Modeling Word Emotion in Historical Language: Quantity Beats Supposed Stability in Seed Word Selection

Johannes Hellrich* **Sven Buechel*** **Udo Hahn**

{firstname.lastname}@uni-jena.de

Jena University Language & Information Engineering (JULIE) Lab
Friedrich-Schiller-Universität Jena, Jena, Germany
julielab.de

Abstract

To understand historical texts, we must be aware that language—including the emotional connotation attached to words—changes over time. In this paper, we aim at estimating the emotion which is associated with a given word in former language stages of English and German. Emotion is represented following the popular Valence-Arousal-Dominance (VAD) annotation scheme. While being more expressive than polarity alone, existing word emotion induction methods are typically not suited for addressing it. To overcome this limitation, we present adaptations of two popular algorithms to VAD. To measure their effectiveness in diachronic settings, we present the first gold standard for historical word emotions, which was created by scholars with proficiency in the respective language stages and covers both English and German. In contrast to claims in previous work, our findings indicate that hand-selecting small sets of seed words with supposedly stable emotional meaning is actually harm- rather than helpful.

1 Introduction

Language change is ubiquitous and, perhaps, most evident in lexical semantics. In this work, we focus on changes in the affective meaning of words over time. Although this problem has been occasionally addressed in previous work (see Section 2.3), most contributions in this area are limited to a rather shallow understanding of human emotion, typically in terms of *semantic polarity* (feelings being either positive, negative or neutral). Another major shortcoming of this area is the lack of appropriate data and methodologies for evaluation. As a result, the aptness of algorithmic contributions has so far only been assessed in terms of face validity rather than quantitative performance figures (Cook and Stevenson, 2010; Buechel et al., 2016; Hamilton et al., 2016a; Hellrich et al., 2018).

To tackle those shortcomings, we first introduce adaptations of algorithms for word polarity induction to vectorial emotion annotation formats, thus enabling a more fine-grained analysis. Second, to put the evaluation of these methods on safer ground, we present two datasets of affective word ratings for English and German, respectively.[1] These have been annotated by scholars in terms of language-stage-specific emotional connotations.

We ran synchronic as well as diachronic experiments to compare different algorithms for modeling historical word emotions—the latter kind of evaluation employs our newly created gold standard. In particular, one prominent claim from previous work has been that *full-sized* emotion lexicons of contemporary language are ill-suited for inducing historical word emotion. Rather, it would be much more beneficial to select a small, *limited* set of seed words of supposedly invariant emotional meaning (Hamilton et al., 2016a). In contrast, our experiments indicate that larger sets of seed words perform better than manually selected ones despite the fact that some of their entries may not be accurate for the target language stage. Our unique historical gold standard is thus an important step towards firmer methodological underpinnings for the computational analysis of textually encoded historical emotions.

* These authors contributed equally to this work. Johannes Hellrich was responsible for selecting historical text corpora and training embedding models. Sven Buechel selected existing emotion lexicons and was responsible for modeling word emotions. The adaptation of polarity-based algorithms (Section 3), the creation of the German and English historical gold standard lexicons (Section 5.1), as well as the overall study design were done jointly.

[1] Publicly available together with experimental code at github.com/JULIELab/HistEmo

Proc. of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, pp. 1–11
Minneapolis, MN, USA, June 7, 2019.

2 Related Work

2.1 Representing Word Emotions

Quantitative models for word emotions can be traced back at least to Osgood (1953) who used questionnaires to gather human ratings for words on a wide variety of dimensional axes including "*good* vs. *bad*". Most previous work focused on varieties of such forms of semantic polarity, a rather simplified representation of the richness of human affective states—an observation increasingly recognized in sentiment analysis (Strapparava, 2016). In contrast to this bi-polar representation, the Valence-Arousal-Dominance (VAD) model of emotion (Bradley and Lang, 1994) is a well-established approach in psychology (Sander and Scherer, 2009) which increasingly attracts interest by NLP researchers (Köper and Schulte im Walde, 2016; Yu et al., 2016; Wang et al., 2016; Shaikh et al., 2016; Buechel and Hahn, 2017; Preoţiuc-Pietro et al., 2016; Mohammad, 2018). The VAD model assumes that affective states can be characterized relative to Valence (corresponding to the concept of polarity), Arousal (the degree of calmness or excitement) and Dominance (perceived degree of control). Formally, VAD spans a three-dimensional real-valued space (see Figure 1) making the prediction of such values a multi-variate regression problem (Buechel and Hahn, 2016).

Another popular line of emotion representation evolved around the notion of *basic emotions*, small sets of discrete, cross-culturally universal affective states (Scherer, 2000). Here, contributions most influential for NLP are Ekman's (1992) six basic emotions as well as Plutchik's (1980) wheel of emotion (Strapparava and Mihalcea, 2007; Mohammad and Turney, 2013; Bostan and Klinger, 2018). In order to illustrate the relationship between Ekman's basic emotions and the VAD affect space the former are embedded into the latter scheme in Figure 1.

The affective meaning of individual words is encoded in so-called *emotion lexicons*. Thanks to over two decades of efforts from psychologists and AI researchers alike, today a rich collection of empirically founded emotion lexicons is available covering both VAD and basic emotion representation for many languages (see Buechel and Hahn (2018b) for an overview). One of the best know resources of this kind are the *Affective Norms for English Words* (ANEW; Bradley and Lang, 1999)

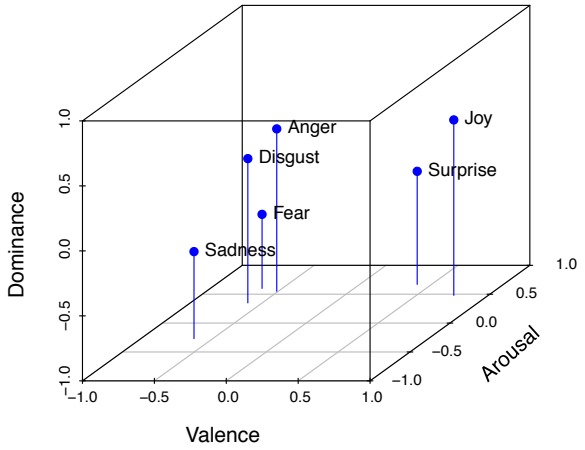

Figure 1: Affective space spanned by the Valence-Arousal-Dominance (VAD) model, together with the position of six basic emotion categories.

Entry	Valence	Arousal	Dominance
rage	2.50	6.62	4.17
orgasm	8.01	7.19	5.84
relaxed	7.25	2.49	7.09

Table 1: Sample Valence-Arousal-Dominance (VAD) ratings from the emotion lexicon by Warriner et al. (2013). The scales span the interval of $[1, 9]$ for each dimension, "5" being the neutral value.

which comprise 1,034 entries in VAD format. ANEW's popular extension by Warriner et al. (2013) comprises roughly 14k entries acquired via crowdsourcing (see Table 1 for examples).

Recently, researchers started to build computational models of the relationship between VAD and discrete categories (illustrated in Figure 1) resulting in techniques to automatically translate ratings between these major representation schemes (Calvo and Kim, 2013; Buechel and Hahn, 2018a).

2.2 Predicting Word Emotions

Word emotion induction—the task of predicting the affective score of unrated words—is an active research area within sentiment analysis (Rosenthal et al., 2015). Most approaches either rely on hand-coded lexical resources, such as WORDNET (Fellbaum, 1998), to propagate sentiment information to unkown words (Shaikh et al., 2016), or employ similarity metrics based on distributional semantics (see below). We deem the former inadequate for diachronic purposes, since almost all lexical resources typically cover contemporary language only. In the following, we focus on

algorithms which have been tested in diachronic settings in previous work. An overview of recent work focusing on applications to contemporary language is given by Buechel and Hahn (2018c).

More than a decade ago, Turney and Littman (2003) introduced a frequently used and often adopted (e.g., Köper and Schulte im Walde (2016); Palogiannidi et al. (2016)) algorithm. It computes a sentiment score based on the similarity of an unrated word to two sets of positive and negative seed words. Bestgen (2008) presented an algorithm which has been prominently put into practice in expanding a VAD lexicon to up to 17,350 entries (Bestgen and Vincze, 2012). Their method employs a k-Nearest-Neighbor methodology where an unrated word inherits the averaged ratings of the surrounding words. Rothe et al. (2016) presented a more recent approach to polarity induction. Based on word embeddings and a set of positive and negative paradigm words, they train an orthogonal transformation of the embedding space so that the encoded polarity information is concentrated in a single vector component whose value then serves as an explicit polarity rating. The algorithm proposed by Hamilton et al. (2016a) employs a random walk within a lexical graph constructed using word similarities. They outperform Rothe et al. (2016) when embeddings are trained on small datasets.

Note that these algorithms differ in the kind of input representation they require. Whereas Turney and Littman (2003), Rothe et al. (2016), and Hamilton et al. (2016a) expect binary class ratings (positive or negative), Bestgen's algorithm (Bestgen, 2008) takes vectorial seed ratings, illustrated in Table 1, as input.

2.3 Historical Sentiment Information

There are several studies using contemporary word emotion information, i.e., emotion lexicons encoding today's emotional meaning, to analyze historical documents. For instance, Acerbi et al. (2013) and Bentley et al. (2014) observed long-term trends in words expressing emotions in the Google Books corpus and linked these to historical (economic) events. Another example are Kim et al. (2017) who investigate emotions in literary texts in search for genre-specific patterns. However, this contemporary emotion information could lead to artifacts, since the emotions connected with a word are not necessarily static over time. This phenomenon is known as elevation & degeneration in historical linguistics, e.g., Old English *cniht* 'boy, servant' was elevated becoming the modern *knight* (Bloomfield, 1984).

Alternatively, algorithms for bootstrapping word emotion information can be used to predict historical emotion values by using word similarity based on historical texts. This was first done for polarity regression with the Turney and Littman (2003) algorithm and a collection of three British English corpora by Cook and Stevenson (2010). Jatowt and Duh (2014) tracked the emotional development of words by averaging the polarity of the words they co-occurred with (assuming the latters' polarity to be stable). Hamilton et al. (2016a) used their novel random walk-based algorithm for polarity regression on COHA. They consider their method especially suited for historical applications.[2] This algorithm was also used by Généreux et al. (2017) to test the temporal validity of inferred word abstractness, a psychological measure akin to the individual VAD dimensions. They used both modern and historical (1960s) psychological datasets rating the same words as gold standards and found a strong correlation with predicted historical abstractness. Buechel et al. (2016) used Bestgen (2008)'s algorithm to investigate emotional profiles of different genres of historical writing. Finally, we used the Turney and Littman (2003) algorithm to induce historical sentiment information which is provided as part of `JeSemE.org`, a website for exploring semantic change in multiple diachronic corpora (Hellrich et al., 2018).

3 Methods

3.1 Word Similarity

We measure word similarity by the cosine between word embeddings, the most recent method in studies of distributional semantics. Their most popular form are Skip-Gram Negative Sampling (SGNS; Mikolov et al., 2013) embeddings which are trained with a very shallow artificial neural network. SGNS processes one word-context pair, i.e., two nearby words, at a time and learns good embeddings by trying to predict the most likely contexts for a given word.

[2] However, the algorithm is sensitive to changes in its training material and thus likely prone to compute artifacts, see their README at `github.com/williamleif/socialsent`

An alternative solution for generating low dimensional vectors is gathering all word-context pairs for a corpus in a large matrix and reducing its dimensionality with singular value decomposition (SVD), a technique very popular in the early 1990's (Deerwester et al., 1990; Schütze, 1993). Levy et al. (2015) propose SVD_{PPMI}, a state-of-the-art algorithm based on combining SVD with the positive pointwise mutual information (PPMI; Niwa and Nitta, 1994) word association metric.

Both SGNS and SVD_{PPMI} have been shown to be adequate for exploring historical semantics (Hamilton et al., 2016b,a). A general downside of existing embedding algorithms other than SVD_{PPMI} is their inherent stochastic behavior during training which makes the resulting embedding models unreliable (Hellrich and Hahn, 2016; Antoniak and Mimno, 2018; Wendlandt et al., 2018). Very recently, contextualized word embeddings, such as ELMo (Peters et al., 2018) and BERT (Devlin et al., 2018), have started to establish themselves as a new family of algorithms for word representation. Those methods achieve enhanced performance on many downstream tasks by taking context into account, both during training and testing, to generate an individual vector representation for each individual *token*. This makes them unsuitable for our contribution, since we address emotion on the *type* level by creating emotion lexicons.

3.2 Word Emotion

Our work employs three algorithms for inducing emotion lexicons, two of which had to be adapted to deal with the more informative vectorial VAD representation instead of a simple binary two-class representation (positive vs. negative polarity):

KNN — The k-Nearest-Neighbor-based algorithm by Bestgen (2008) which already supports vectorial input.

PARASIMNUM — An adaptation of the classical PARASIM algorithm by Turney and Littman (2003) which is based on the similarity of two opposing sets of paradigm words.

RANDOMWALKNUM — An adaptation of the RANDOMWALK algorithm proposed by Hamilton et al. (2016a) which propagates affective information of seed words via a random walk through a lexical graph.

KNN sets the emotion values of each word w to the average of the emotion values of the k most similar seed words. For any given seed word s, let $e(s)$ denote its three-dimensional emotion vector corresponding to its VAD value in our seed lexicon. Furthermore, let $\text{nearest}(w, k)$ denote the set of the k seed word most similar to a given word w. Then the predicted emotion of word w according to KNN is defined as follows:

$$e_{\text{KNN}}(w, k) := \frac{1}{k} \sum_{s \in \text{nearest}(w,k)} e(s) \quad (1)$$

PARASIM computes the emotion of word w by comparing its similarity with a set of positive and negative paradigm words (POS and NEG, respectively):

$$e_{\text{PARASIM}}(w) := \sum_{p \in \text{POS}} \text{sim}(w, p) - \sum_{n \in \text{NEG}} \text{sim}(w, n) \quad (2)$$

where $\text{sim}(\cdot, \cdot)$ denotes the cosine similarly between two embedding vectors.

Let $e(s)$ map to '1', if word $s \in$ POS, and to '-1', if $s \in$ NEG, then Equation (2) can be rewritten as

$$e_{\text{PARASIM}}(w) = \sum_{s \in \text{POS} \cup \text{NEG}} \text{sim}(w, s) \times e(s). \quad (3)$$

For PARASIMNUM, our adaptation of PARASIM, we change $e(s)$ to map to a three-dimensional vector corresponding to the VAD entry of a word in our set of seed words $\mathcal{S} :=$ POS $\cup$ NEG. We also introduce a normalization factor so that the predictions according to PARASIMNUM take the form of a weighted average:

$$e_{\text{PARASIMNUM}}(w) := \frac{\sum_{s \in \mathcal{S}} \text{sim}(w, s) \times e(s)}{\sum_{s \in \mathcal{S}} \text{sim}(w, s)} \quad (4)$$

RANDOMWALK propagates sentiment scores through a graph, with vertices representing words and edge weights denoting word similarity. Let $\mathcal{V}$ represent the set of words in this lexical graph, and let the vector $p \in \mathbb{R}^{|\mathcal{V}|}$ represent the induced sentiment score for each word in the graph. To compute word emotions, p is iteratively updated by applying a transition matrix T:

$$p^{(t+1)} := \beta T p^{(t)} + (1 - \beta) s \quad (5)$$

Here $s \in \mathbb{R}^{|\mathcal{V}|}$ is the vector representing the seed sentiment scores and the β-parameter balances

between assigning similar scores to neighbors and correct scores to seeds. The vector p is initialized so that the i-th element $p_i = 1/|\mathcal{V}|$, whereas s is initialized with $s_i = 1/|\mathcal{S}|$ ($\mathcal{S}$ being the set of seed words), if the corresponding word w_i is a seed word and 0, otherwise. Details how the transition matrix is initialized can be found in Zhou et al. (2004).

To obtain the final sentiment scores p_{final}, the process is independently run until convergence for both a positive and a negative seed set, before the resulting values p^+ and p^- are normalized by performing a z-transformation on:

$$p_{\text{final}} := \frac{p^+}{p^+ + p^-} \tag{6}$$

We now provide a simple adaptation for vectorial emotion values, RANDOMWALKNUM: p and s are replaced by $|\mathcal{V}| \times 3$ matrices P and S, respectively. All entries of P are initialized with $1/|\mathcal{V}|$. For the positive seed set, S is populated with the original VAD values of each word in the seed lexicon and 0, otherwise. For the negative seed set all values are inverted relative to the center of the numerical VAD rating scales. For instance, the valence score of *relaxed* in Table 1 is transformed from 7 to 3, because 5 is the center of the respective scale. Finally, S is normalized so that each column adds up to 1. P_{final} can then be calculated analogously to the original algorithm.

4 External Datasets

4.1 Diachronic Corpora

We rely on two well curated diachronic corpora—the Corpus of Historical American English[3] (COHA; Davies, 2012) and the core corpus of the Deutsches Text Archiv[4] ['German Text Archive'] (DTA; Geyken, 2013; Geyken and Gloning, 2015). They are smaller than some alternative diachronic corpora, especially the Google Books N-gram subcorpora (Lin et al., 2012), yet their balanced nature and transparent composition should make results more resilient against artifacts (Pechenick et al., 2015). Both corpora contain metadata in the form of automatically generated POS annotations and lemmatizations. The latter appears to be more consistent in DTA, possibly due to the inclusion of an orthographic normalization step (Jurish, 2013).

COHA is relatively large for a structured corpus (Davies, 2012, p. 122) containing over 100k long and short texts from the 1810s to the 2000s. It is conceptually centered around decades and aims at providing equally sized and genre-balanced data for each decade. The only deviations are an increase in size between the 1810s and 1830s to a then stable level, as well as the inclusion of newspaper texts from the 1860s onwards. COHA is based on post-processed texts from several pre-existing collections, e.g., Project Gutenberg (Davies, 2012, p. 125), digitized with optical character recognition (OCR) software.

DTA is the closest German equivalent to COHA and the result of an ongoing effort to create a digital full-text corpus of printed German documents from the 15th to the 19th century. It is smaller than COHA, containing only about 1.3k long texts, yet of higher quality, based on extensive manual transcription (mostly double keying, in some cases corrected OCR). It contains texts from different genres, and individual texts were chosen with an eye toward cultural (not statistical) representativeness. Balance between genres is limited for some timespans, e.g., non-fiction is strongly over-represented in the early 17th century. However, the texts used in our experiments (see below) are well balanced between fictional and non-fictional texts (101 vs. 91 texts, respectively).

For both, COHA and DTA, we selected all texts from particular timespans as basis for our experiments. Those timespans served two purposes: (a) when building our gold standard of historical word emotions (Section 5.1) the annotators were requested to rate word emotions according to the respective target language stage; (b) documents associated with the respective timespan were used to train language stage-specific word embeddings (Section 6.1) in order to model those gold ratings.

The 2000s decade of COHA was an obvious fit for our synchronic experiments in Section 6.2, as it is the most recent one. For our diachronic experiments in Section 6.3 we aimed at sufficiently sized training material (10M+ tokens) to ensure high quality word embeddings. We also wanted to use data as distant from the present time as possible. We thus picked the 1830s decade of COHA for English and combined thirty years of DTA texts (1810–1839) for German—earlier COHA decades, as well as all individual DTA decades, are of insufficient size.

[3] english-corpora.org/coha/

[4] deutschestextarchiv.de — we used the May 2016 snapshot.

4.2 Emotion Lexicons

We now describe the VAD lexicons which were used to provide seed words for both synchronic and diachronic experiments. Based on its size and popularity, we chose the extended version of ANEW (Warriner et al., 2013; see Section 2) for English. Concerning German emotion lexicons, we chose the *Affective Norms for German Sentiment Terms* (ANGST; Schmidtke et al., 2014) which contain 1,003 words and largely follows ANEW's acquisition methodology.

5 Historical Gold Standard

5.1 Dataset Construction

In general, native speakers fluent in the respective (sub)language are the only viable option for acquiring a gold standard lexicon of emotional meaning for any language or domain. In the case of historical language older than about a century, this option is off the table due to biological reasons—we simply lack native speakers competent for that specific language period.

As the best conceivable surrogate, we rely on historical language experts for constructing our dataset. The gold standard consists of two parts, an English and a German one, each with 100 words. We recruited three annotators for German and two for English, all doctoral students experienced in interpreting 19th century texts.

We selected high-frequency words for the annotation to ensure high quality of the associated word embeddings. The selection was done by, first, extracting adjectives, common nouns and lexical verbs from the 1830s COHA and the 1810–1839 DTA subcorpus and then, second, randomly sampling 100 words out of the 1000 most frequent ones. We manually excluded two cases of ordinal numerals misclassified as adjectives.

The actual rating process was set up as a questionnaire study following established designs from psychological research (Bradley and Lang, 1999; Warriner et al., 2013). The participants were requested to put themselves in the position of a person living between 1810 and 1839 for the German data set, or a person living in the 1830s for the English one. They were then presented with stimulus words and used the so-called Self-Assessment Manikin (SAM; Bradley and Lang, 1994) to judge the kind of feeling evoked by these lexical items. SAM consists of three individual nine-point scales, one for each VAD dimension.

	Valence	Arousal	Dominance	Mean
goldEN	1.20	1.08	1.41	1.23
goldDE	1.72	1.56	2.31	1.86
Warriner	1.68	2.30	2.16	2.05

Table 2: Inter-annotator agreement for our English (goldEN) and German (goldDE) gold standard, as well as the lexicon by Warriner et al. (2013) for comparision; Averaged standard deviation of ratings for each VAD dimension and mean over all dimensions.

Each of the 27 rating points is illustrated by an cartoon-like anthropomorphic figure serving as a non-verbal description of the scale. Moreover, these figures are supplemented by verbal anchors for the low and high end points of the scales e.g., the rating point "9" of the Valence scale represents "complete happiness". They were not provided with or instructed to use any further material or references, e.g., dictionaries. The final ratings for each word were derived by averaging the individual ratings of the annotators.

5.2 Dataset Analysis

We measure inter-annotator agreement (IAA) by calculating the standard deviation (SD) for each word and dimension and averaging these, first, for each dimension alone, and then over these aggregate values, thus constituting an error-based score (the lower the better). Results are provided in Table 2. In comparison with the lexicon by Warriner et al. (2013), our gold standard displays higher rating consistency. As average over all three VAD dimensions, our lexicon displays an IAA of 1.23 and 1.86 for English and German, respectively, compared to 2.05 as reported by Warriner et al. (2013). This suggests that experts show higher consensus, even when judging word emotions for a historical language period, than crowdworkers for contemporary language. An alternative explanation might be differences in word material, i.e., our random sample of frequent words.

Next, we provide a short comparison of historical and modern emotion ratings. This analysis is restricted to the English language, because the overlap of the historical and modern German lexicons is really small (13 words compared to 97 for English). This difference is most likely due to the fact that the English modern lexicon is more than an order of magnitude larger than the German one.

	historical			modern		
	V	A	D	V	A	D
daughter	3.5	4.0	4.0	6.7	5.0	5.1
divine	7.0	7.0	2.0	7.2	3.0	6.0
strange	2.0	6.5	1.0	4.7	3.5	5.3

Table 3: Illustrative example words with large deviation between historical and modern affective meaning; Valence-Arousal-Dominance (VAD) of newly created gold standard compared to Warriner et al. (2013).

The Pearson correlation between modern and historical lexicons is 0.66, 0.51, and 0.31 for Valence, Arousal, and Dominance, respectively. Table 3 displays illustrative examples from our newly created gold standard where historical and modern affective meaning differ strongly. We conducted a post-facto interview on annotator motivation for those cases. Explanations—which match observations described in common reference textbooks (e.g., Brinkley (2003))—range from the influence of feminism leading to an increase in Valence for *"daughter"* up to secularization that might explain a drop in Arousal and rise in Dominance for *"divine"*. The annotation for *"strange"* was motivated by several now obsolete senses indicating foreignness or alienness.[5]

In summary, we recruited historical language experts as best conceivable surrogate to compensate for the lack of actual native speakers in order to create a gold standard for historical word emotions. To the best of our knowledge, no comparable dataset is elsewhere available, making this contribution unique and hopefully valuable for future research, despite its obvious size limitation.

6 Modeling Word Emotions

This section describes how we trained time period-specific word embeddings and used these to evaluate the algorithms presented in Section 3.2 on both a contemporary dataset and our newly created historical gold standard.

6.1 Word Embedding Training

COHA and DTA were preprocessed by using the lemmatization provided with each corpus, as well as removing punctuation and converting all text to lower case.

We used the HYPERWORDS toolkit (Levy et al., 2015) to create one distinct word embedding model for each of those subcorpora. Hyperparameter choices follow Hamilton et al. (2016a). In particular, we trained 300-dimensional word vectors, with a context window of up to four words. Context windows were limited by document boundaries while ignoring sentence boundaries. We modeled words with a minimum token frequency of 10 per subcorpus, different from Hamilton et al. (2016a). For SVD_{PPMI}, eigenvectors were discarded, no negative sampling was used and word vectors were combined with their respective context vectors.

6.2 Synchronic Evaluation

Our first evaluation of lexicon induction algorithms compares the ability of the three different algorithms described in Section 3 to predict ratings of a modern, contemporary VAD lexicon, i.e., the one by Warriner et al. (2013), using two different types of seed sets (see below). For this experiment, we used word embeddings trained on the 2000s COHA subcorpus. We call this evaluation setup *synchronic* in the linguistic sense, since seed lexicon, target lexicon and word embeddings belong to the same language period. A unique feature of our work here is that we also take into account possible interaction effects between lexicon induction algorithms and word embedding algorithms, i.e., SGNS and SVD_{PPMI}.

We use two different seed lexicons, both are based on the word ratings by Warriner et al. (2013). The *full* seed lexicon corresponds to all the entries of words which are also present in ANEW (about 1,000 words; see Section 2). In contrast, the *limited* seed lexicon is restricted to 19 words[6] which were identified as temporally stable by Hamilton et al. (2016a).

The first setup is thus analogous to the polarity experiments performed by Cook and Stevenson (2010), whereas the second one corresponds to the settings from Hamilton et al. (2016a). We use Pearson's r between actual and predicted values for each emotion dimension (Valence, Arousal and Dominance) for quantifying performance[7] and a

[5] See the Oxford English Dictionary: oed.com/view/Entry/191244

[6] One of the 20 words given by Hamilton et al. (2016a), *"hated"*, is not present in the Warriner lexicon and was therefore eliminated.

[7] Some other studies use the rank correlation coefficient Kendall's τ. We found that for our experiments the results are overall consistent between both metrics. In the following we only report Pearson's r as it is specifically designed for

Induction Method	Seed Selection	SVD_{PPMI}	SGNS
kNN	full	**0.548**	0.487
ParaSimNum	full	**0.557**	0.489
RandomWalkNum	full	**0.544**	0.436
kNN	limited	0.181	0.166
ParaSimNum	limited	0.249	0.191
RandomWalkNum	limited	**0.330**	0.181

Table 4: Results of the synchronic evaluation in Pearson's r averaged over all three VAD dimensions. The best system for each seed lexicon and those with statistically non-significant differences ($p \geq 0.05$) are in **bold**.

Language	Induction Method	Seed Selection	SVD_{PPMI}	SGNS
English	kNN	full	0.307	**0.365**
	ParaSimNum	full	0.348	0.361
	RandomWalkNum	full	0.351	0.361
	kNN	limited	0.273	0.153
	ParaSimNum	limited	0.295	0.232
	RandomWalkNum	limited	**0.305**	$0.039^{\triangle}$
German	kNN	full	0.366	0.263
	ParaSimNum	full	**0.384**	0.214
	RandomWalkNum	full	0.302	0.273

Table 5: Results of the diachronic evaluation in Pearson's r averaged over all three VAD dimensions. The best system for each language and seed selection strategy (*full* vs. *limited*) is in **bold**. Only the system marked with '$^{\triangle}$' is significantly different from the best system ($p < 0.05$).

Fisher transformation followed by a Z-test for significance testing (Cohen, 1995, pp. 130–131).

Table 4 provides the average values of these VAD correlations for each seed lexicon, embedding method and induction algorithm. SGNS embeddings are worse than SVD_{PPMI} embeddings for both full and limited seed lexicons. SVD_{PPMI} embeddings seem to be better suited for induction based on the full seed set, leading to the highest observed correlation with ParaSimNum. However, results with other induction algorithms are not significantly different. For the limited seed set, consistent with claims by Hamilton et al. (2016a), RandomWalkNum is significantly better than all alternative approaches. However, all results with the limited seed set are far (and significantly) worse than those with the full seed lexicon.

Performance is known to differ between VAD dimensions, i.e., Valence is usually the easiest one to predict. For the full seed lexicon and the best induction method, ParaSimNum with SVD_{PPMI} embeddings, we found Pearson's r correlation to range between 0.679 for Valence, 0.445 for Arousal and 0.547 for Dominance.

6.3 Diachronic Evaluation

The second evaluation set-up utilizes our historical gold standard described in Section 5.1. We call this set-up *diachronic*, since the emotion lexicons generated in our experiments aim to match word use of *historical* language stages, whereas the seed values used for this process stem from *contemporary* language. This approach allows us to test the recent claim that artificially *limiting* seed lexicons to words assumed to be semantically stable over long time spans is beneficial for generating historical emotion lexicons (Hamilton et al., 2016a). We used Pearson's r correlation and the Z-test, as in Section 6.2.

Again, we investigate interactions between lexicon induction algorithms and embedding types. For English, we evaluate with both *full* and *limited* seed lexicons, whereas for German, we evaluate only using the *full* seed lexicon (Angst, see Section 2) since most entries of the English *limited* lexicon have no corresponding entry in Angst. Embeddings are based on the 1830s COHA subcorpus for English and on the 1810–1839 DTA subcorpus for German, thus matching the time frames featured by our gold standard.

The results of this experiment are given in Table 5. For English, using the full seed lexicons, we achieve performance figures around $r = .35$. In contrast, using the *limited* seed lexicon we find that the performance is markedly weaker in each of our six conditions compared to using the full seed lexicon. This observation directly opposes the claims from Hamilton et al. (2016a) who

numerical values. In contrast, Kendall's τ only captures ordinal information and is therefore less suited for VAD.

argued that their hand selected set of emotionally stable seed words would boost performance relative to using the full, contemporary dataset as seeds.

Our finding is statistically significant in only one of all cases (the combination of SGNS and RANDOMWALKNUM). However, the fact that we get the *identical* outcomes for all the other five combinations of embedding and induction algorithm strongly indicates that using the full seed set is virtually superior, even though the differences are not statistically significant when looking at the individual conditions in isolation, due to the size[8] of our gold standard. Note that this outcome is also consistent with our results from the synchronic evaluation where we did find significant differences.

German results with the full seed lexicon are similar to those for English. Here, however, the SGNS embeddings are outperformed by SVD_{PPMI}, whereas for English both are competitive. A possible explanation for this result might be differences in pre-processing between the two data sets which were necessary due to the more complex morphology of the German language.

7 Conclusion

In this contribution, we addressed the task of constructing emotion lexicons for historical language stages. We presented adaptations of two existing polarity lexicon induction algorithms to the multidimensional VAD model of emotion, which provides deeper insights than common bipolar approaches. Furthermore, we constructed the first gold standard for affective lexical semantics in historical language. In our experiments, we investigated the interaction between word embedding algorithm, word emotion induction algorithm and seed word selection strategy. Most importantly, our results suggest that limiting seed words to supposedly temporally stable ones does not improve performance as suggested in previous work but rather turns out to be harmful. Regarding the compared algorithms for emotion lexicon induction and embedding generation, we recommend using SVD_{PPMI} together with PARASIMNUM (our adaption of the Turney and Littman (2003) algorithm), as this set-up yields strong and stable performance, and requires few hyperparameter choices. We will continue to work on further solutions to get around data sparsity issues when working with historical language, hopefully allowing for more advanced machine learning approaches in the near future.

[8] Typical emotion lexicons are one or even two orders of magnitude larger, as discussed in Section 2.1. Given the current correlation values, we would need to increase the size of our gold standard by a factor of about 40—a challenging task, given its expert reliant nature—to ensure $p < .05$.

Acknowledgments

We thank our emotion gold standard annotators for volunteering. This research was partially funded by the Deutsche Forschungsgemeinschaft (DFG) within the Graduate School *The Romantic Model* (GRK 2041/1).

References

Alberto Acerbi, Vasileios Lampos, Philip Garnett, and R. Alexander Bentley. 2013. The expression of emotions in 20th century books. *PLoS ONE*, 8(3):e59030.

Maria Antoniak and David Mimno. 2018. Evaluating the stability of embedding-based word similarities. *Transactions of the Association for Computational Linguistics*, 6:107–120.

R. Alexander Bentley, Alberto Acerbi, Paul Ormerod, and Vasileios Lampos. 2014. Books average previous decade of economic misery. *PLoS ONE*, 9(1):e83147.

Yves Bestgen. 2008. Building affective lexicons from specific corpora for automatic sentiment analysis. In *LREC 2008*, pages 496–500.

Yves Bestgen and Nadja Vincze. 2012. Checking and bootstrapping lexical norms by means of word similarity indexes. *Behavior Research Methods*, 44(4):998–1006.

Leonard Bloomfield. 1984. *Language*. University of Chicago Press. [Reprint, first published 1933].

Laura Ana Maria Bostan and Roman Klinger. 2018. An analysis of annotated corpora for emotion classification in text. In *COLING 2018, Technical Papers*, pages 2104–2119.

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

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

Alan Brinkley. 2003. *American History. A Survey*, 11th edition. McGraw Hill.

Sven Buechel and Udo Hahn. 2016. Emotion analysis as a regression problem — Dimensional models and their implications on emotion representation and metrical evaluation. In *ECAI 2016*, pages 1114–1122.

Sven Buechel and Udo Hahn. 2017. EMOBANK: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *EACL 2017, Short Papers*, pages 578–585.

Sven Buechel and Udo Hahn. 2018a. Emotion representation mapping for automatic lexicon construction (mostly) performs on human level. In *COLING 2018, Technical Papers*, pages 2892–2904.

Sven Buechel and Udo Hahn. 2018b. Representation mapping: A novel approach to generate high-quality multi-lingual emotion lexicons. In *LREC 2018*, pages 184–191.

Sven Buechel and Udo Hahn. 2018c. Word emotion induction for multiple languages as a deep multi-task learning problem. In *NAACL-HLT 2018, Long Papers*, pages 1907–1918.

Sven Buechel, Johannes Hellrich, and Udo Hahn. 2016. Feelings from the past: adapting affective lexicons for historical emotion analysis. In *LT4DH @ COLING 2016*, pages 54–61.

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

Paul R. Cohen. 1995. *Empirical Methods for Artificial Intelligence*. MIT Press.

Paul Cook and Suzanne Stevenson. 2010. Automatically identifying changes in the semantic orientation of words. In *LREC 2010*, pages 28–34.

Mark Davies. 2012. Expanding horizons in historical linguistics with the 400-million word Corpus of Historical American English. *Corpora*, 7:121–157.

Scott C. Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard A. Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. http://arxiv.org/abs/1810.04805.

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

Christiane Fellbaum, editor. 1998. WORDNET*: An Electronic Lexical Database*. MIT Press, Cambridge/MA; London/England.

Michel Généreux, Bryor Snejfella, and Marta Maslej. 2017. Big data in psychology: Using word embeddings to study theory-of-mind. In *IEEE BigData 2017*, pages 4747–4749.

Alexander Geyken. 2013. Wege zu einem historischen Referenzkorpus des Deutschen: das Projekt Deutsches Textarchiv. In Ingelore Hafemann, editor, *Perspektiven einer corpusbasierten historischen Linguistik und Philologie*, pages 221–234.

Alexander Geyken and Thomas Gloning. 2015. A living text archive of 15^{th}-19^{th}-century German. Corpus strategies, technology, organization. In Jost Gippert and Ralf Gehrke, editors, *Historical Corpora. Challenges and Perspectives*, pages 165–180. Narr.

William L. Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky. 2016a. Inducing domain-specific sentiment lexicons from unlabeled corpora. In *EMNLP 2016*, pages 595–605.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016b. Diachronic word embeddings reveal statistical laws of semantic change. In *ACL 2016, Long Papers*, pages 1489–1501.

Johannes Hellrich, Sven Buechel, and Udo Hahn. 2018. JESEME: a website for exploring diachronic changes in word meaning and emotion. In *COLING 2018, System Demonstrations*, pages 10–14.

Johannes Hellrich and Udo Hahn. 2016. Bad company—Neighborhoods in neural embedding spaces considered harmful. In *COLING 2016, Technical Papers*, pages 2785–2796.

Adam Jatowt and Kevin Duh. 2014. A framework for analyzing semantic change of words across time. In *JCDL 2014*, pages 229–238.

Bryan Jurish. 2013. Canonicalizing the Deutsches Textarchiv. In Ingelore Hafemann, editor, *Perspektiven einer corpusbasierten historischen Linguistik und Philologie*, pages 235–244.

Evgeny Kim, Sebastian Padó, and Roman Klinger. 2017. Investigating the relationship between literary genres and emotional plot development. In *LaTeCH-CLfL @ ACL 2017*, pages 17–26.

Maximilian Köper and Sabine Schulte im Walde. 2016. Automatically generated affective norms of abstractness, arousal, imageability and valence for 350,000 German lemmas. In *LREC 2016*, pages 2595–2598.

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

Yuri Lin, Jean-Baptiste Michel, Erez Lieberman Aiden, Jon Orwant, William Brockman, and Slav Petrov. 2012. Syntactic annotations for the GOOGLE BOOKS NGRAM corpus. In *ACL 2012, System Demonstrations*, pages 169–174.

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *ICLR 2013*.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *ACL 2018, Long Papers*, pages 174–184.

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

Yoshiki Niwa and Yoshihiko Nitta. 1994. Co-occurrence vectors from corpora vs. distance vectors from dictionaries. In *COLING 1994*, pages 304–309.

Charles E. Osgood. 1953. *Method and Theory in Experimental Psychology*. Oxford University Press.

Elisavet Palogiannidi, Polychronis Koutsakis, Elias Iosif, and Alexandros Potamianos. 2016. Affective lexicon creation for the Greek language. In *LREC 2016*, pages 2867–2872.

Eitan Adam Pechenick, Christopher M. Danforth, and Peter Sheridan Dodds. 2015. Characterizing the GOOGLE BOOKS corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *PLoS One*, 10(10):e0137041.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher T. Clark, Kenton Lee, and Luke S. Zettlemoyer. 2018. Deep contextualized word representations. In *NAACL-HLT 2018, Long Papers*, pages 2227–2237.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. *Emotion: Theory, Research and Experience*, 1(3):3–33.

Daniel Preoţiuc-Pietro, Hansen Andrew Schwartz, Gregory Park, Johannes C. Eichstaedt, Margaret L. Kern, Lyle H. Ungar, and Elizabeth P. Shulman. 2016. Modelling valence and arousal in FACEBOOK posts. In *WASSA @ NAACL-HLT 2016*, pages 9–15.

Sara Rosenthal, Preslav I. Nakov, Svetlana Kiritchenko, Saif M. Mohammad, Alan Ritter, and Veselin Stoyanov. 2015. SemEval 2015 Task 10: Sentiment analysis in Twitter. In *SemEval 2015*, pages 451–463.

Sascha Rothe, Sebastian Ebert, and Hinrich Schütze. 2016. Ultradense word embeddings by orthogonal transformation. In *NAACL 2016*, pages 767–777.

David Sander and Klaus R. Scherer, editors. 2009. *The Oxford Companion to Emotion and the Affective Sciences*. Oxford University Press, Oxford, U.K., New York, N.Y.

Klaus R. Scherer. 2000. Psychological models of emotion. In Joan C. Borod, editor, *The Neuropsychology of Emotion*, pages 137–162. Oxford University Press.

David S. Schmidtke, Tobias Schröder, Arthur M. Jacobs, and Markus Conrad. 2014. ANGST: Affective norms for German sentiment terms, derived from the affective norms for English words. *Behavior Research Methods*, 46(4):1108–1118.

Hinrich Schütze. 1993. Part-of-speech induction from scratch. In *ACL 1993*, pages 251–258.

Samira Shaikh, Kit Cho, Tomek Strzalkowski, Laurie Feldman, John Lien, Ting Liu, and George Aaron Broadwell. 2016. ANEW+: Automatic expansion and validation of affective norms of words lexicons in multiple languages. In *LREC 2016*, pages 1127–1132.

Carlo Strapparava. 2016. Emotions and NLP: Future directions. In *WASSA @ NAACL 2016*, page 180.

Carlo Strapparava and Rada Mihalcea. 2007. SemEval-2007 Task 14: Affective text. In *SemEval 2007*, pages 70–74.

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

Jin Wang, Liang-Chih Yu, K. Robert Lai, and Xuejie Zhang. 2016. Dimensional sentiment analysis using a regional CNN-LSTM model. In *ACL 2016, Long Papers*, pages 225–230.

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

Laura Wendlandt, Jonathan K. Kummerfeld, and Rada Mihalcea. 2018. Factors influencing the surprising instability of word embeddings. In *NAACL-HLT 2018, Long Papers*, pages 2092–2102.

Liang-Chih Yu, Lung-Hao Lee, Shuai Hao, Jin Wang, Yunchao He, Jun Hu, K. Robert Lai, and Xuejie Zhang. 2016. Building Chinese affective resources in valence-arousal dimensions. In *NAACL 2016*, pages 540–545.

Denny Zhou, Olivier Bousquet, Thomas N Lal, Jason Weston, and Bernhard Schölkopf. 2004. Learning with local and global consistency. In *NIPS 2004*, pages 321–328.