PO-EMO: Conceptualization, Annotation, and Modeling of Aesthetic Emotions in German and English Poetry

Thomas Haider[1,3], **Steffen Eger**[2], **Evgeny Kim**[3], **Roman Klinger**[3], **Winfried Menninghaus**[1]
[1]Department of Language and Literature, Max Planck Institute for Empirical Aesthetics
[2]NLLG, Department of Computer Science, Technische Universitat Darmstadt
[3]Institut für Maschinelle Sprachverarbeitung, University of Stuttgart
{thomas.haider, w.m}@ae.mpg.de, eger@aiphes.tu-darmstadt.de
{roman.klinger, evgeny.kim}@ims.uni-stuttgart.de

Abstract

Most approaches to emotion analysis of social media, literature, news, and other domains focus exclusively on basic emotion categories as defined by Ekman or Plutchik. However, art (such as literature) enables engagement in a broader range of more complex and subtle emotions. These have been shown to also include mixed emotional responses. We consider emotions in poetry as they are *elicited in the reader*, rather than what is *expressed in the text* or *intended by the author*. Thus, we conceptualize a set of *aesthetic emotions* that are predictive of aesthetic appreciation in the reader, and allow the annotation of multiple labels per line to capture mixed emotions within their context. We evaluate this novel setting in an annotation experiment both with carefully trained experts and via crowdsourcing. Our annotation with experts leads to an acceptable agreement of $\kappa = .70$, resulting in a consistent dataset for future large scale analysis. Finally, we conduct first emotion classification experiments based on BERT, showing that identifying aesthetic emotions is challenging in our data, with up to .52 F1-micro on the German subset. Data and resources are available at `https://github.com/tnhaider/poetry-emotion`.

Keywords: Emotion, Aesthetic Emotions, Literature, Poetry, Annotation, Corpora, Emotion Recognition, Multi-Label

1. Introduction

Emotions are central to human experience, creativity and behavior. Models of affect and emotion, both in psychology and natural language processing, commonly operate on predefined categories, designated either by *continuous scales* of, e.g., *Valence*, *Arousal* and *Dominance* (Mohammad, 2016) or *discrete emotion labels* (which can also vary in intensity). Discrete sets of emotions often have been motivated by theories of basic emotions, as proposed by Ekman (1992)—*Anger*, *Fear*, *Joy*, *Disgust*, *Surprise*, *Sadness*—and Plutchik (1991), who added *Trust* and *Anticipation*. These categories are likely to have evolved as they motivate behavior that is directly relevant for survival. However, *art reception* typically presupposes a situation of safety and therefore offers special opportunities to engage in a broader range of more complex and subtle emotions. These differences between real-life and art contexts have not been considered in natural language processing work so far.

To emotionally move readers is considered a prime goal of literature since Latin antiquity (Johnson-Laird and Oatley, 2016; Menninghaus et al., 2019; Menninghaus et al., 2015). Deeply moved readers shed tears or get chills and goosebumps even in lab settings (Wassiliwizky et al., 2017). In cases like these, the emotional response actually implies an aesthetic evaluation: narratives that have the capacity to move readers are evaluated as good and powerful texts for this very reason. Similarly, feelings of suspense experienced in narratives not only respond to the trajectory of the plot's content, but are also directly predictive of aesthetic liking (or disliking). Emotions that exhibit this dual capacity have been defined as "aesthetic emotions" (Menninghaus et al., 2019). Contrary to the negativity bias of classical emotion catalogues, emotion terms used for aesthetic evaluation purposes include far more positive than negative emotions. At the same time, many overall positive aesthetic emotions encompass negative or mixed emotional ingredients (Menninghaus et al., 2019), e.g., feelings of suspense include both hopeful and fearful anticipations.

For these reasons, we argue that the analysis of literature (with a focus on poetry) should rely on specifically selected emotion items rather than on the narrow range of basic emotions only. Our selection is based on previous research on this issue in psychological studies on art reception and, specifically, on poetry. For instance, Knoop et al. (2016) found that *Beauty* is a major factor in poetry reception.

We primarily adopt and adapt emotion terms that Schindler et al. (2017) have identified as aesthetic emotions in their study on how to measure and categorize such particular affective states. Further, we consider the aspect that, when selecting specific emotion labels, the perspective of annotators plays a major role. Whether emotions are *elicited in the reader*, *expressed in the text*, or *intended by the author* largely changes the permissible labels. For example, feelings of *Disgust* or *Love* might be intended or expressed in the text, but the text might still fail to elicit corresponding feelings as these concepts presume a strong reaction in the reader. Our focus here was on the actual emotional experience of the readers rather than on hypothetical intentions of authors. We opted for this reader perspective based on previous research in NLP (Buechel and Hahn, 2017a; Buechel and Hahn, 2017b) and work in empirical aesthetics (Menninghaus et al., 2017), that specifically measured the reception of poetry. Our final set of emotion labels consists of *Beauty/Joy, Sadness, Uneasiness, Vitality/Energy, Suspense, Awe/Sublime, Humor, Annoyance*, and *Nostalgia*.[1]

[1]The concepts *Beauty* and *Awe/Sublime* primarily define object-based aesthetic virtues. Kant (2001) emphasized that such virtues are typically intuitively felt rather than rationally computed. Such

In addition to selecting an adapted set of emotions, the annotation of poetry brings further challenges, one of which is the choice of the appropriate unit of annotation. Previous work considers words[2] (Mohammad and Turney, 2013; Strapparava and Valitutti, 2004), sentences (Alm et al., 2005; Aman and Szpakowicz, 2007), utterances (Cevher et al., 2019), sentence triples (Kim and Klinger, 2018), or paragraphs (Liu et al., 2019) as the units of annotation. For poetry, reasonable units follow the logical document structure of poems, i.e., verse (line), stanza, and, owing to its relative shortness, the complete text. The more coarse-grained the unit, the more difficult the annotation is likely to be, but the more it may also enable the annotation of emotions in context. We find that annotating fine-grained units (lines) that are hierarchically ordered within a larger context (stanza, poem) caters to the specific structure of poems, where emotions are regularly mixed and are more interpretable within the whole poem. Consequently, we allow the mixing of emotions already at line level through multi-label annotation.

The remainder of this paper includes (1) a report of the annotation process that takes these challenges into consideration, (2) a description of our annotated corpora, and (3) an implementation of baseline models for the novel task of aesthetic emotion annotation in poetry. In a first study, the annotators work on the annotations in a closely supervised fashion, carefully reading each verse, stanza, and poem. In a second study, the annotations are performed via crowdsourcing within relatively short time periods with annotators not seeing the entire poem while reading the stanza. Using these two settings, we aim at obtaining a better understanding of the advantages and disadvantages of an expert vs. crowdsourcing setting in this novel annotation task. Particularly, we are interested in estimating the potential of a crowdsourcing environment for the task of self-perceived emotion annotation in poetry, given time and cost overhead associated with in-house annotation process (that usually involve training and close supervision of the annotators).

We provide the final datasets of German and English language poems annotated with reader emotions on verse level at `https://github.com/tnhaider/poetry-emotion`.

2. Related Work

2.1. Poetry in Natural Language Processing

Natural language understanding research on poetry has investigated *stylistic variation* (Kaplan and Blei, 2007; Kao and Jurafsky, 2015; Voigt and Jurafsky, 2013), with a focus on broadly accepted formal features such as *meter* (Greene et al., 2010; Agirrezabal et al., 2016; Estes and Hench, 2016) and *rhyme* (Reddy and Knight, 2011; Haider and Kuhn, 2018), as well as *enjambement* (Ruiz et al., 2017; Baumann et al., 2018) and *metaphor* (Kesarwani et al., 2017; Reinig and Rehbein, 2019). Recent work has also explored the relationship of poetry and prose, mainly on a syntactic level (Krishna et al., 2019; Gopidi and Alam, 2019). Furthermore, poetry also lends itself well to semantic (change) analysis (Haider, 2019; Haider and Eger, 2019), as linguistic invention (Underwood and Sellers, 2012; Herbelot, 2014) and succinctness (Roberts, 2000) are at the core of poetic production.

Corpus-based analysis of emotions in poetry has been considered, but there is no work on German, and little on English. Kao and Jurafsky (2015) analyze English poems with word associations from the Harvard Inquirer and LIWC, within the categories *positive/negative outlook*, *positive/negative emotion* and *phys./psych. well-being*. Hou and Frank (2015) examine the binary sentiment polarity of Chinese poems with a weighted personalized PageRank algorithm. Barros et al. (2013) followed a tagging approach with a thesaurus to annotate words that are similar to the words 'Joy', 'Anger', 'Fear' and 'Sadness' (moreover translating these from English to Spanish). With these word lists, they distinguish the categories 'Love', 'Songs to Lisi', 'Satire' and 'Philosophical-Moral-Religious' in Quevedo's poetry. Similarly, Alsharif et al. (2013) classify unique Arabic 'emotional text forms' based on word unigrams.

Mohanty et al. (2018) create a corpus of 788 poems in the Indian Odia language, annotate it on text (poem) level with binary negative and positive sentiment, and are able to distinguish these with moderate success. Sreeja and Mahalakshmi (2019) construct a corpus of 736 Indian language poems and annotate the texts on Ekman's six categories + Love + Courage. They achieve a Fleiss Kappa of .48.

In contrast to our work, these studies focus on basic emotions and binary sentiment polarity only, rather than addressing aesthetic emotions. Moreover, they annotate on the level of complete poems (instead of fine-grained verse and stanza-level).

2.2. Emotion Annotation

Emotion corpora have been created for different tasks and with different annotation strategies, with different units of analysis and different foci of emotion perspective (reader, writer, text). Examples include the ISEAR dataset (Scherer and Wallbott, 1994) (document-level); emotion annotation in children stories (Alm et al., 2005) and news headlines (Strapparava and Mihalcea, 2007) (sentence-level); and fine-grained emotion annotation in literature by Kim and Klinger (2018) (phrase- and word-level). We refer the interested reader to an overview paper on existing corpora (Bostan and Klinger, 2018).

We are only aware of a limited number of publications which look in more depth into the emotion perspective. Buechel and Hahn (2017a) report on an annotation study that focuses both on writer's and reader's emotions associated with English sentences. The results show that the reader perspective yields better inter-annotator agreement. Yang et al. (2009) also study the difference between writer and reader emotions, but not with a modeling perspective. The authors find that positive reader emotions tend to be linked to positive writer emotions in online blogs.

feelings of *Beauty* and *Sublime* have therefore come to be subsumed under the rubrique of *aesthetic emotions* in recent psychological research (Menninghaus et al., 2019). For this reason, we refer to the whole set of category labels as *emotions* throughout this paper.

[2] to create emotion dictionaries

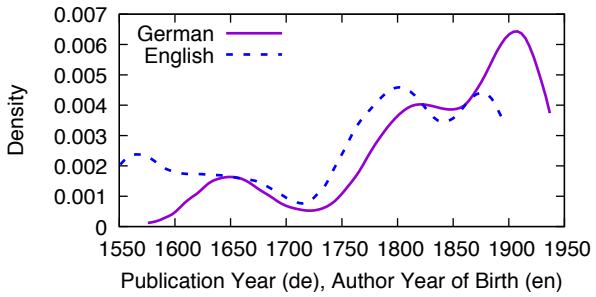

Figure 1: Temporal distribution of poetry corpora (Kernel Density Plots with bandwidth = 0.2).

	German	English
# tokens	20403	8082
# lines	3650	1240
# stanzas	731	174
# poems	158	64
# authors	51	22

Table 1: Statistics on our poetry corpora *PO-EMO*. Tokens without punctuation.

2.3. Emotion Classification

The task of emotion classification has been tackled before using rule-based and machine learning approaches. Rule-based emotion classification typically relies on lexical resources of emotionally charged words (Strapparava and Valitutti, 2004; Esuli and Sebastiani, 2006; Mohammad and Turney, 2013) and offers a straightforward and transparent way to detect emotions in text.

In contrast to rule-based approaches, current models for emotion classification are often based on neural networks and commonly use word embeddings as features. Schuff et al. (2017) applied models from the classes of CNN, BiLSTM, and LSTM and compare them to linear classifiers (SVM and MaxEnt), where the BiLSTM shows best results with the most balanced precision and recall. Abdul-Mageed and Ungar (2017) claim the highest F_1 with gated recurrent unit networks (Chung et al., 2015) for Plutchik's emotion model. More recently, shared tasks on emotion analysis (Mohammad et al., 2018; Klinger et al., 2018) triggered a set of more advanced deep learning approaches, including BERT (Devlin et al., 2019) and other transfer learning methods (Dankers et al., 2019).

3. Data Collection

For our annotation and modeling studies, we build on top of two poetry corpora (in English and German), which we refer to as *PO-EMO*. This collection represents important contributions to the literary canon over the last 400 years. We make this resource available in TEI P5 XML[3] and an easy-to-use tab separated format. Table 1 shows a size overview of these data sets. Figure 1 shows the distribution of our data over time via density plots. Note that both corpora show a relative underrepresentation before the onset of the romantic period (around 1750).

[3] https://tei-c.org/guidelines/p5/

3.1. German

The German corpus contains poems available from the website lyrik.antikoerperchen.de (ANTI-K), which provides a platform for students to upload essays about poems. The data was available in the Hypertext Markup Language, with clean line and stanza segmentation, which we transformed into TEI P5. ANTI-K also has extensive metadata, including author names, years of publication, numbers of sentences, poetic genres, and literary periods, that enable us to gauge the distribution of poems according to periods. The 158 poems we consider (731 stanzas) are dispersed over 51 authors and the New High German timeline (1575–1936 A.D.). This data has been annotated, besides emotions, for meter, rhythm, and rhyme in other studies (Haider and Kuhn, 2018; Haider et al., 2020).

3.2. English

The English corpus contains 64 poems of popular English writers. It was partly collected from Project Gutenberg with the GutenTag tool,[4] and, in addition, includes a number of hand selected poems from the modern period and represents a cross section of popular English poets. We took care to include a number of female authors, who would have been underrepresented in a uniform sample. Time stamps in the corpus are organized by the birth year of the author, as assigned in Project Gutenberg.

4. Expert Annotation

In the following, we will explain how we compiled and annotated three data subsets, namely, (1) 48 German poems with gold annotation. These were originally annotated by three annotators. The labels were then aggregated with majority voting and based on discussions among the annotators. Finally, they were curated to only include one gold annotation. (2) The remaining 110 German poems that are used to compute the agreement in table 3 and (3) 64 English poems contain the raw annotation from two annotators.

We report the genesis of our annotation guidelines including the emotion classes. With the intention to provide a language resource for the computational analysis of emotion in poetry, we aimed at maximizing the consistency of our annotation, while doing justice to the diversity of poetry. We iteratively improved the guidelines and the annotation workflow by annotating in batches, cleaning the class set, and the compilation of a gold standard. The final overall cost of producing this expert annotated dataset amounts to approximately €3,500.

4.1. Workflow

The annotation process was initially conducted by three female university students majoring in linguistics and/or literary studies, which we refer to as our "expert annotators". We used the INCePTION platform for annotation[5] (Klie et al., 2018). Starting with the German poems, we annotated in batches of about 16 (and later in some cases 32) poems.

[4] https://gutentag.sdsu.edu/

[5] https://www.informatik.tu-darmstadt.de/ukp/research_6/current_projects/inception/index.en.jsp

Factor	Items
Negative emotions	anger/distasteful
Prototypical Aesthetic Emotions	beauty/sublime/being moved
Epistemic Emotions	interest/insight
Animation	motivation/inspiration
Nostalgia / Relaxation	nostalgic/calmed
Sadness	sad/melancholic
Amusement	funny/cheerful

Table 2: Aesthetic Emotion Factors (Schindler et al., 2017).

After each batch, we computed agreement statistics including heatmaps, and provided this feedback to the annotators. For the first three batches, the three annotators produced a gold standard using a majority vote for each line. Where this was inconclusive, they developed an adjudicated annotation based on discussion. Where necessary, we encouraged the annotators to aim for more consistency, as most of the frequent switching of emotions within a stanza could not be reconstructed or justified.

In poems, emotions are regularly mixed (already on line level) and are more interpretable within the whole poem. We therefore annotate lines hierarchically within the larger context of stanzas and the whole poem. Hence, we instruct the annotators to read a complete stanza or full poem, and then annotate each line in the context of its stanza. To reflect on the emotional complexity of poetry, we allow a maximum of two labels per line while avoiding heavy label fluctuations by encouraging annotators to reflect on their feelings to avoid 'empty' annotations. Rather, they were advised to use fewer labels and more consistent annotation. This additional constraint is necessary to avoid "wild", non-reconstructable or non-justified annotations.

All subsequent batches (all except the first three) were only annotated by two out of the three initial annotators, coincidentally those two who had the lowest initial agreement with each other. We asked these two experts to use the generated gold standard (48 poems; majority votes of 3 annotators plus manual curation) as a reference ("if in doubt, annotate according to the gold standard"). This eliminated some systematic differences between them[6] and markedly improved the agreement levels, roughly from 0.3–0.5 Cohen's κ in the first three batches to around 0.6–0.8 κ for all subsequent batches. This annotation procedure relaxes the *reader* perspective, as we encourage annotators (if in doubt) to annotate how they think the other annotators would annotate. However, we found that this formulation improves the usability of the data and leads to a more consistent annotation.

4.2. Emotion Labels

We opt for measuring the *reader perspective* rather than the text surface or author's intent. To closer define and support conceptualizing our labels, we use particular 'items', as they are used in psychological self-evaluations. These items consist of adjectives, verbs or short phrases. We build on top of Schindler et al. (2017) who proposed 43 items that were then grouped by a factor analysis based on self-evaluations of participants. The resulting factors are shown in Table 2. We attempt to cover all identified factors and supplement with basic emotions (Ekman, 1992; Plutchik, 1991), where possible.

[6]One person labeled lines with more negative emotions such as *Uneasiness* and *Annoyance* and the person labeled more positive emotions such as *Vitality/Energy* and *Beauty/Joy*.

	κ		**Ann. 1 %**		**Ann. 2 %**	
	en	de	en	de	en	de
Beauty / Joy	.77	.74	.31	.30	.26	.30
Sadness	.72	.77	.21	.20	.20	.18
Uneasiness	.84	.77	.15	.19	.15	.18
Vitality / Energy	.50	.63	.12	.11	.18	.13
Awe / Sublime	.71	.61	.07	.06	.07	.06
Suspense	.58	.65	.04	.07	.07	.08
Humor	.81	.68	.04	.05	.04	.05
Nostalgia	.81	—	.03	—	.03	—
Annoyance	.62	.65	.03	.04	.02	.02

Table 3: Cohen's kappa agreement levels and normalized line-level emotion frequencies for expert annotators (Nostalgia is not available in the German data).

	English	German
avg. κ	0.707	0.688
F1	0.775	0.774
F1 Majority	0.323	0.323
F1 Random	0.108	0.119

Table 4: Top: averaged kappa scores and micro-F1 agreement scores, taking one annotator as gold. Bottom: Baselines.

We started with a larger set of labels to then delete and substitute (tone down) labels during the initial annotation process to avoid infrequent classes and inconsistencies. Further, we conflate labels if they show considerable confusion with each other. These iterative improvements particularly affected *Confusion*, *Boredom* and *Other* that were very infrequently annotated and had little agreement among annotators ($\kappa < .2$). For German, we also removed *Nostalgia* ($\kappa = .218$) after gold standard creation, but after consideration, added it back for English, then achieving agreement. *Nostalgia* is still available in the gold standard (then with a second label *Beauty/Joy* or *Sadness* to keep consistency). However, *Confusion*, *Boredom* and *Other* are not available in any sub-corpus.

Our final set consists of nine classes, i.e., (in order of frequency) *Beauty/Joy, Sadness, Uneasiness, Vitality/Energy, Suspense, Awe/Sublime, Humor, Annoyance,* and *Nostalgia*. In the following, we describe the labels and give further details on the aggregation process.

Annoyance (annoys me/angers me/felt frustrated): Annoyance implies feeling annoyed, frustrated or even angry while reading the line/stanza. We include the class *Anger* here, as this was found to be too strong in intensity.

Awe/Sublime (found it overwhelming/sense of greatness): *Awe/Sublime* implies being overwhelmed by the line/stanza, i.e., if one gets the impression of facing something sublime

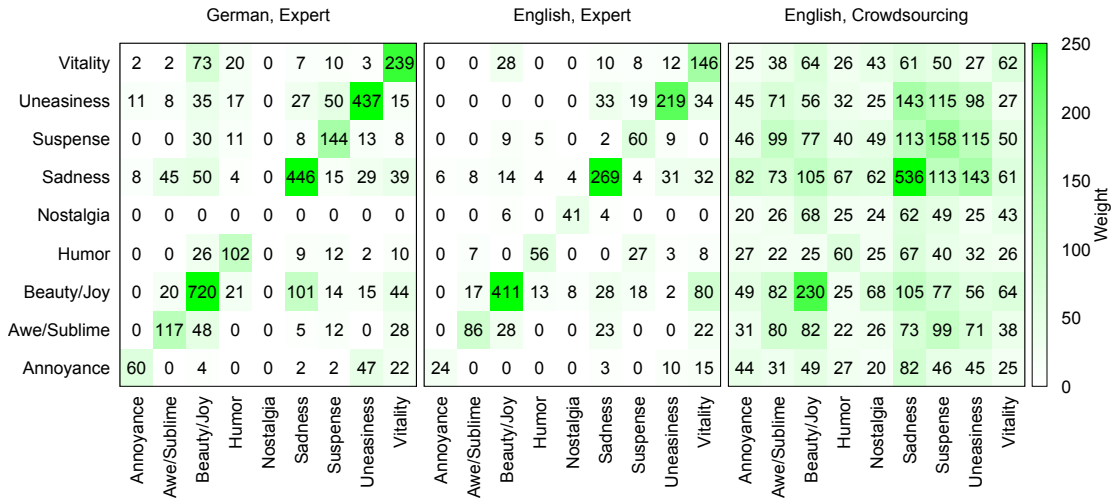

Figure 2: Emotion cooccurrence matrices for the German and English expert annotation experiments and the English crowdsourcing experiment.

or if the line/stanza inspires one with awe (or that the expression itself is sublime). Such emotions are often associated with subjects like god, death, life, truth, etc. The term *Sublime* originated with Kant (2001) as one of the first aesthetic emotion terms. *Awe* is a more common English term.

Beauty/Joy (found it beautiful/pleasing/makes me happy/joyful): Kant (2001) already spoke of a "feeling of beauty", and it should be noted that it is not a 'merely pleasing emotion'. Therefore, in our pilot annotations, *Beauty* and *Joy* were separate labels. However, Schindler et al. (2017) found that items for *Beauty* and *Joy* load into the same factors. Furthermore, our pilot annotations revealed, while *Beauty* is the more dominant and frequent feeling, both labels regularly accompany each other, and they often get confused across annotators. Therefore, we add *Joy* to form an inclusive label *Beauty/Joy* that increases consistency.

Humor (found it funny/amusing): Implies feeling amused by the line/stanza or if it makes one laugh.

Nostalgia (makes me nostalgic): Nostalgia is defined as a sentimental longing for things, persons or situations in the past. It often carries both positive and negative feelings. However, since this label is quite infrequent, and not available in all subsets of the data, we annotated it with an additional *Beauty/Joy* or *Sadness* label to ensure annotation consistency.

Sadness (makes me sad/touches me): If the line/stanza makes one feel sad. It also includes a more general 'being touched / moved'.

Suspense (found it gripping/sparked my interest): Choose *Suspense* if the line/stanza keeps one in suspense (if it excites one or triggers one's curiosity). We removed *Anticipation* from the earlier *Suspense/Anticipation* label, as *Anticipation* appeared to us as being a more cognitive prediction whereas Suspense is a far more straightforward emotion item.

Uneasiness (found it ugly/unsettling/disturbing / frightening/distasteful): This label covers situations when one feels discomfort, when the line/stanza feels distasteful/ugly, unsettling/disturbing or frightens one. The labels *Ugliness* and *Disgust* were conflated into *Uneasiness*, as both are seldom felt in poetry (being inadequate/too strong/high in arousal), and typically lead to *Uneasiness*.

Vitality/Energy (found it invigorating/spurs me on/inspires me): This label is meant for a line/stanza that has an inciting, encouraging effect (if it conveys a feeling of movement, energy and vitality which animates to action). Other terms: *Animated*, *Inspiration*, *Stimulation* and *Activation*.[7]

4.3. Agreement

Table 3 shows the Cohen's κ agreement scores among our two expert annotators for each emotion category e as follows. We assign each instance (a line in a poem) a binary label indicating whether or not the annotator has annotated the emotion category e in question. From this, we obtain vectors v_i^e, for annotators $i = 0, 1$, where each entry of v_i^e holds the binary value for the corresponding line. We then apply the κ statistics to the two binary vectors v_i^e. Additionally to averaged κ, we report micro-F1 values in Table 4 between the multi-label annotations of both expert annotators as well as the micro-F1 score of a random baseline as well as of the majority emotion baseline (which labels each line as *Beauty/Joy*).

We find that Cohen κ agreement ranges from .84 for *Uneasiness* in the English data, .81 for *Humor* and *Nostalgia*, down to German *Suspense* (.65), *Awe/Sublime* (.61) and *Vitality/Energy* for both languages (.50 English, .63 German). Both annotators have a similar emotion frequency profile, where the ranking is almost identical, especially for German. However, for English, Annotator 2 annotates more *Vitality/Energy* than *Uneasiness*. Figure 2 shows the confusion matrices of labels between annotators as heatmaps. Notably, *Beauty/Joy* and *Sadness* are confused across annotators more often than other labels. This is topical for poetry, and therefore not surprising: One might argue that the beauty of beings and situations is only beautiful because it is not enduring and therefore not to divorce from the sadness of the vanishing of beauty (Benjamin, 2016). We also find considerable confusion of *Sadness* with *Awe/Sublime* and

[7] Activation appears stable across cultures (Jackson et al., 2019)

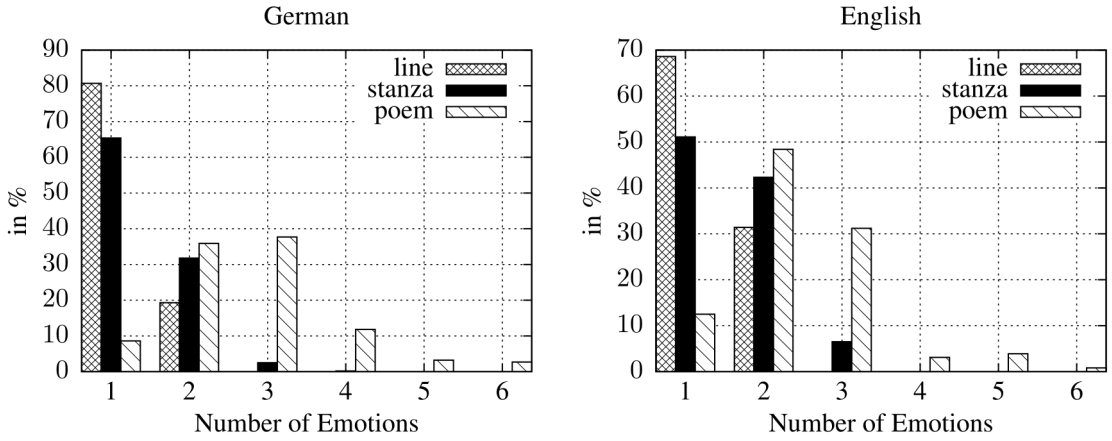

Figure 3: Distribution of number of distinct emotion labels per logical document level in the expert-based annotation. No whole poem has more than 6 emotions. No stanza has more than 4 emotions.

Vitality/Energy, while the latter is also regularly confused with *Beauty/Joy*.

Furthermore, as shown in Figure 3, we find that no single poem aggregates to more than six emotion labels, while no stanza aggregates to more than four emotion labels. However, most lines and stanzas prefer one or two labels. German poems seem more emotionally diverse where more poems have three labels than two labels, while the majority of English poems have only two labels. This is however attributable to the generally shorter English texts.

5. Crowdsourcing Annotation

After concluding the expert annotation, we performed a focused crowdsourcing experiment, based on the final label set and items as they are listed in Table 5 and Section 4.2. With this experiment, we aim to understand whether it is possible to collect reliable judgements for aesthetic perception of poetry from a crowdsourcing platform. A second goal is to see whether we can replicate the expensive expert annotations with less costly crowd annotations.

We opted for a maximally simple annotation environment, where we asked participants to annotate English 4-line stanzas with self-perceived reader emotions. We choose English due to the higher availability of English language annotators on crowdsourcing platforms. Each annotator rates each stanza independently of surrounding context.

5.1. Data and Setup

For consistency and to simplify the task for the annotators, we opt for a trade-off between completeness and granularity of the annotation. Specifically, we subselect stanzas composed of four verses from the corpus of 64 hand selected English poems. The resulting selection of 59 stanzas is uploaded to Figure Eight[8] for annotation.

The annotators are asked to answer the following questions for each instance.

Question 1 (single-choice): Read the following stanza and decide for yourself which emotions it evokes.

Question 2 (multiple-choice): Which additional emotions does the stanza evoke?

[8] https://www.figure-eight.com/

The answers to both questions correspond to the emotion labels we defined to use in our annotation, as described in Section 4.2. We add an additional answer choice "None" to Question 2 to allow annotators to say that a stanza does not evoke any additional emotions.

Each instance is annotated by ten people. We restrict the task geographically to the United Kingdom and Ireland and set the parameters on Figure Eight to only have the highest quality annotators join the task. We pay €0.09 per instance. The final cost of the crowdsourcing experiment is €74.

5.2. Results

In the following, we determine the best aggregation strategy regarding the 10 annotators with bootstrap resampling. For instance, one could assign the label of a specific emotion to an instance if just one annotators picks it, or one could assign the label only if all annotators agree on this emotion. To evaluate this, we repeatedly pick two sets of 5 annotators each out of the 10 annotators for each of the 59 stanzas, 1000 times overall (i.e., 1000×59 times, bootstrap resampling). For each of these repetitions, we compare the agreement of these two groups of 5 annotators. Each group gets assigned with an adjudicated emotion which is accepted if at least one annotator picks it, at least two annotators pick it, etc. up to all five pick it.

We show the results in Table 5. The κ scores show the average agreement between the two groups of five annotators, when the adjudicated class is picked based on the particular threshold of annotators with the same label choice. We see that some emotions tend to have higher agreement scores than others, namely *Annoyance* (.66), *Sadness* (up to .52), and *Awe/Sublime*, *Beauty/Joy*, *Humor* (all .46). The maximum agreement is reached mostly with a threshold of 2 (4 times) or 3 (3 times).

We further show in the same table the average numbers of labels from each strategy. Obviously, a lower threshold leads to higher numbers (corresponding to a disjunction of annotations for each emotion). The drop in label counts is comparably drastic, with on average 18 labels per class. Overall, the best average κ agreement (.32) is less than half of what we saw for the expert annotators (roughly .70).

	κ					Counts				
Threshold	≥ 1	≥ 2	≥ 3	≥ 4	≥ 5	≥ 1	≥ 2	≥ 3	≥ 4	≥ 5
Beauty / Joy	.21	.41	.46	.28	–	34.58	15.98	7.51	3.23	1.43
Sadness	.43	.47	.52	.02	−.04	43.34	28.99	17.77	9.52	2.82
Uneasiness	.18	.25	.08	−.01	–	36.47	16.33	5.49	1.54	1.04
Vitality	.15	.26	.19	–	–	25.62	7.34	2.02	1.05	1.00
Awe / Sublime	.31	.17	.37	.46	–	29.8	11.36	3.4	1.31	1.00
Suspense	.11	.29	.21	.26	–	39.12	17.8	6.54	1.97	1.04
Humor	.19	.46	.39	≈0	–	19.26	5.36	2.1	1.22	1.07
Nostalgia	.23	.01	−.02	–	–	30.52	10.16	1.95	1.00	1.00
Annoyance	.01	.07	.66	0	–	26.54	6.17	1.35	1.00	1.00
Average	0.20	0.27	0.32	0.14	−0.04	31.69	13.28	5.35	2.43	1.27

Table 5: Results obtained via boostrapping for annotation aggregation. The row *Threshold* shows how many people within a group of five annotators should agree on a particular emotion. The column labeled *Counts* shows the average number of times certain emotion was assigned to a stanza given the threshold. Cells with '–' mean that neither of two groups satisfied the threshold.

Crowds especially disagree on many more intricate emotion labels (Uneasiness, Vitality/Energy, Nostalgia, Suspense).
We visualize how often two emotions are used to label an instance in a confusion table in Figure 2. Sadness is used most often to annotate a stanza, and it is often confused with Suspense, Uneasiness, and Nostalgia. Further, Beauty/Joy partially overlaps with Awe/Sublime, Nostalgia, and Sadness.
On average, each crowd annotator uses two emotion labels per stanza (56% of cases); only in 36% of the cases the annotators use one label, and in 6% and 1% of the cases three and four labels, respectively. This contrasts with the expert annotators, who use one label in about 70% of the cases and two labels in 30% of the cases for the same 59 four-liners.
Concerning frequency distribution for emotion labels, both experts and crowds name Sadness and Beauty/Joy as the most frequent emotions (for the 'best' threshold of 3) and Nostalgia as one of the least frequent emotions. The Spearman rank correlation between experts and crowds is about 0.55 with respect to the label frequency distribution, indicating that crowds could replace experts to a moderate degree when it comes to extracting, e.g., emotion distributions for an author or time period. Now, we further compare crowds and experts in terms of whether crowds could replicate expert annotations also on a finer stanza level (rather than only on a distributional level).

5.3. Comparing Experts with Crowds

To gauge the quality of the crowd annotations in comparison with our experts, we calculate agreement on the emotions between experts and an increasing group size from the crowd. For each stanza instance s, we pick N crowd workers, where $N \in \{4, 6, 8, 10\}$, then pick their majority emotion for s, and additionally pick their second ranked majority emotion if at least $\frac{N}{2} - 1$ workers have chosen it.[9] For the experts, we aggregate their emotion labels on stanza level, then perform the same strategy for selection of emotion labels. Thus, for s, both crowds and experts have 1 or 2 emotions. For each emotion, we then compute Cohen's κ as before. Note that, compared to our previous experiments in Section 5.2 with a threshold, each stanza now receives an emotion annotation (exactly one or two emotion labels), both by the experts and the crowd-workers.

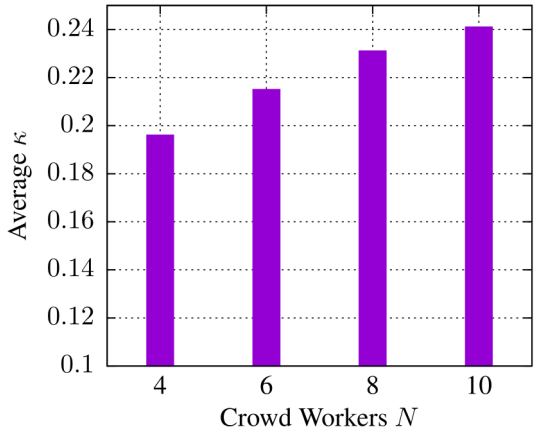

Figure 4: Agreement between experts and crowds as a function of the number N of crowd workers.

In Figure 4, we plot agreement between experts and crowds on stanza level as we vary the number N of crowd workers involved. On average, there is roughly a steady linear increase in agreement as N grows, which may indicate that $N = 20$ or $N = 30$ would still lead to better agreement.
Concerning individual emotions, *Nostalgia* is the emotion with the least agreement, as opposed to *Sadness* (in our sample of 59 four-liners): the agreement for this emotion grows from .47 κ with $N = 4$ to .65 κ with $N = 10$. *Sadness* is also the most frequent emotion, both according to experts and crowds. Other emotions for which a reasonable agreement is achieved are *Annoyance, Awe/Sublime, Beauty/Joy, Humor* ($\kappa > 0.2$). Emotions with little agreement are *Vitality/Energy, Uneasiness, Suspense, Nostalgia* ($\kappa < 0.2$).
By and large, we note from Figure 2 that expert annotation is more restrictive, with experts agreeing more often on particular emotion labels (seen in the darker diagonal).

[9] For workers, we additionally require that an emotion has been chosen by at least 2 workers.

The results of the crowdsourcing experiment, on the other hand, are a mixed bag as evidenced by a much sparser distribution of emotion labels. However, we note that these differences can be caused by 1) the disparate training procedure for the experts and crowds, and 2) the lack of opportunities for close supervision and on-going training of the crowds, as opposed to the in-house expert annotators.

In general, however, we find that substituting experts with crowds is possible to a certain degree. Even though the crowds' labels look inconsistent at first view, there appears to be a good signal in their *aggregated* annotations, helping to approximate expert annotations to a certain degree. The average κ agreement (with the experts) we get from $N = 10$ crowd workers (0.24) is still considerably below the agreement among the experts (0.70).

6. Modeling

To estimate the difficulty of automatic classification of our data set, we perform multi-label[10] document classification (of stanzas) with BERT (Devlin et al., 2019). For this experiment we aggregate all labels for a stanza and sort them by frequency, both for the gold standard and the raw expert annotations. As can be seen in Figure 3, a stanza bears a minimum of one and a maximum of four emotions. Unfortunately, the label *Nostalgia* is only available 16 times in the German data (the gold standard) as a second label (as discussed in Section 4.2). None of our models was able to learn this label for German. Therefore we omit it, leaving us with eight proper labels.

We use the code and the pre-trained BERT models of FARM,[11] provided by `deepset.ai`. We test the multilingual-uncased model (MULTILING), the german-base-cased model (BASE),[12] the german-dbmdz-uncased model (DBMDZ),[13] and we tune the BASE model on 80k stanzas of the German Poetry Corpus DLK (Haider and Eger, 2019) for 2 epochs, both on token (masked words) and sequence (next line) prediction ($\text{BASE}_{\text{TUNED}}$).

We split the randomized German dataset so that each label is at least 10 times in the validation set (63 instances, 113 labels), and at least 10 times in the test set (56 instances, 108 labels) and leave the rest for training (617 instances, 946 labels).[14] We train BERT for 10 epochs (with a batch size of 8), optimize with entropy loss, and report F1-micro on the test set. See Table 6 for the results.

We find that the multilingual model cannot handle infrequent categories, i.e., *Awe/Sublime, Suspense* and *Humor*. However, increasing the dataset with English data improves the results, suggesting that the classification would largely benefit from more annotated data. The best model overall is DBMDZ (.520), showing a balanced response on both validation and test set. See Table 7 for a breakdown of all emotions as predicted by the this model. Precision is mostly higher than recall. The labels *Awe/Sublime*, *Suspense* and *Humor* are harder to predict than the other labels.

The BASE and $\text{BASE}_{\text{TUNED}}$ models perform slightly worse than DBMDZ. The effect of tuning of the BASE model is questionable, probably because of the restricted vocabulary (30k). We found that tuning on poetry does not show obvious improvements. Lastly, we find that models that were trained on lines (instead of stanzas) do not achieve the same F1 (~.42 for the German models).

	German		Multiling.	
Model	dev	test	dev	test
Majority	.212	.167	.176	.150
MULTILING	.409	.341	.461	.384
BASE	.500	.439	–	–
$\text{BASE}_{\text{TUNED}}$	.477	.514	–	–
DBMDZ	.520	**.520**	–	–

Table 6: BERT-based multi-label classification on stanza-level.

Label	Precision	Recall	F1	Support
Beauty/Joy	0.5000	0.5556	0.5263	18
Sadness	0.5833	0.4667	0.5185	15
Uneasiness	0.6923	0.5625	0.6207	16
Vitality/Energy	1.0000	0.5333	0.6957	15
Annoyance	1.0000	0.4000	0.5714	10
Awe/Sublime	0.5000	0.3000	0.3750	10
Suspense	0.6667	0.1667	0.2667	12
Humor	1.0000	0.2500	0.4000	12
micro avg	0.6667	0.4259	0.5198	108
macro avg	0.7428	0.4043	0.4968	108
weighted avg	0.7299	0.4259	0.5100	108
samples avg	0.5804	0.4464	0.4827	108

Table 7: Recall and precision scores of the best model (dbmdz) for each emotion on the test set. 'Support': number of instances with this label.

7. Concluding Remarks

In this paper, we presented a dataset of German and English poetry annotated with reader response to reading poetry. We argued that basic emotions as proposed by psychologists (such as Ekman and Plutchik) that are often used in emotion analysis from text are of little use for the annotation of poetry reception. We instead conceptualized aesthetic emotion labels and showed that a closely supervised annotation task results in substantial agreement—in terms of κ score—on the final dataset.

The task of collecting reader-perceived emotion response to poetry in a crowdsourcing setting is not straightforward. In contrast to expert annotators, who were closely supervised and reflected upon the task, the annotators on crowdsourcing platforms are difficult to control and may lack necessary background knowledge to perform the task at hand. However, using a larger number of crowd annotators may lead to finding an aggregation strategy with a better trade-off between quality and quantity of adjudicated labels. For future work, we thus propose to repeat the experiment with larger

[10] We found that single-label classification had only marginally better performance, even though the task is simpler.

[11] `https://github.com/deepset-ai/FARM`

[12] There was no uncased model available.

[13] `https://github.com/dbmdz` a model by the Bavarian state library that was also trained on literature.

[14] We do the same for the English data (at least 5 labels) and add the stanzas to the respective sets.

number of crowdworkers, and develop an improved training strategy that would suit the crowdsourcing environment. The dataset presented in this paper can be of use for different application scenarios, including multi-label emotion classification, style-conditioned poetry generation, investigating the influence of rhythm/prosodic features on emotion, or analysis of authors, genres and diachronic variation (e.g., how emotions are represented differently in certain periods). Further, though our modeling experiments are still rudimentary, we propose that this data set can be used to investigate the intra-poem relations either through multi-task learning (Schulz et al., 2018) and/or with the help of hierarchical sequence classification approaches.

Acknowledgements

A special thanks goes to Gesine Fuhrmann, who created the guidelines and tirelessly documented the annotation progress. Also thanks to Annika Palm and Debby Trzeciak who annotated and gave lively feedback. For help with the conceptualization of labels we thank Ines Schindler. This research has been partially conducted within the CRETA center (`http://www.creta.uni-stuttgart.de/`) which is funded by the German Ministry for Education and Research (BMBF) and partially funded by the German Research Council (DFG), projects SEAT (Structured Multi-Domain Emotion Analysis from Text, KL 2869/1-1). This work has also been supported by the German Research Foundation as part of the Research Training Group Adaptive Preparation of Information from Heterogeneous Sources (AIPHES) at the Technische Universität Darmstadt under grant No. GRK 1994/1.

Appendix

We illustrate two examples of our German gold standard annotation, a poem each by Friedrich Hölderlin and Georg Trakl, and an English poem by Walt Whitman. Hölderlin's text stands out, because the mood changes starkly from the first stanza to the second, from *Beauty/Joy* to *Sadness*. Trakl's text is a bit more complex with bits of *Nostalgia* and, most importantly, a mixture of *Uneasiness* with *Awe/Sublime*. Whitman's poem is an example of *Vitality* and its mixing with *Sadness*. The English annotation was unified by us for space constraints. For the full annotation please see `https://github.com/tnhaider/poetry-emotion/`

Friedrich Hölderlin: Hälfte des Lebens (1804)

Mit gelben Birnen hänget	[Beauty/Joy]
Und voll mit wilden Rosen	[Beauty/Joy]
Das Land in den See,	[Beauty/Joy]
Ihr holden Schwäne,	[Beauty/Joy]
Und trunken von Küssen	[Beauty/Joy]
Tunkt ihr das Haupt	[Beauty/Joy]
Ins heilignüchterne Wasser.	[Beauty/Joy]
Weh mir, wo nehm' ich, wenn	[Sadness]
Es Winter ist, die Blumen, und wo	[Sadness]
Den Sonnenschein,	[Sadness]
Und Schatten der Erde?	[Sadness]
Die Mauern stehn	[Sadness]
Sprachlos und kalt, im Winde	[Sadness]
Klirren die Fahnen.	[Sadness]

Georg Trakl: In den Nachmittag geflüstert (1912)

Sonne, herbstlich dünn und zag,	[Beauty/Joy] [Nostalgia]
Und das Obst fällt von den Bäumen.	[Beauty/Joy] [Nostalgia]
Stille wohnt in blauen Räumen	[Beauty/Joy]
Einen langen Nachmittag.	[Beauty/Joy]
Sterbeklänge von Metall;	[Sadness] [Uneasiness]
Und ein weißes Tier bricht nieder.	[Sadness] [Uneasiness]
Brauner Mädchen rauhe Lieder	[Sadness] [Nostalgia]
Sind verweht im Blätterfall.	[Sadness] [Nostalgia]
Stirne Gottes Farben träumt,	[Uneasiness] [Awe/Sublime]
Spürt des Wahnsinns sanfte Flügel.	[Uneasiness] [Awe/Sublime]
Schatten drehen sich am Hügel	[Uneasiness] [Awe/Sublime]
Von Verwesung schwarz umsäumt.	[Uneasiness] [Awe/Sublime]
Dämmerung voll Ruh und Wein;	[Beauty/Joy]
Traurige Guitarren rinnen.	[Beauty/Joy]
Und zur milden Lampe drinnen	[Beauty/Joy]
Kehrst du wie im Traume ein.	[Beauty/Joy]

Walt Whitman: O Captain! My Captain! (1865)

O Captain! my Captain! our fearful trip is done,	[Beauty/Joy]
The ship has weather'd every rack, the prize we sought is won,	[Beauty/Joy]
The port is near, the bells I hear, the people all exulting,	[Beauty/Joy]
While follow eyes the steady keel, the vessel grim and daring;	[Beauty/Joy]
But O heart! heart! heart!	[Sadness]
O the bleeding drops of red,	[Sadness]
Where on the deck my Captain lies,	[Sadness]
Fallen cold and dead.	[Sadness]
O Captain! my Captain! rise up and hear the bells;	[Vitality]
Rise up – for you the flag is flung – for you the bugle trills,	[Vitality]
For you bouquets and ribbon'd wreaths –	
for you the shores a-crowding,	[Vitality]
For you they call, the swaying mass, their eager faces turning;	[Vitality]
Here Captain! dear father!	[Vitality]
This arm beneath your head!	[Vitality]
It is some dream that on the deck,	[Sadness]
You've fallen cold and dead.	[Sadness]
My Captain does not answer, his lips are pale and still,	[Sadness]
My father does not feel my arm, he has no pulse nor will,	[Sadness]
The ship is anchor'd safe and sound, its voyage closed and done,	[Vitality] [Sadn.]
From fearful trip the victor ship comes in with object won;	[Vitality] [Sadn.]
Exult O shores, and ring O bells!	[Vitality] [Sadn.]
But I with mournful tread,	[Sadness]
Walk the deck my Captain lies,	[Sadness]
Fallen cold and dead.	[Sadness]

8. Bibliographical References

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

Agirrezabal, M., Alegria, I., and Hulden, M. (2016). Machine learning for metrical analysis of English poetry. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 772–781, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Alm, C. O., Roth, D., and 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*, pages 579–586, Vancouver, British Columbia, Canada, October. Association for Computational Linguistics.

Alsharif, O., Alshamaa, D., and Ghneim, N. (2013). Emotion classification in arabic poetry using machine learning. *International Journal of Computer Applications*, 65(16).

Aman, S. and Szpakowicz, S. (2007). Identifying expressions of emotion in text. In Václav Matoušek et al., editors, *Text, Speech and Dialogue*, pages 196–205, Berlin, Heidelberg. Springer Berlin Heidelberg.

Barros, L., Rodriguez, P., and Ortigosa, A. (2013). Automatic classification of literature pieces by emotion detection: A study on quevedo's poetry. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 141–146. IEEE.

Baumann, T., Hussein, H., and Meyer-Sickendiek, B. (2018). Style detection for free verse poetry from text and speech. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1929–1940.

Benjamin, W. (2016). *Goethes Wahlverwandtschaften*. BoD–Books on Demand.

Bostan, L.-A.-M. and Klinger, R. (2018). An analysis of annotated corpora for emotion classification in text. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2104–2119, Santa Fe, New Mexico, USA, aug. Association for Computational Linguistics.

Buechel, S. and 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*, pages 578–585, Valencia, Spain, April. Association for Computational Linguistics.

Buechel, S. and Hahn, U. (2017b). Readers vs. writers vs. texts: Coping with different perspectives of text understanding in emotion annotation. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 1–12, Valencia, Spain, April. Association for Computational Linguistics.

Cevher, D., Zepf, S., and Klinger, R. (2019). Towards multimodal emotion recognition in german speech events in cars using transfer learning. In *Conference on Natural Language Processing (KONVENS)*.

Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2015). Gated feedback recurrent neural networks. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 2067–2075. JMLR.org.

Dankers, V., Rei, M., Lewis, M., and Shutova, E. (2019). Modelling the interplay of metaphor and emotion through multitask learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2218–2229, Hong Kong, China, November. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and 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)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

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

Estes, A. and Hench, C. (2016). Supervised machine learning for hybrid meter. In *Proceedings of the Fifth Workshop on Computational Linguistics for Literature*, pages 1–8.

Esuli, A. and Sebastiani, F. (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06*, pages 417–422.

Gopidi, A. and Alam, A. (2019). Computational analysis of the historical changes in poetry and prose. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 14–22.

Greene, E., Bodrumlu, T., and Knight, K. (2010). Automatic analysis of rhythmic poetry with applications to generation and translation. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 524–533.

Haider, T. and Eger, S. (2019). Semantic change and emerging tropes in a large corpus of new high german poetry. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 216–222.

Haider, T. and Kuhn, J. (2018). Supervised rhyme detection with siamese recurrent networks. In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature at COLING 2018, Santa Fe New Mexico*, pages 81–86.

Haider, T., Trzeciak, D., and Kentner, G. (2020). Speech rhythm and syntax in poetry and prose. In *In Proceedings of the International Digital Humanities Conference DH2020, Ottawa*. accepted.

Haider, T. (2019). Diachronic topics in new high german poetry. In *In Proceedings of the International Digital Humanities Conference DH2019, Utrecht*.

Herbelot, A. (2014). The semantics of poetry: a distributional reading. *Digital Scholarship in the Humanities*, 30(4):516–531.

Hou, Y. and Frank, A. (2015). Analyzing sentiment in classical Chinese poetry. In *Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, pages 15–24, Beijing, China, July. Association for Computational Linguistics.

Jackson, J. C., Watts, J., Henry, T. R., List, J.-M., Forkel, R., Mucha, P. J., Greenhill, S. J., Gray, R. D., and Lindquist, K. A. (2019). Emotion semantics show both cultural variation and universal structure. *Science*, 366(6472):1517–1522.

Johnson-Laird, P. N. and Oatley, K., (2016). *Handbook of emotions*, chapter Emotions in Music, Literature, and Film, pages 82–97. Guilford Publications.

Kant, I. (2001). *Critique of the Power of Judgment*. (P.Guyer & E. Matthews, Trans.). Cambridge, England: Cambridge University Press (Original work published 1790).

Kao, J. T. and Jurafsky, D. (2015). A computational analy-

sis of poetic style. *LiLT (Linguistic Issues in Language Technology)*, 12.

Kaplan, D. M. and Blei, D. M. (2007). A computational approach to style in american poetry. In *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, pages 553–558. IEEE.

Kesarwani, V., Inkpen, D., Szpakowicz, S., and Tanasescu, C. (2017). Metaphor detection in a poetry corpus. In *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 1–9.

Kim, E. and Klinger, R. (2018). Who feels what and why? annotation of a literature corpus with semantic roles of emotions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1345–1359, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Klie, J.-C., Bugert, M., Boullosa, B., de Castilho, R. E., and Gurevych, I. (2018). The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9.

Klinger, R., De Clercq, O., Mohammad, S., and Balahur, A. (2018). IEST: WASSA-2018 implicit emotions shared task. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 31–42, Brussels, Belgium, October. Association for Computational Linguistics.

Knoop, C. A., Wagner, V., Jacobsen, T., and Menninghaus, W. (2016). Mapping the aesthetic space of literature "from below". *Poetics*, 56:35–49.

Krishna, A., Sharma, V. D., Santra, B., Chakraborty, A., Satuluri, P., and Goyal, P. (2019). Poetry to prose conversion in sanskrit as a linearisation task: A case for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1160–1166.

Liu, C., Osama, M., and De Andrade, A. (2019). DENS: A dataset for multi-class emotion analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6294–6299, Hong Kong, China, November. Association for Computational Linguistics.

Menninghaus, W., Wagner, V., Hanich, J., Wassiliwizky, E., Kuehnast, M., and Jacobsen, T. (2015). Towards a psychological construct of being moved. *PloS one*, 10(6):e0128451.

Menninghaus, W., Wagner, V., Wassiliwizky, E., Jacobsen, T., and Knoop, C. A. (2017). The emotional and aesthetic powers of parallelistic diction. *Poetics*, 63:47–59.

Menninghaus, W., Wagner, V., Wassiliwizky, E., Schindler, I., Hanich, J., Jacobsen, T., and Koelsch, S. (2019). What are aesthetic emotions? *Psychological review*, 126(2):171.

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

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

Mohammad, S. M. (2016). Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In *Emotion measurement*, pages 201–237. Elsevier.

Mohanty, G., Mishra, P., and Mamidi, R. (2018). Kabithaa: An annotated corpus of odia poems with sentiment polarity information. In Girish Nath Jha, et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, may. European Language Resources Association (ELRA).

Plutchik, R. (1991). *The Emotions*. University Press of America.

Reddy, S. and Knight, K. (2011). Unsupervised discovery of rhyme schemes. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 77–82.

Reinig, I. and Rehbein, I. (2019). Metaphor detection for german poetry. Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019).

Roberts, P. (2000). *How Poetry Works*. Penguin UK.

Ruiz, P., Cantón, C. M., Poibeau, T., and González-Blanco, E. (2017). Enjambment detection in a large diachronic corpus of spanish sonnets. In *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 27–32.

Scherer, K. R. and Wallbott, H. G. (1994). Evidence for universality and cultural variation of differential emotion response patterning. *Journal of personality and social psychology*, 66(2):310.

Schindler, I., Hosoya, G., Menninghaus, W., Beermann, U., Wagner, V., Eid, M., and Scherer, K. R. (2017). Measuring aesthetic emotions: A review of the literature and a new assessment tool. *PloS one*, 12(6):e0178899.

Schuff, H., Barnes, J., Mohme, J., Padó, S., and 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*, pages 13–23, Copenhagen, Denmark, September. Association for Computational Linguistics.

Schulz, C., Eger, S., Daxenberger, J., Kahse, T., and Gurevych, I. (2018). Multi-task learning for argumentation mining in low-resource settings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 35–41, New Orleans, Louisiana, June. Association for Computational Linguistics.

Sreeja, S. P. and Mahalakshmi, G. S. (2019). Perc-an emotion recognition corpus for cognitive poems. In *2019 International Conference on Communication and Signal Processing (ICCSP)*, pages 0200–0207, April.

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

Strapparava, C. and Valitutti, A. (2004). WordNet affect: an affective extension of WordNet. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. European Language Resources Association (ELRA).

Underwood, T. and Sellers, J. (2012). The emergence of literary diction. *The Journal of Digital Humanities, 1(2)*, pages http://journalofdigitalhumanities.org/1–2/theemergence–of–literary–diction–by–ted–underwoodand–jordan–sellers/.

Voigt, R. and Jurafsky, D. (2013). Tradition and modernity in 20th century chinese poetry. In *Proceedings of the Workshop on Computational Linguistics for Literature*, pages 17–22.

Wassiliwizky, E., Jacobsen, T., Heinrich, J., Schneiderbauer, M., and Menninghaus, W. (2017). Tears falling on goosebumps: Co-occurrence of emotional lacrimation and emotional piloerection indicates a psychophysiological climax in emotional arousal. *Frontiers in Psychology*, 8:41.

Yang, C., Lin, K. H., and Chen, H. (2009). Writer meets reader: Emotion analysis of social media from both the writer's and reader's perspectives. In *2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, volume 1, pages 287–290, Sep.