms, we also compute performances for all datasets

when the different lexica are combined. We explore three options: using a naive concatenation of all lexica (resulting in a combined feature vector of dimension 40), using the VAE joint lexicon, and using a naive combination where the VAE lexicon is concatenated as an additional lexicon (feature vector with dimension 40 + number of latent dimensions).

We then explore the use of the combined lexica in a neural network approach, using a bi-directional LSTM. Our data is transformed to lexicon vectors (using the naively concatenated representation, the VAE lexicon and the naive concatenation plus VAE lexicon). Each data instance is thus again represented as a feature vector where the words are represented by their lexicon scores. We compare two versions of the network: one where we update the weights (lexicon scores) while training, and one where we keep the weights fixed.

We further test whether lexica can offer complementary gains to neural approaches, which typically rely solely on embeddings. We do this by concatenating the lexicon features in the Bi-LSTM with GloVe word embeddings (200-dimensional embeddings, pre-trained on Twitter data) (Pennington et al., 2014) and the state-of-the-art BERT embeddings (Devlin et al., 2019) as input features. Because we are merely interested in comparing different approaches and not in finding the best model per se, we do not perform a large grid search over hyperparameters for our networks. For BERT, we simply use the pretrained BERT model and the PyTorch interface for BERT by Hugging Face (Wolf et al., 2019) and use the word vectors of the last layer (768-dimensional embeddings) in the BERT model as input word vectors. We investigate how our Bi-LSTM performs with only word embeddings as features and how the performance alters when lexicon features are added (again with the three scenarios discussed above). We both try fine-tuning the embeddings and keeping the pre-trained embeddings fixed.

We are interested in a) the strengths of individual lexica, for example regarding agreement of framework between lexicon and dataset or the effect of lexicon size and construction method; b) the effect of combining lexica compared to using individual lexica, more specifically when using latent representations from a

VAE compared to a naive concatenation of lexica; c) the performance of different machine learning methods (although we limit ourselves to basic approaches and do not heavily tune those) and word representations; d) the performance of using lexica in combination with word embeddings compared to word embeddings or lexica on their own and e) the performance of using fixed lexicon scores compared to trainable inputs.

4.2. Results

4.2.1 VAE dimensionality First, we determine which dimensionality of the VAE latent space is most appropriate for using it as features to predict emotions in thirteen datasets (logistic regression for classification and linear regression for regression). Different dimensionalities are investigated: $N = \{3, 6, 8, 10, 20, 30, 40\}$. Table 5 shows the performance of the different VAE versions when used as features in the linear model. Differences in performance seem to be only minor. Indeed, when testing for significance using Welch ANOVA, no significant difference in mean between the performances of different VAE dimensionalities (using performance on separate datasets as data points) is found ($F = 0.053, P = 0.999$ for single label datasets; $F = 0.020, P = 1.000$ for multi-label datasets; $F = 0.073, P = 0.998$ for regression datasets)[1]. As $N = 8$ often leads to the highest results and it was shown we can correlate this dimensionality with Plutchik framework, we choose $N = 8$ as final dimensionality.

4.2.2 Individual lexica We train linear and logistic regression classifiers for each dataset separately with lexicon scores as features (see Table 4 for an overview of the target labels in each dataset). Table 6 (upper part) reports the average accuracy, aggregated over the different single-label datasets, average

[1] We took the performance on each dataset as data points per VAE dimensionality. For the single-label datasets, we have $8 * 7$ datapoints (8 datasets and 7 VAE versions); for the multi-label datasets, we take each label as a separate data point, resulting in $19 * 7$ data points; for regression, we again take each dimension as a separate data point, resulting in $11 * 7$ data points. We compare means by Welch ANOVA for unequal variances.

Dataset	Metric	#classes	3-dim	6-dim	8-dim	10-dim	20-dim	30-dim	40-dim
Blogs	Acc.	7	0.695	0.709	0.707	0.708	0.708	**0.714**	0.708
Emotion in Text		13	0.271	0.280	0.286	0.290	**0.291**	0.289	0.278
DailyDialog		7	0.817	0.819	**0.820**	0.818	**0.820**	**0.820**	0.819
ElectoralTweets		19	0.250	0.253	**0.264**	0.259	0.255	0.253	0.247
Emotion-Stimulus		7	0.536	0.646	0.644	0.636	0.604	**0.665**	0.633
ISEAR		7	0.271	0.378	0.374	0.372	0.388	**0.390**	0.353
Tales		5	0.467	0.554	**0.574**	0.562	0.550	0.537	0.500
TEC		6	0.417	0.451	0.454	**0.466**	0.458	0.450	0.437
Affect in Tweets	Jacc.	11	0.332	0.381	0.394	0.388	**0.401**	0.381	0.356
SSEC		8	0.425	0.434	0.432	**0.439**	0.432	0.433	0.426
Affective Text	r		0.308	0.336	0.323	**0.347**	0.304	0.303	0.294
EmoBank			0.262	0.282	0.298	0.308	0.320	**0.344**	0.343
Facebook-VA			0.382	0.384	0.385	0.386	**0.393**	0.386	0.392

Table 5: Emotion prediction results per dataset for the VAE lexicon with different dimensions. Accuracy is reported for single-label classification, Jaccard accuracy for multi-label classification and Pearson's r for regression. The best results per dataset are marked in bold.
See Table 4 for an overview of the target labels in each dataset.

Jaccard accuracy for the multi-label datasets and average Pearson correlation for the regression datasets. The results for individual datasets are shown in Table 8 (Segment 1). First, the individual lexica are used separately, and overall, the `NRC Hashtag` lexicon is the most predictive one. More specifically, `NRC Hashtag` was the best lexicon for nine out of thirteen datasets (Emotion in Text, DailyDialog, ElectoralTweets, ISEAR, Tales, Affect In Tweets, SSEC, Affective Text and Facebook-VA). In three datasets, `NRC Affect Intensity` is the best lexicon overall. `Stevenson` gives the best performance on one dataset (Tales).

`Affective Norms`, `NRC VAD`, `ANEW` and `WordNet Affect` are most often the least predictive lexica. This indicates that lexica with a VAD-framework are less suited for emotion prediction than lexica annotated with (scores for) categories. Moreover, even for the datasets that are annotated with dimensions (*valence* and *arousal* in Facebook-VA and *valence*, *arousal* and *dominance* in EmoBank), `NRC Hashtag` and `NRC Affect Intensity` are respectively the best lexica. Although one could suggest that VAD lexica perform better on dimensional datasets than on categorical datasets (with `Affective Norms` performing second best on Facebook-VA and `NRC VAD` second best on EmoBank)

there is no sign that VAD lexica are more suitable for datasets with dimensional annotations than categorical lexica.

One factor that could influence the performance of the lexicon is the lexicon size. However, we find that this is not at all decisive. The second best performing lexicon is `NRC Affect Intensity`, but with its 4,192 unique words, this lexicon is rather small. Also `Stevenson` performs fairly well, although only containing 1,034 words. On the other hand, the largest lexicon is `NRC VAD`, but this lexicon performs rather badly (probably because it has VAD annotations instead of categorical annotations).

4.2.3 Combining lexica in linear classifiers Again using linear and logistic regression classifiers, we test combinations of the different lexica for the emotion analysis tasks. Here we get the chance to evaluate the (8-dimensional) joint VAE lexicon and compare it with a naive concatenation of the original lexica. The results are given in the lower part of Table 6 (results averaged over datasets) and in Table 8 (results for individual datasets, Segment 2).

Contrary to what we expected, the VAE lexicon performs better for only four datasets compared to the naive concatenation (ElectoralTweets, Tales, EmoBank and Facebook-VA). However, adding the VAE dimensions to the naive concatenation (resulting in a 48-dimensional feature vector), resulted in the best accuracy score for ten out of the thirteen datasets. Table 6 shows that, on average, this combination approach works best for the single-label and multi-label datasets. This seems to suggest that the VAE lexicon and the original lexica on their own capture complementary information, in the same way that unigram and bigram features can capture different aspects of useful information.

4.2.4 Combining lexica in a Bi-LSTM We compare the same three scenarios as in the previous section (naive concatenation, VAE lexicon, and naive concatenation with VAE included), but now we use a neural network with Bi-LSTM layers. Table 7 (first three rows) shows the results for these experiments,

	Single-Label	Multi-Label	Regression
	Accuracy (micro-F1)	Jaccard accuracy	Pearson's *r*
△ NRC Hashtag	**0.468**	**0.361**	**0.268**
△ NRC Affect Intensity	0.459	0.311	0.265
△ WordNet Affect	0.450	0.246	0.122
△ Stevenson	0.444	0.274	0.176
△ NRC Emotion	0.441	0.305	0.207
○ Affective Norms	0.420	0.297	0.244
○ NRC VAD	0.414	0.269	0.245
○ ANEW	0.410	0.246	0.137
combi (-vae)	0.539	0.415	0.321
vae	0.515	0.413	**0.335**
combi (+vae)	**0.549**	**0.426**	0.329

Table 6: Results aggregated over datasets (macro-average) for separate lexica and combinations of lexica with logistic/linear regression.

△ categorical lexicon

○ dimensional lexicon

Combinations of lexica: combi = naive concatenation, vae = combination with VAE lexicon; vae+combi = naive concatenation with VAE lexicon included.

Single-Label datasets = Blogs, Emotion in Text, DailyDialog, ElectoralTweets, Emotion-Stimulus, ISEAR, Tales, TEC.

Multi-Label datasets = Affect in Tweets, SSEC.

Regression datasets = Affective Text, EmoBank, Facebook-VA.

See Table 4 for an overview of the target labels in each dataset.

	Single-Label	Multi-Label	Regression
	Accuracy (micro-F1)	Jaccard accuracy	Pearson's r
combi	**0.427**	0.166	0.107
vae	0.405	0.181	**0.115**
combi+vae	0.421	**0.272**	-0.064
GloVe	0.580	0.432	0.259
GloVe+combi	**0.604**	**0.475**	0.110
GloVe+vae	0.588	0.463	**0.274**
GloVe+combi+vae	0.595	0.442	0.232
BERT	0.644	0.512	**0.397**
BERT+combi	0.637	**0.538**	0.275
BERT+vae	**0.648**	0.507	0.347
BERT+combi+vae	0.643	0.499	0.370

Table 7: Results aggregated over datasets (macro-average) for combinations of lexica in Bi-LSTM.
See Table 6 for datasets and abbreviations.

aggregated over the single-label datasets, over the multi-label ones and the regression datasets. Here, only the results where the weights of the sentence vector were updated while training are reported. Results for individual datasets and fixed inputs are reported in Table 8 (Segment 3). Overall, updating the lexicon weights performs better. This might be due to the domain discrepancy between datasets and lexica (even though we combined different lexica). Therefore, we hypothesise that training the VAE jointly with the classification network would perform better. This is something future research will need to confirm.

In general, when only lexicon features are used, the linear/logistic regression classifier (see Table 6) performs better than the Bi-LSTM, probably because the datasets are rather small and the classifier has (in this setup at least) few features to fit to, which makes it far from optimal for neural network based

approaches. We again see that the naive concatenation with the VAE lexicon included works best for the multi-label datasets and the VAE lexicon on its own works best for the regression datasets. However, on average, the naive concatenation works best for the single-label datasets.

		Blogs	Emotion In Text	Daily Dialog	Elect. Tweets	Emot. Stim.	ISEAR	Tales	TEC	AffectIn Tweets	SSEC	Affect. Text	Emo Bank	Facebook VA
		Accuracy								Jaccard accuracy		Pearson's r		
Segm. 1	△ NRC Hashtag	0.688	**0.280**	**0.820**	**0.229**	0.549	**0.379**	0.347	**0.451**	**0.311**	**0.411**	**0.294**	0.212	**0.296**
	△ NRC Affect Intensity	**0.690**	0.275	0.818	0.220	**0.641**	0.314	0.318	0.395	0.227	0.395	0.266	**0.272**	0.256
	△ WordNet Affect	0.685	0.259	0.817	0.220	0.604	0.298	0.318	0.395	0.109	0.383	0.056	0.162	0.149
	△ Stevenson	0.688	0.252	0.818	0.225	0.414	0.329	**0.405**	0.424	0.184	0.364	0.160	0.138	0.231
	△ NRC Emotion	0.689	0.272	0.817	0.220	0.520	0.290	0.335	0.387	0.235	0.374	0.175	0.218	0.226
	○ Affective Norms	0.685	0.263	0.816	0.217	0.398	0.208	0.376	0.398	0.205	0.390	0.248	0.202	0.280
	○ NRC VAD	0.684	0.252	0.816	0.220	0.401	0.196	0.335	0.406	0.164	0.374	0.279	0.224	0.230
	○ ANEW	0.687	0.250	0.816	0.224	0.330	0.214	0.368	0.392	0.129	0.364	0.101	0.126	0.185
Segm. 2.	combi (-vae)	0.718	0.308	0.821	0.255	0.678	**0.472**	0.562	0.501	0.394	0.436	0.296	0.328	0.339
	vae	0.707	0.286	0.820	0.264	0.644	0.374	0.574	0.454	0.394	0.432	**0.323**	0.298	**0.385**
	combi (+vae)	**0.726**	**0.315**	**0.822**	**0.278**	**0.683**	0.459	**0.603**	**0.506**	**0.412**	**0.440**	0.279	**0.347**	0.361
Segm. 3	fxd combi	0.677	0.249	0.816	0.221	0.267	0.224	0.285	0.404	0.018	0.282	**0.062**	**0.350**	-0.111
	trn combi	0.683	0.277	0.810	**0.270**	0.267	**0.341**	0.331	0.434	0.016	0.316	0.045	0.130	**0.145**
	fxd vae	0.685	0.249	0.817	0.222	0.254	0.188	0.289	0.406	0.016	0.269	0.044	0.133	-0.010
	trn vae	**0.687**	0.272	**0.828**	0.198	0.289	0.206	**0.351**	0.409	0.019	0.342	0.035	0.196	0.113
	fxd combi+vae	**0.687**	0.252	0.809	0.216	0.254	0.196	0.306	0.404	0.039	0.295	0.021	0.246	-0.011
	trn combi+vae	0.643	**0.294**	0.783	0.264	**0.349**	0.266	0.331	**0.438**	**0.203**	**0.342**	0.037	-0.307	0.079
Segm. 4	fxd GloVe	0.689	0.287	0.817	0.263	0.483	0.429	0.388	0.437	0.111	0.393	0.236	-0.167	0.330
	trn GloVe	0.674	0.342	0.838	0.269	0.911	0.538	0.508	0.563	0.396	0.469	0.244	0.190	0.343
	fxd GloVe+combi	0.692	0.259	0.821	0.271	0.461	0.373	0.397	0.450	0.145	0.483	0.257	-0.396	0.235
	trn GloVe+combi	**0.748**	**0.360**	**0.850**	**0.295**	**0.948**	0.515	**0.541**	**0.577**	**0.463**	0.487	0.127	-0.142	0.346
	fxd GloVe+vae	0.705	0.288	0.822	0.285	0.605	0.454	0.360	0.460	0.210	0.399	0.182	-0.178	0.040
	trn GloVe+vae	0.735	0.348	0.842	0.263	0.913	**0.541**	0.508	0.556	0.437	**0.490**	0.243	**0.230**	0.350
	fxd GloVe+combi+vae	0.686	0.263	0.812	0.281	0.475	0.406	0.405	0.473	0.282	0.440	0.298	0.089	0.305
	trn GloVe+combi+vae	0.724	0.359	0.847	0.284	0.934	0.511	0.529	0.570	0.447	0.437	**0.316**	-0.082	**0.461**
Segm. 5	fxd BERT	0.731	0.336	0.836	0.290	0.671	0.499	0.603	0.555	0.398	0.530	**0.376**	-0.200	0.611
	trn BERT	0.800	**0.389**	**0.854**	0.279	**0.857**	**0.634**	0.727	0.614	0.497	0.528	0.353	0.095	0.744
	fxd BERT+combi	0.737	0.350	0.835	0.322	0.651	0.505	0.707	0.583	0.394	0.525	0.324	**0.216**	0.646
	trn BERT+combi	0.819	0.379	0.851	0.294	0.835	0.614	0.707	0.598	**0.530**	**0.546**	0.289	-0.201	0.737
	fxd BERT+vae	0.727	0.336	0.835	0.270	0.671	0.577	0.616	0.579	0.417	0.498	0.222	-0.077	0.631
	trn BERT+vae	**0.829**	0.381	0.850	**0.328**	0.851	0.629	0.698	**0.618**	0.499	0.514	0.236	0.053	**0.753**
	fxd BERT+combi+vae	0.749	0.356	0.834	0.293	0.612	0.575	0.640	0.580	0.417	0.525	0.307	-0.112	0.629
	trn BERT+combi+vae	0.803	0.378	0.850	0.293	0.841	0.634	**0.736**	0.613	0.512	0.486	0.299	0.116	0.696

Table 8: Results per dataset for separate lexica, combinations of lexica and GloVe/BERT embeddings with logistic/linear regression or Bi-LSTM. fxd = fixed embeddings, trn = trainable embeddings. See Table 6 for other abbreviations.

Most differences in performance are not significant when the datasets are viewed as data points. Significance was tested for difference in performance between the approaches in each segment by taking the performance on each dataset as data points with separate tests for single-label, multi-label and regression datasets, using Kruskal-Wallis H test. Only for the regression datasets in Segment 1, we found a statistically significant difference in performance. See Table C.11 in Appendix C for $\boldsymbol{H}$– and $\boldsymbol{P}$-values and more information about the calculation.

4.2.5 GloVe and BERT Results aggregated over datasets are shown in Table 7 (rows 4-11). The results for individual datasets are sown in Table 8 (Segments 4 and 5). We find that, when using GloVe embeddings, adding lexicon information always boosts performance. In most cases (especially for the single-label and regression datasets), adding the naive lexicon concatenation works best, but in some adding the VAE lexicon performs better. Overall, the models with GloVe (strongly) outperform the models with only lexicon features, although models with only GloVe embeddings (without lexicon information) do not perform better than when lexicon information is added.

For BERT, we see a different pattern. Here, adding lexicon information still performs better for the majority of datasets, but not for every single one. In four cases (Emotion in Text, DailyDialog, ISEAR, Affective Text), a model with only BERT embeddings as input performs best. For the other datasets, the best performing combination was often BERT combined with the VAE lexicon. Variants of the BERT model work best for all datasets except for Emotion-Stimulus and EmoBank, where, respectively, trainable GloVe with the naive concatenation and the fixed naive concatenation in a Bi-LSTM work best. This pattern is in line with findings of related work: state-of-the-art models such as BERT lessen the need for lexicon-based features. However, we show that for the majority of datasets, adding lexicon information still offers additional gains compared to plain embedding models.

4.2.6 Comparison with benchmark results For reference, we report the highest metrics achieved in other studies dealing with the datasets of interest. Table 9 shows these scores in the metric as reported in the referred study. For two datasets – DailyDialog and Emotion-Stimulus – we have not found any benchmark results, as these datasets were originally not developed for the task of emotion detection, but for response retrieval/generation and detection of the causes of emotions respectively. Most of these results are not directly comparable with our results as different train and test splits were used.

We find higher results on four datasets, namely ISEAR, Tales, TEC and

	Metric	Ours		SOTA	
		Model	Score	Score	Reference
Blogs*	Macro-F1	BERT+combi+vae$^{\Delta}$	0.616	**0.667**	Hosseini (2017)
Emotion in Text*	Acc.	BERT	0.389	**0.415****	Li et al. (2018)
DailyDialog	Acc.	BERT	0.854		–
Electoral Tweets*	Acc.	BERT+vae	0.328	**0.568**	Mohammad et al. (2015)
Emotion-Stimulus*	Acc.	GloVe+combi	0.948		–
ISEAR*	Acc.	BERT	**0.634**	0.56	Atmaja (2019)
Tales*	Macro-F1	BERT+combi+vae$^{\Delta}$	**0.700**	0.661	Agrawal et al. (2018)
TEC*	Macro-F1	BERT	**0.535**	0.499	Purpura et al. (2019)
Affect in Tweets	Jaccard acc.	BERT+combi	0.530	**0.59**	Jabreel & Moreno (2019)
SSEC	Micro-F1	BERT+combi	**0.691**	0.62	Schuff et al. (2017)
Affective Text	Pearson's r	BERT$^{\Delta}$	0.376	**0.67**	Buechel et al. (2018)
EmoBank*	Pearson's r	combi$^{\Delta}$	0.350	**0.487*****	Wu et al. (2019)
Facebook-VA*	Pearson's r	BERT+vae	0.753	**0.794*****	Wu et al. (2019)

Table 9: Comparison of our best models with state-of-the-art results.

* No train and test split in original dataset. We used an 80:20 train-test split. 10% from all data in the training set was set apart for validation. For Blogs, Electoral Tweets and Tales, the studies we compared our results with employed 10-fold cross validation for evaluation, while In TEC, 5-fold cross validation was used. For ISEAR, the reported state-of-the-art metric was obtained in an 80:20 train-test split.

** Only 50% of data used, with a 3:1:1 train-dev-test ratio.

*** Only 40% of data used for training, 10% used for testing and 50% regarded as unlabeled data for semi-supervised learning.

$^{\Delta}$Fixed instead of trainable embeddings gave a slightly better performance here.

– For DailyDialog and Emotion-Stimulus we have not found any benchmark results, as these datasets were originally not developed for the task of emotion detection, but for response retrieval/generation and detection of emotion *stimuli* respectively.

See Table 4 for an overview of the target labels in each dataset.

SSEC, but only SSEC is directly comparable. For all of these four datasets, BERT was the best performing model, although not necessarily with the VAE lexicon. Note that we did not perform a large grid search over hyperparameters for our networks, so it is very likely that our results can be improved further by hyperparameter optimization and fine-tuning BERT representations.

5. Discussion

In this section, we discuss some insights about the use of lexica provided by the emotion detection experiments performed in Section 4. We zoom in on factors that have a potential impact on the performance of lexica, namely lexicon size, construction method and quality, label set and dimensionality, trainability of the input representation and lexicon combination strategy.

5.1. Effect of lexicon size

Different factors play a role in the performance of a lexicon. Vocabulary coverage is a crucial aspect, as lexicon features can only be useful when enough words in the text to be classified have lexicon annotations. Of course, lexicon size and vocabulary coverage are correlated, as the comparison of the individual lexica point out: although the three VAD lexica all perform rather poorly, `ANEW` is clearly worse than the other two lexica. With only 1,034 words, `ANEW` has only a limited size and a lot of words in the texts to be classified are not found in the lexicon. On the other hand, the best lexicon, `NRC Hashtag`, is fairly extended (16,862 words), supporting the hypothesis that large lexica perform better. However, the (regarding label set) rather similar datasets `Stevenson` and `NRC Affect Intensity` contain only 1,034 and 4,192 words respectively and also perform reasonably to very well.

The best scores are obtained when all lexica are combined, either as a naively combined lexicon or the joint VAE lexicon. These are by far the largest lexica used in our experiments and indeed, they perform signficantly better than the individual lexica. However, the gain given by combining lexica is much

more substantial than the gain of using the approximately 17,000 words in `NRC Hashtag Emotion Lexicon` compared to the smaller `Stevenson` or `NRC Affect Intensity`. This suggests that the benefit of combining lexica not only lies in expanding the vocabulary size, but also in combining the signals coming from various emotion frameworks to build a richer emotion representation for words.

5.2. Effect of construction method and quality

It is compelling to link the origin and quality of the lexica to their performance. One may hypothesize that lexica created under lab conditions are of higher quality and thus would be more useful when used as features in machine learning tasks. `Anew`, `Stevenson` and `Affective Norms` were created manually under lab conditions, but show some notable variability among raters: `Anew` and `Affective Norms` have ratings from 1 to 7 and have average standard deviations of 1.65, 2.37 and 2.06 for *valence*, *arousal* and *dominance* in `Anew` and 1.68, 2.30 and 2.16 in `Affective Norms`; `Stevenson` has ratings from 1 to 5 for *anger*, *fear*, *sadness*, *joy* and *disgust* and has standard deviations between 0.9 and 1 for all dimensions (standard deviations were calculated per word across the ratings of all raters and then the average was taken per dimension). `NRC Emotion` and `NRC Affect Intensity` have been collected through crowdsourcing, which could have an affect on quality as well. `WordNet Affect` was annotated manually but then automatically expanded with `WordNet` relations. There is no information about the annotators or quality. Interestingly, the best performing lexicon has been constructed automatically. Lexica created under lab conditions do not necessarily perform well (`ANEW` performs badly and `Affective Norms` only average), while crowdsourced lexicon annotations can give fairly good results (as in the case of `NRC Affect Intensity`).

5.3. Effect of label set

We find that lexica with categorical annotations perform better than VAD lexica, on the condition that the categorical lexica have real-valued annotations instead of binary values.

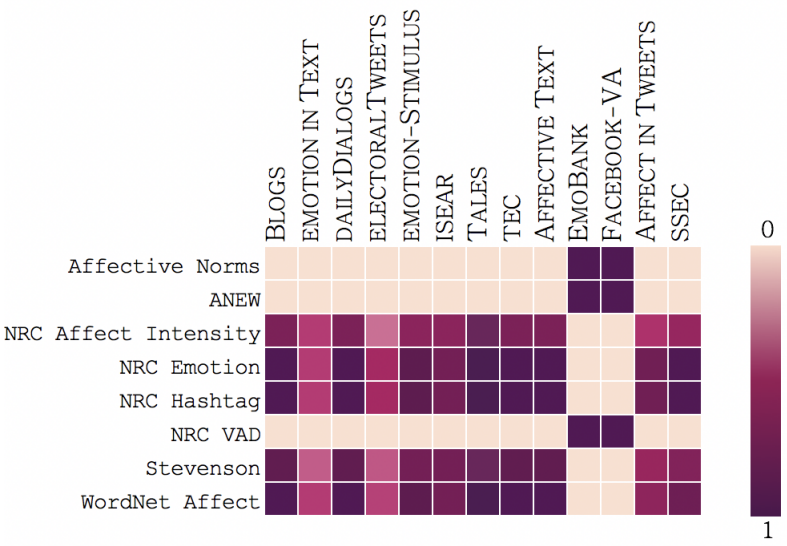


Figure 3: Visualization of overlap between lexica and datasets (number of labels that are shared between the lexicon's and the dataset's label set, normalized by label set size in the dataset).

We figured label overlap could play a role in the performance (meaning that the more labels in the lexicon that overlap with the target labels of the dataset, the better the lexicon would perform on that dataset). We define label overlap as the number of labels that are shared between the lexicon's and the dataset's label set, normalized by label set size in the dataset. For example: label overlap between the `NRC Affect Intensity` lexicon and the Electoral Tweets dataset is 0.21 (Electoral Tweets has 19 labels, of which 4 are shared with the `NRC Affect Intensity` lexicon). Figure 3 visualizes the label overlap between each lexicon and each dataset.

For each dataset, we have eight label overlap scores and eight accuracy scores (one for each lexicon), which allows us to calculate Pearson correlation between the label overlaps and accuracies per dataset (see Table 10). For regression datasets, the label overlap hypothesis does not hold: VAD lexica are not better in predicting VAD scores in datasets than categorical lexica are (see Table 6 in

Dataset	r	Dataset	r
Blogs	0.243	Affect in Tweets	0.642
Emotion in Text	0.651	SSEC	0.326
DailyDialog	0.724	Affective Text	-0.179
Electoral Tweets	0.416	EmoBank	-0.170
Emotion-Stimulus	0.736	Facebook-VA	0.000
ISEAR	0.910		
Tales	0.447		
TEC	0.315		

Table 10: Pearson correlation between overlap and performance per dataset.

Section 4.2). This might be due to the complexity of annotating VAD (Mohammad, 2018a), resulting in different interpretations of the concepts in the lexica and target datasets and making other – non-VAD – lexica more informative. For half of the remaining datasets, there is a high correlation between label overlap and accuracy. This translates into the claim that the more labels are annotated in the lexicon (and thus the higher the chance of a big label overlap), the better the lexicon performs.

We also had a look at the coefficients in the linear and logistic regression classifiers to get some more intuition into which lexica were most important in the naive concatenation. The coefficients are shown in Table C.12 in Appendix C. The influence of different lexica is not equal for all datasets. Even per dataset, it does not seem the case that there are particular datasets that have more importance than others, but we rather see particular emotion categories or dimensions that have consistently more weight (*anger* is the most pronounced one). However, we can not draw conclusions for lexica as a whole by studying these coefficients.

5.4. Effect of training embeddings

In the neural network approaches, the input representation of our data consists of word embeddings and/or lexicon vectors. The lexicon vectors can be

seen as a pre-trained word embedding, which are concatenated or not with the pre-trained GloVe or BERT embeddings. We perform experiments with fixed pre-trained embeddings and investigate updating the learned weights of the words while training the Bi-LSTM model.

When lexicon vectors are used on their own, updating the word vectors increases performance in more than half of the datasets. When combined with GloVe or BERT embeddings, the trainable setting performs better in almost all cases. This means that tailoring the model to a specific dataset is valuable. Moreover, this also suggests that emotions (or the association of words with certain emotions) are not universal, but rather dataset-dependent or domain-specific. Further research where we get more insights on how emotion scores alter across domains is therefore desirable.

5.5. Effect of lexicon combination strategy

We consistently test the difference in performance of a naively concatenated lexicon, the joint VAE lexicon, and a naively concatenated lexicon where the VAE lexicon is included.

In the logistic/linear regression approaches, the naive concatenation with the VAE lexicon included performs best on average. This seems to suggest that the VAE space and the original lexica on their own capture complementary information, in the same way that unigram and bigram features can capture different aspects of useful information.

However, in the neural network approaches, the kind of lexicon information that performs best strongly varies over datasets. Combined with GloVe embeddings, it is often the naive concatenation that works best, but combined with BERT, the VAE lexicon results in the best accuracy on several datasets.

While adding lexicon information almost always outperforms the GloVe-embedding-only approach (regardless of the lexicon combination strategy), this observation does not completely hold for BERT. In around half of the cases, the performance of BERT embeddings could not be improved by adding lexicon information, probably because large pre-trained models are already very strong

and already encode some kind of emotion information. However, since lexicon information does improve performance in the other cases, we believe employing lexica still has value, especially when there is no access to large pre-trained models like in low-resource languages.

Contrary to what we expected, the VAE lexicon is not unambiguously better than a naive concatenation when used as features in supervised machine learning for emotion detection. We see different explanations for this. Firstly, it could be that the latent emotion representation contains relevant information that is lost during mapping the latent emotion representation from the VAE to the specific emotion framework from the target dataset. Secondly, the conflicting information in the naive concatenation could be less problematic than anticipated. We based this hypothesis on the results of a similar comparison for sentiment lexica, where it was shown that a VAE outperformed a naive concatenation (Hoyle et al., 2019). However, when merging sentiment lexica, only one concept (sentiment polarity) is playing, while in emotion lexica, multiple concepts are quantified (like *valence*, *arousal*, *anger*, *sadness*, etc.). In merging sentiment lexica with a VAE, the concept is thus preserved (it is rather the scales that are unified instead), while in merging emotion lexica with a VAE, some concepts could get underrepresented. This could lead to lower performance than a naive concatenation, where all concepts are preserved (and 'conflicting information' could even be useful). However, in other applications of lexica (in psychological experiments or straightforward keyword-based emotion tagging), such conflicting information is unwanted and a unified label set is desired. Therefore, we believe the VAE joint emotion lexicon is still a valuable resource.

6. Conclusion

This paper addressed the problem of disparate label spaces in emotion lexica and presented an extended, unified lexicon containing 30,273 unique entries. The lexicon was obtained by merging eight existing emotion lexica with a multi-view variational autoencoder. We showed that we can choose the dimension of

the VAE latent space so that it is still interpretable, corresponding to emotional dimensions present in the source lexica.

We evaluated the VAE lexicon by using it as features in the downstream task of emotion detection on thirteen datasets and compared it with the use of the individual source lexica and a naive concatenation thereof, allowing us to explore the impact of different lexicon characteristics like construction approaches, label sets and vocabulary coverage.

We found that lexica with categorical annotations perform better than VAD lexica, on the condition that the categorical lexica have real-valued annotations instead of binary values. Generally, it seems that the more labels are annotated in the lexicon, the better the classification performance on the dataset. In practice, this means that out of the existing emotion lexica, the Plutchik-annotated `NRC Hashtag Emotion Lexicon` is best. Also in the VAE lexicon, we found that a latent dimension of $N = 8$ works best, and linked these dimensions to the Plutchik emotions.

We trained linear and logistic regression classifiers with lexicon scores as features and Bi-LSTMs with GloVe or BERT embeddings and lexicon vectors as input representations. We found that the VAE lexicon outperforms individual lexica, but in contrast to what we expected, the VAE approach is not always convincingly better than the naive concatenation. However, it does contribute to the naive concatenation when added as an extra lexicon, suggesting that it captures complementary information to the individual lexica.

Overall, the BERT model with lexicon features from emotion lexica performs best on average, with the best lexicon combination strategy varying over datasets. This means that emotion lexica can offer complimentary information to even extremely large pre-trained models. Moreover, when large pre-trained models are not available, as in the case of low-resource languages, we believe lexica are especially useful. It would thus be interesting to explore these approaches for low-resource languages in future research.

Although the VAE lexicon performed not unambiguously better as features in a supervised machine learning systems for emotion detection, it does have

advantages compared to a naive concatenation of lexica. In contrast to a naive concatenation, the VAE lexicon has a unified label set and can thus be used in other applications of lexica, like in psychological experiments or keyword-based emotion tagging.

Acknowledgements

This research was carried out with funding of the Research Foundation - Flanders under a Strategic Basic Research fellowship and supported with a travel grant from Research Foundation - Flanders.

References

Aggarwal, C. C. et al. (2018). *Neural networks and deep learning*. Springer.

Agrawal, A., An, A., & Papagelis, M. (2018). Learning emotion-enriched word representations. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 950–961). Santa Fe, New Mexico, USA: Association for Computational Linguistics.

Alm, C. O., Roth, D., & Sproat, R. (2005). Emotions from text: Machine learning for text-based emotion prediction. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing* (pp. 579–586). Vancouver, British Columbia, Canada: Association for Computational Linguistics.

Aman, S., & Szpakowicz, S. (2007). Identifying expressions of emotion in text. In V. Matoušek, & P. Mautner (Eds.), *Text, Speech and Dialogue* (pp. 196–205). Berlin, Heidelberg: Springer Berlin Heidelberg.

Atmaja, B. T. (2019). Deep learning-based categorical and dimensional emotion recognition for written and spoken text. Unpublished. INA-Rxiv. June 7. doi:10.31227/osf.io/fhu29.

Baziotis, C., Nikolaos, A., Chronopoulou, A., Kolovou, A., Paraskevopoulos, G., Ellinas, N., Narayanan, S., & Potamianos, A. (2018). NTUA-SLP at SemEval-2018 task 1: Predicting affective content in tweets with deep attentive RNNs and transfer learning. In *Proceedings of The 12th International Workshop on Semantic Evaluation* (pp. 245–255). New Orleans, Louisiana: Association for Computational Linguistics.

Bostan, L. A. M., & Klinger, R. (2018). An analysis of annotated corpora for emotion classification in text. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 2104–2119). Association for Computational Linguistics.

Bradley, M. M., & Lang, P. J. (1999). *Affective norms for English words (ANEW): Instruction manual and affective ratings*. Technical Report University of Florida.

Bravo-Marquez, F., Mendoza, M., & Poblete, B. (2014). Meta-level sentiment models for big social data analysis. *Knowledge-Based Systems*, *69*, 86–99.

Buechel, S., & Hahn, U. (2017a). Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers* (pp. 578–585).

Buechel, S., & Hahn, U. (2017b). A flexible mapping scheme for discrete and dimensional emotion representations. In *CogSci 2017 Proceedings*.

Buechel, S., & Hahn, U. (2018). Emotion representation mapping for automatic lexicon construction (mostly) performs on human level. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 2892–2904). Santa Fe, New Mexico, USA: Association for Computational Linguistics.

Buechel, S., Rücker, S., & Hahn, U. (2020). Learning and evaluating emotion lexicons for 91 languages. In *Proceedings of the 58th Annual Meeting of*

the Association for Computational Linguistics (pp. 1202–1217). Online: Association for Computational Linguistics. URL: https://www.aclweb.org/anthology/2020.acl-main.112. doi:10.18653/v1/2020.acl-main.112.

Buechel, S., Sedoc, J., Schwartz, H. A., & Ungar, L. (2018). Learning neural emotion analysis from 100 observations: The surprising effectiveness of pretrained word representations. *arXiv preprint arXiv:1810.10949*, .

Cambria, E., Das, D., Bandyopadhyay, S., & Feraco, A. (2017). *A practical guide to sentiment analysis*. Springer.

Chaffar, S., & Inkpen, D. (2011). Using a heterogeneous dataset for emotion analysis in text. In *Canadian conference on artificial intelligence* (pp. 62–67). Springer.

Chaumartin, F.-R. (2007). UPAR7: A knowledge-based system for headline sentiment tagging. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)* (pp. 422–425). Prague, Czech Republic: Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics.

Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion*, *6*, 169–200.

Emerson, G., & Declerck, T. (2014). SentiMerge: Combining sentiment lexicons in a Bayesian framework. In *Proceedings of Workshop on Lexical and Grammatical Resources for Language Processing* (pp. 30–38). Dublin, Ireland: Association for Computational Linguistics and Dublin City University.

Esuli, A., & Sebastiani, F. (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. In *LREC* (pp. 417–422). Citeseer volume 6.

Fontaine, J. R., Scherer, K. R., Roesch, E. B., & Ellsworth, P. C. (2007). The world of emotions is not two-dimensional. *Psychological Science*, *18*, 1050–1057.

Ghazi, D., Inkpen, D., & Szpakowicz, S. (2015). Detecting emotion stimuli in emotion-bearing sentences. In A. Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing* (pp. 152–165). Cham: Springer International Publishing.

Giulianelli, M., & de Kok, D. (2018). Semi-supervised emotion lexicon expansion with label propagation. *Computational Linguistics in the Netherlands Journal*, *8*, 99–121.

Hosseini, A. S. (2017). Sentence-level emotion mining based on combination of adaptive meta-level features and sentence syntactic features. *Engineering Applications of Artificial Intelligence*, *65*, 361–374.

Hoyle, A. M., Wolf-Sonkin, L., Wallach, H., Cotterell, R., & Augenstein, I. (2019). Combining Sentiment Lexica with a Multi-View Variational Autoencoder. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 635–640). Minneapolis, Minnesota: Association for Computational Linguistics.

Ide, N., Baker, C., Fellbaum, C., & Passonneau, R. (2010). The manually annotated sub-corpus: A community resource for and by the people. In *Proceedings of the ACL 2010 Conference Short Papers* (pp. 68–73). Uppsala, Sweden: Association for Computational Linguistics.

Izard, C. E. (1971). *The Face of Emotion*. Appleton-Century-Crofts.

Jabreel, M., & Moreno, A. (2019). A deep learning-based approach for multi-label emotion classification in tweets. *Applied Sciences*, *9*, 1123.

Kirange, D., & Deshmukh, R. (2012). Emotion classification of news headlines using svm. *Asian Journal of Computer Science and Information Technology*, (pp. 104–106).

Li, Y., Pan, Q., Wang, S., Yang, T., & Cambria, E. (2018). A generative model for category text generation. *Information Sciences*, *450*, 301–315.

Li, Y., Su, H., Shen, X., Li, W., Cao, Z., & Niu, S. (2017). Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 986–995). Taipei, Taiwan: Asian Federation of Natural Language Processing.

Ma, Y., Peng, H., & Cambria, E. (2018). Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive lstm. In *AAAI Conference on Artificial Intelligence*.

Mehrabian, A., & Russell, J. A. (1974). *An Approach to Environmental Psychology*. MIT Press.

Meisheri, H., & Dey, L. (2018). TCS research at SemEval-2018 task 1: Learning robust representations using multi-attention architecture. In *Proceedings of The 12th International Workshop on Semantic Evaluation* (pp. 291–299). New Orleans, Louisiana: Association for Computational Linguistics.

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, *38*, 39–41.

Mohammad, S. (2012a). #emotional tweets. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)* (pp. 246–255). Montréal, Canada: Association for Computational Linguistics.

Mohammad, S. (2012b). Portable features for classifying emotional text. In *Proceedings of the 2012 Conference of the North American Chapter of the*

Association for Computational Linguistics: Human Language Technologies (pp. 587–591). Montréal, Canada: Association for Computational Linguistics.

Mohammad, S., Bravo-Marquez, F., Salameh, M., & Kiritchenko, S. (2018). SemEval-2018 task 1: Affect in tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation* (pp. 1–17). New Orleans, Louisiana: Association for Computational Linguistics.

Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X., & Cherry, C. (2016). SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)* (pp. 31–41). San Diego, California: Association for Computational Linguistics.

Mohammad, S., Zhu, X., Kiritchenko, S., & Martin, J. (2015). Sentiment, emotion, purpose, and style in electoral tweets. *Information Processing & Management*, *51*, 480 – 499.

Mohammad, S. M. (2018a). Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of The Annual Conference of the Association for Computational Linguistics (ACL)*. Melbourne, Australia.

Mohammad, S. M. (2018b). Word affect intensities. In *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference (LREC-2018)*. Miyazaki, Japan.

Mohammad, S. M., & Kiritchenko, S. (2015). Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence*, *31*, 301–326.

Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, *29*, 436–465.

Ohana, B., & Tierney, B. (2009). Sentiment classification of reviews using sentiwordnet. In *9th. IT&T Conference*. Dublin Institute of Technology.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543).

Plutchik, R. (1980). A general psychoevolutionary theory of emotion. In R. Plutchik, & H. Kellerman (Eds.), *Theories of Emotion* (pp. 3–33). Academic Press.

Preoţiuc-Pietro, D., Schwartz, H. A., Park, G., Eichstaedt, J., Kern, M., Ungar, L., & Shulman, E. (2016). Modelling valence and arousal in facebook posts. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* (pp. 9–15).

Purpura, A., Masiero, C., Silvello, G., & Antonio Susto, G. (2019). Supervised lexicon extraction for emotion classification. In *Companion Proceedings of The 2019 World Wide Web Conference* (pp. 1071–1078). ACM.

Roseman, I. J. (1984). Cognitive determinants of emotion: A structural theory. *Review of Personality & Social Psychology*, *5*, 11–36.

Ruiz, F. J. R., Titsias, M. K., & Blei, D. M. (2016). The generalized reparameterization gradient. In *Advances in neural information processing systems* (pp. 460–468).

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

Scherer, K. R., & Wallbott, H. G. (1994). Evidence for universality and cultural variation of differential emotion response patterning. *Journal of personality and social psychology*, *66*, 310.

Schuff, H., Barnes, J., Mohme, J., Padó, S., & Klinger, R. (2017). Annotation, modelling and analysis of fine-grained emotions on a stance and sentiment detection corpus. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* (pp. 13–23). Copenhagen, Denmark: Association for Computational Linguistics.

Stevenson, R. A., Mikels, J. A., & James, T. W. (2007). Characterization of the affective norms for english words by discrete emotional categories. *Behavior Research Methods*, *39*, 1020–1024.

Strapparava, C., & Mihalcea, R. (2007). SemEval-2007 task 14: Affective text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)* (pp. 70–74). Prague, Czech Republic: Association for Computational Linguistics.

Strapparava, C., & Mihalcea, R. (2008). Learning to identify emotions in text. In *Proceedings of the 2008 ACM symposium on Applied computing* (pp. 1556–1560). ACM.

Strapparava, C., & Valitutti, A. (2004). Wordnet-affect: an affective extension of wordnet. *Vol 4.*, *4*.

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, *45*, 1191–1207.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., & Brew, J. (2019). Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, *abs/1910.03771*.

Wu, C., Wu, F., Wu, S., Yuan, Z., Liu, J., & Huang, Y. (2019). Semi-supervised dimensional sentiment analysis with variational autoencoder. *Knowledge-Based Systems*, *165*, 30–39.

Zhu, X., & Ghahramani, Z. (2002). Learning from labeled and unlabeled data with label propagation, .

Appendix A.

Overview of emotion lexica

`ANEW` The Affective Norms for English Words (`ANEW`) by Bradley & Lang (1999) is the oldest set of normative emotional ratings for English words that is still influential in emotion (analysis) studies. 1,034 words have been rated for *valence*, *arousal* and *dominance* on a 9-point scale. The ratings were obtained under lab conditions and originate from the field of psychology.

`Stevenson` Stevenson et al. (2007) provide complementary ratings for the words in `ANEW` on the five discrete emotions *anger*, *fear*, *sadness*, *joy* and *disgust*, on a scale of 1 to 5. Just as `ANEW`, these ratings were obtained under lab conditions.

`Affective Norms` With as much as 13,915 lemmas rated for *valence*, *arousal* and *dominance*, the affective norms of Warriner et al. (2013) form a substantial expansion of `ANEW`. Although originating from the psychology field, ratings were not obtained in the lab, but through crowdsourcing with Amazon Mechanical Turk.

`WordNet Affect` This resource was created by NLP researchers. Strapparava & Valitutti (2004) developed an extension of WordNet (Miller, 1995) by manually assigning affective labels to a subset of WordNet synsets, containing information about emotions, moods, attitudes, etc. This set was then expanded by marking WordNet synonyms as having the same emotion. In the SemEval-2007 Affective Text task, Strapparava & Mihalcea (2007) extracted a list of words relevant to the Ekman emotions from `WordNet Affect`. This list contains 1,116 words with binary association scores (0 or 1) for the Ekman emotions.

`NRC Emotion` These ratings were created by Mohammad & Turney (2013), specifically with the aim of using them in an NLP context. The ratings were

obtained by calling in the crowd, which resulted in 14,182 words annotated with one or multiple of the Plutchik emotions (binary assocation scores).

`NRC VAD` This lexicon is another crowdsourced resource for emotion analysis in NLP. Mohammad (2018a) obtained ratings for *valence*, *arousal* and *dominance* for 20,007 words, resulting in the largest emotion lexicon that is openly available.

`NRC Affect Intensity` Mohammad (2018b) also provide a lexicon with ratings for (the intensity of) the emotions *anger*, *fear*, *sadness* and *joy*. The lexicon contains 4,192 unique words, where each word gets a rating between 0 and 1 for one or more of the emotion categories. This was again a result of a crowdsourcing effort.

`NRC Hashtag Emotion` Unlike the previously discussed lexica, which were manually created, this lexicon was constructed automatically by computing the strength of association between a word and an emotion (based on the HASHTAG EMOTION CORPUS) (Mohammad & Kiritchenko, 2015). The lexicon contains real-valued scores for the eight Plutchik emotions for 16,862 words.

Appendix B.

Overview of emotion datasets

Blogs This is one of the oldest emotion datasets used in NLP. It was created by Aman & Szpakowicz (2007). They labeled 5,025 sentences (single-label) from blogs with the Ekman emotions, with additional labels *mixed emotion* and *no emotion*. Supplementary information like emotion intensity (low, medium, high) and emotion markers was given as well, but we will only use the single-label emotion information. Moreover, following other studies, we will only use the sentences with high agreement, resulting in a dataset with 4,090 sentences where the *mixed emotion* category is discarded.

Emotion in Text This dataset was published by CrowdFlower (currently known as Figure Eight), an online data annotation platform. Crowdsourced annotations were collected for 40,000 tweets on the emotion categories *anger*, *boredom*, *empty*, *enthusiasm*, *fun*, *happiness*, *hate*, *love*, *relief*, *sadness*, *surprise*, *worry* and a *neutral* category (single-label).

DailyDialog This fairly recent dataset was published by Li et al. (2017) and consists of 13,118 sentences from dialogs. The dataset is developed for the task of response retrieval and generation, but additionally, emotion information was annotated. The sentences are labeled following the Ekman emotions (with an additional *no emotion* label) in a single-label manner.

ElectoralTweets Mohammad et al. (2015) collected 4,058 tweets in the political domain, more specifically with the aim to analyse how public sentiment is shaped when it comes to elections. 4,058 tweets were annotated via crowdsourcing for the categories *acceptance*, *admiration*, *amazement*, *anger* (including *annoyance*, *hostility* and *fury*), *anticipation* (including *expectancy* and *interest*), *calmness* (or *serenity*), *disappointment*, *disgust*, *dislike*, *fear* (including *apprehension*, *panic* and *terror*), *hate*, *indifference*, *joy* (including *happiness* and *elation*), *like*, *sadness* (including *gloominess*, *grief* and *sorrow*), *surprise*, *trust*, *uncertainty* (or *indecision*, *confusion*) and *vigilance*. The annotations are single-label.

Emotion-Stimulus Originally, the purpose of this dataset was to identify emotion causes in texts. However, these data can also be used as an emotion detection dataset, as it contains emotion labels for 2,414 sentences. The annotations are done in a single-label manner with the Ekman categories and the additional category *shame* as labels.

ISEAR In the International Survey on Emotion Antecedents and Reactions, Scherer & Wallbott (1994) asked people to report on emotional events for the seven emotions *anger*, *disgust*, *fear*, *guilt*, *joy*, *sadness* and *shame*. The sentences

from these reports were extracted and linked to the emotion of interest, resulting in a dataset of 7,665 sentences with one out of seven labels.

Tales Although being the oldest emotion dataset in the NLP field, this dataset from Alm et al. (2005) is still a popular resource. The full dataset consists of 15,302 sentences from 185 fairy tales, annotated with the Ekman emotions, where the *surprise* category is broken up into *positive surprise* and *negative surprise* and a *neutral* label is added as well. The annotation happened in a single-label way. However, the 'high-agreement' version of this dataset, where *anger* and *disgust* are merged, no distinction is made between the kinds of *surprise* and *neutral* sentences are ignored, is used more frequently. We will therefore also rely on this reduced dataset, which comprises 1,207 sentences.

TEC The Twitter Emotion Corpus or TEC was automatically created by Mohammad (2012a) via distant supervision. Emotion word hashtags were used to collect tweets, and the hashtags were used for self-labeling. This resulted in a set of 21,051 tweets with (single-label) Ekman tags.

Affect in Tweets In contrast to the previous datasets, the instances in the Affect in Tweets dataset can have multiple labels. The annotations were obtained via crowdsourcing for the Plutchik emotions and three additional labels *love*, *optimism* and *pessimism*. The dataset was used for one of the subtasks in SemEval-2018: Affect in Tweets (Mohammad et al., 2018).

SSEC The Stance Sentiment Emotion Corpus is another multi-label dataset, published by Schuff et al. (2017). It is an extension of the stance and sentiment dataset from SemEval-2016 (Mohammad et al., 2016) and has annotations for the Plutchik emotions for 4,868 tweets.

Affective Text While the aforementioned datasets all contain discrete labels and are intended for emotion classification, the Affective Text dataset from

SemEval-2007 by Strapparava & Mihalcea (2007) can be used in regression tasks. 1,250 news headlines were scored for Ekman emotions on a 0 to 100 scale.

EmoBank This dataset by Buechel & Hahn (2017a) is also intended for emotion regression tasks. 10,548 sentences were annotated for the dimensions *valence*, *arousal* and *dominance* on a scale from 1 to 5. The sentences originate from various genres and domains, including the sentences from AFFECTIVE TEXT and subsets (blogs, essays, fiction, travel guides, ...) of the Manually Annotated Sub-Corpus of the American National Corpus Ide et al. (2010).

Facebook-VA Preoţiuc-Pietro et al. (2016) published a dataset consisting of 2,895 Facebook posts. The posts are annotated for the dimensions *valence* and *arousal* on a 9-point scale and thus are intended for regression tasks.

Appendix C.

Supplementary tables

Table C.11 shows $H-$ and P-values of Kruskal-Wallis H tests for testing significance for difference in performance between different emotion detection approaches. The performances on each dataset were taken as data points for each approach. For the multi-label and regression datasets, the performance for each dimension is taken as a separate data point.

Table C.12 shows the coefficients of the linear and logistic regression classifiers, this in order to get more intuition into which lexica were most important in the naive concatenation.

Segment		H	df	P	Segment		H	df	P
	SL	1.791	7	0.970		SL	0.315	3	0.957
Segm. 1	ML	2.71	7	0.910	Segm. 4	ML	0.054	3	0.997
	Reg	17.044	7	**0.017**		Reg	3.305	3	0.347
	SL	0.32	2	0.852		SL	0.185	3	0.980
Segm. 2	ML	0.09	2	0.956	Segm. 5	ML	0.512	3	0.916
	Reg	0.041	2	0.980		Reg	0.598	3	0.897
	SL	0.099	2	0.952					
Segm. 3	ML	0.274	2	0.872					
	Reg	0.683	2	0.711					

Table C.11: H-values, degrees of freedom and P-values of Kruskal-Wallis H tests for testing significance for difference in performance between different approaches within segments (see Table 8).

Lexicon	Emotional cat./dim.	Blogs In Text	Emotion Dialog	Daily Tweets	Elect. Stim.	Emot.	ISEAR	Tales	TEC Tweets	AffectIn	SSEC Text	Affect. Bank	Emo VA	Facebook
nrchashtag	anger	4.676	1.189	2.492	-0.581	5.071	-0.238	1.937	11.308	-0.714	0.822	-0.159	-0.121	-3.204
nrchashtag	anticipation	-0.946	-0.911	2.745	-0.502	0.355	-0.203	0.057	-1.428	0.143	-0.752	-7.347	0.040	-0.481
nrchashtag	disgust	0.856	0.231	1.916	-0.832	-0.416	-0.259	0.062	-0.730	-0.289	1.152	21.539	-0.166	-2.417
nrchashtag	fear	-0.805	-0.289	4.309	-0.193	-0.590	-0.203	-1.309	-4.102	-0.175	0.074	24.605	-0.180	-0.168
nrchashtag	joy	-1.095	-0.966	-5.290	-0.326	-3.031	-0.163	-2.199	-6.015	-0.379	-0.702	26.894	0.118	1.428
nrchashtag	sadness	-0.695	-1.150	-0.546	-0.098	-2.705	-0.160	-1.717	-1.099	-0.410	0.142	8.219	-0.340	-3.552
nrchashtag	surprise	-0.818	0.054	-2.768	-0.495	-0.142	-0.024	-0.715	-2.740	0.121	-0.499	7.010	-0.131	-2.190
nrchashtag	trust	-0.294	-0.286	-1.987	-0.081	0.338	-0.132	0.549	1.740	0.545	0.201	-6.679	0.148	-1.751
nrcaffect	anger	1.141	-0.144	-3.700	-0.055	3.890	-0.041	1.095	3.111	-0.477	0.294	60.006	0.643	6.871
nrcaffect	fear	-0.678	-0.088	-0.648	-0.238	-1.061	-0.054	-0.108	0.087	0.162	0.306	48.413	0.070	-7.338
nrcaffect	joy	-0.884	-0.235	-2.355	0.184	-2.073	-0.071	-1.036	-2.083	-0.204	-0.177	9.285	0.908	4.694
nrcaffect	sadness	-0.270	-0.234	-2.295	-0.083	-2.485	-0.079	-0.392	-0.627	-0.286	0.226	48.177	-0.726	-0.192
wordnetaffect	anger	1.625	-0.143	-2.010	0.064	6.940	-0.009	1.158	2.453	-0.192	0.073	0.000	-0.096	-3.532
wordnetaffect	disgust	-0.193	0.212	-0.919	-0.045	-0.361	-0.007	0.105	0.031	0.094	-0.069	64.406	-0.094	-1.044
wordnetaffect	fear	-0.911	-0.108	-3.958	0.014	-0.264	-0.007	-0.897	-1.328	0.137	0.109	10.673	-0.385	-2.137
wordnetaffect	joy	-0.352	-0.076	-2.265	0.036	-1.645	-0.030	-0.843	-0.357	-0.562	0.079	-7.936	0.490	0.318
wordnetaffect	sadness	0.029	-0.146	-2.870	0.014	-2.931	-0.024	-0.940	1.222	-0.085	0.056	2.903	-0.435	-0.581
wordnetaffect	surprise	0.368	0.174	-2.493	-0.033	-0.949	-0.008	0.378	0.147	0.155	-0.127	337.291	0.625	2.142
stevenson	happiness	-1.092	-0.695	-1.727	-0.064	-0.691	-0.445	-1.557	-0.517	-0.219	0.085	-15.913	-0.020	-1.081
stevenson	anger	1.307	0.167	0.940	-0.508	1.844	-0.295	1.788	2.459	-0.148	0.060	-2.709	-0.203	-0.992
stevenson	sadness	-0.854	-0.092	0.912	-0.531	-1.512	-0.311	-0.589	-0.957	-0.201	-0.009	-14.954	0.119	0.066
stevenson	fear	-0.822	0.478	-0.950	-0.390	-0.899	-0.310	-0.406	-1.039	0.035	0.221	11.274	-0.221	-0.576
stevenson	disgust	-0.077	0.233	-0.424	-0.386	-0.204	-0.248	1.198	-2.192	-0.071	-0.122	17.204	0.139	2.084
nrcemotion	anger	1.151	-0.063	0.553	0.103	4.661	-0.082	1.231	1.366	-0.644	0.426	-40.169	-0.372	-2.967
nrcemotion	anticipation	-0.688	-0.278	-0.842	-0.170	-0.493	-0.102	-0.587	-1.429	-0.093	-0.038	3.405	0.004	-0.188
nrcemotion	disgust	0.437	-0.013	-0.690	-0.435	0.913	-0.069	1.089	-1.198	-0.111	-0.055	0.266	0.224	-2.080
nrcemotion	fear	-0.876	-0.205	-0.219	-0.261	-1.276	-0.096	0.092	0.162	-0.053	0.320	-12.310	-0.076	2.435
nrcemotion	joy	-0.917	-0.397	-3.104	0.112	-1.762	-0.077	-0.893	-0.635	-0.509	-0.106	0.292	0.265	-1.384
nrcemotion	sadness	-0.144	-0.208	-0.192	-0.086	-1.236	-0.113	-0.126	0.334	-0.414	0.438	-20.880	-0.042	-2.079
nrcemotion	surprise	-0.560	-0.195	0.927	0.110	-1.840	-0.046	-0.922	-0.184	-0.156	-0.004	-3.125	-0.022	4.592
nrcemotion	trust	-0.478	0.070	2.923	0.026	0.489	-0.135	-0.269	0.859	0.062	0.053	1.593	-0.174	0.933
affectivenorms	valence	-0.908	-0.824	-1.353	0.374	-1.194	-0.116	-1.391	-0.830	0.410	-0.463	-3.247	0.075	0.819
affectivenorms	arousal	1.717	0.737	0.445	-1.220	-0.184	-1.087	0.542	0.030	-0.210	0.411	2.239	-0.021	-0.080
affectivenorms	dominance	-0.294	-0.198	1.236	0.395	1.149	0.271	1.073	0.806	-0.325	0.229	1.418	-0.068	-0.847
nrcvad	valence	-1.330	-0.045	-1.729	0.373	-1.006	-0.250	-1.080	-2.817	0.287	-0.414	9.758	0.156	-0.944
nrcvad	arousal	0.575	0.697	-2.563	-0.065	0.866	-0.418	0.396	0.988	-0.123	0.273	10.473	-0.137	4.428
nrcvad	dominance	-0.267	0.550	3.861	0.591	0.837	-0.352	0.029	2.216	0.321	0.070	-35.464	0.119	-3.249
anew	valence	-0.046	0.075	1.488	0.334	-1.107	-0.896	-1.343	-0.232	0.310	0.145	16.234	-0.036	0.018
anew	arousal	-0.441	-0.105	-1.402	-0.439	-0.417	-0.823	-0.197	0.710	-0.091	0.075	-6.186	0.128	0.376
anew	dominance	1.311	-0.002	0.459	0.249	1.963	-0.826	1.103	0.329	-0.179	-0.246	-7.020	-0.024	0.030

Table C.12: Coefficients of the logistic and linear classifiers per dataset (naive concatenation as input features).