Emotion Annotation: Rethinking Emotion Categorization

Emily Öhman[1][0000−0003−1363−7361]

University of Helsinki, Finland
Tampere University, Finland
emily.ohman@helsinki.fi

Abstract. One of the biggest hurdles for the utilization of machine learning in interdisciplinary projects is the need for annotated training data which is costly to create. Emotion annotation is a notoriously difficult task, and the current annotation schemes which are based on psychological theories of human interaction are not always the most conducive for the creation of reliable emotion annotations, nor are they optimal for annotating emotions in the modality of text. This paper discusses the theory, history, and challenges of emotion annotation, and proposes improvements for emotion annotation tasks based on both theory and case studies. These improvements focus on rethinking the categorization of emotions and the overlap and disjointedness of emotion categories.

Keywords: Emotion Annotation, Textual Expressions of Emotions, Theories of Emotion.

1 Introduction

Sentiment analysis has progressed along with general developments in Natural Language Processing (NLP) and machine learning in the past two decades [35], with more and more advanced models and algorithms aiding in the detection of sentiments and emotions in text. Most of these machine learning models are supervised, which means that they require large manually annotated datasets. Annotation tasks range in difficulty based on the data being annotated, the annotation scheme, and the training received by the annotators. Emotion annotation is notoriously difficult, a notion shared by many emotion researchers particularly in the field of NLP (see e.g. [4, 7, 29, 34, 48, 50]).

To the best of my knowledge, most emotion detection papers use some variation of Ekman's [19–21] six core emotions (*anger, disgust, fear, happiness, sadness, surprise*), or Plutchik's [43] eight core emotions (*anger, anticipation, disgust, fear, joy, sadness, surprise, trust*). These are based on well-known psychological theories that have been researched extensively for decades and were therefore natural starting points for computational emotion detection. However, whether these categories are the best at describing human emotions is a question still debated in the emotion community [45] and recently whether these emotions

correspond to human emotions as expressed in text has been asked in several sentiment analysis and NLP papers (see e.g. [16, 40]).

One of the reasons that make emotion annotation difficult is the modality, i.e. text versus speech versus video and so forth. In typical human interactions, emotions are expressed through multiple modalities simultaneously [15], but in annotation tasks, the focus is usually on one single modality, most commonly text. Emotions are expressed in text through various means that are limited by linguistic, cultural, and social constraints. In contrast, the most commonly used emotion annotation schemes are based on psychological theories that are in turn based on human interactions, not text. Therefore annotating outside the originally intended modality or environment makes the emotion annotation task harder.

In the next section the theory of emotions are discussed from a cultural, linguistic, and computational aspect. Previous work related to emotion annotation and annotation theory is presented in section 3. Finally, the future of emotion annotation is discussed in terms of alternative annotation schemes and other steps to be taken in order to improve the annotation process.

2 Emotion Theories in NLP

The one thing all emotion analysis studies, both quantitative and qualitative, have in common, is that they try to analyze human feelings. These feelings can be defined differently, using different psychological, or even physiological, theories of emotion and labeled as affect, feeling, emotion, sentiment, or opinion ([5, 40]). What exactly these terms mean is interpreted differently in different fields and sometimes even between researchers in the same field. As there is no consensus on what human emotions are [45], the first step in any sentiment analysis or emotion annotation task is to define the terms and the theory that is being relied on in that specific study.

As recent survey studies show, most modern research on emotions, particularly in NLP, is at least to some extent based on the work of Ekman [19]. This includes the work of Robert Plutchik, especially his Wheel of Emotions [43], as well as SenticNet [11] (see e.g. [10, 27, 30, 32, 42]).

For SenticNet [9, 10, 12] Plutchik's wheel was reworked to show the change in emotional intensity on a Gaussian curve. The idea is that this would fit better with human-computer interaction and studies in affective computing. The emotions are further categorized into *pleasantness, attention, sensitivity*, and *aptitude*. Although SenticNet is a well-known model in the field of NLP (based on citation counts and sources), it has not become as prevalent as one would assume.

Most recently, the work of Keltner and Cowen [14, 15, 25, 26] have tried to tackle the categorization of emotions in a number of studies and have come up with an emotion categorization consisting of 27 distinct emotions by studying emotion responses to a number of different stimuli such as videos, music, facial expressions, speech prosody and even nonverbal vocalization [15, 16]. This is the

emotion annotation scheme partly relied on in GoEmotions [16]. But although this categorization has its benefits, it too suffers from some of the same issues other categorization schemes suffer from, namely that it is not designed for emotion detection in **text**.

There are some significant differences in the surface realization of emotion in different languages. These differences do not mean that certain emotions are not present in that language, but the fact that the emotion words available in different languages are so different makes exploring emotions in text particularly difficult. The concept of a certain emotion might only exist in one language or some emotions might be separate categories in one language and not be differentiated at all in another.

The same is true for text type. Narrative texts in particular, such as novels or movie scripts and subtitles, have been shown to have lower annotator agreement scores than other text types [3, 46, 47]. In traditional literary analysis *mood* is often studied [37, 44]. This might be another alternative for other types of narrative texts.

The joint modeling of emotion classification and semantic role labeling has also been shown to be beneficial to emotion detection tasks [38]. Similarly, assigning appraisal dimensions to event descriptions improves the classification accuracies of discrete emotion categories [23].

3 Emotion Annotation and Classification

Supervised machine learning tasks rely on annotated data, but the annotation of datasets can be very costly and time consuming [4, 17]. Crowd-sourcing can often be a cheaper alternative to hiring expert annotators, and has been used successfully in several projects to create different types of annotated datasets, including sentiment and emotion annotated ones [49, 22, 31, 32, 39]. One issue with using non-experts to solicit annotations is that there is a risk of the quality suffering. This risk can be mitigated by carefully controlling selection criteria [24], and being aware of typically difficult instances to annotate such as requests, the speaker's emotional state, neutral reporting of positive or negative facts and so forth [29].

Due to the subjectivity of most annotation tasks, however, humans do not always agree with each other on how to annotate. Whether a tweet, for example, expresses surprise or fear can be ambiguous, and heavily depends on the reader's interpretation. There is rarely one single reading that can be judged as being correct [13] as the annotation task is generally highly subjective even with careful annotator guidelines [29]. It should also be noted that emotions do not have distinct and clear boundaries that separate them and they often occur together with other emotions [32].

Table 1 shows the distribution of emotion combinations in the multilabel XED [41] dataset. The most common annotation is single-label, however, after that the combination of emotions becomes more intriguing. *Anger* and *disgust* are the most common pair, followed closely by all the possible pairs consisting of

anticipation, joy and *trust*. Statistically speaking it is not surprising to see *anger* and *anticipation* co-occur to such a high degree, but intuitively speaking this is something that warrants a closer look. It might be linked to the source data which is movie subtitles and therefore a true reflection of the emotions expressed by the subtitles and suggestive of an emotion akin to *suspense* or *nervousness*. The fact that *anticipation* co-occurs with all emotions makes training an algorithm to detect anticipation specifically, a difficult task.

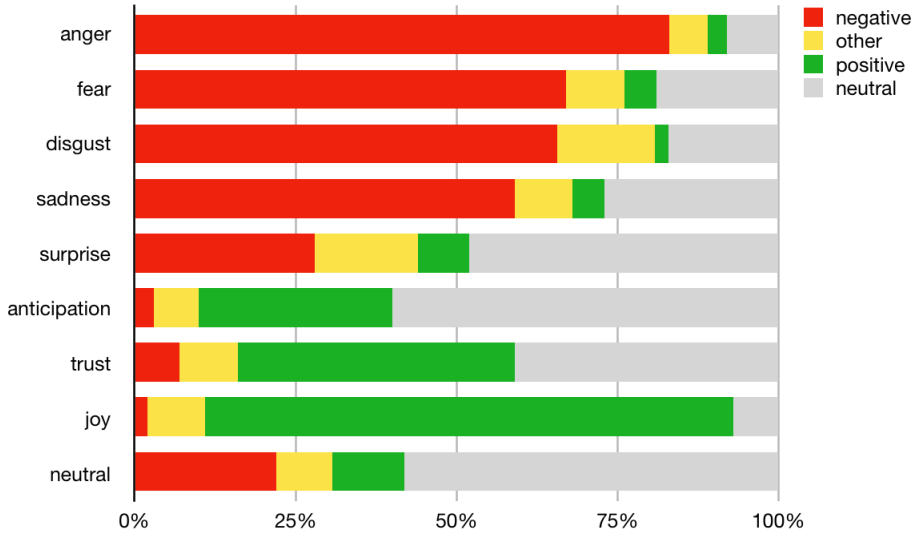
Table 1. Label combinations in the XED dataset

Number of unique label combinations: 147

Total # of combinations								
	anger	antic.	disg.	fear	joy	sad.	surpr.	trust
2393	■							
2028		■						
1721					■			
1617								■
1529				■				
1527						■		
1429							■	
1413			■					
407	■		■					
354					■			■
316		■			■			
200		■						■
196	■	■						
177						■	■	
139				■		■		
127	■			■				
114	■						■	
113			■			■		
110	■					■		
109		■			■			■
106		■					■	
105					■		■	

When a part of the XED dataset was re-annotated into *positive*, *negative*, *neutral*, and *other* by expert annotators at the University of Turku, the results for *anticipation* were by far the worst (see figure 1). Here the corresponding sentiment for what was originally annotated as *anticipation* seems to for the majority of cases be *neutral*, with *positive* a close second and *negative* and *other* only marginally present. This not only highlights the difficulty of emotion annotation, but also how particular emotions are much harder to categorize as well as generalize. When a category such as *suspense* is missing from the annotation scheme, there seems to be significant overflow into adjacent categories regardless of the typically perceived congruence or polarity of that category.

Fig. 1. Re-annotation by expert annotators. Image courtesy of Associate Professor Sampo Pyysalo of the University of Turku.



Annotation reliability is usually measured by calculating inter-annotator agreement, sometimes referred to as inter-rater correlation. These scores have not been that good for emotional labeling. For a sentence-level sentiment polarity annotation task (positive/negative), inter-annotator agreements stayed at around 57% with Krippendorf's α scores of $\alpha = 0.4219$, well below even tentative reliability ($\alpha = 0.666...$) [7]. In a word sense disambiguation task, κ values were around 0.3, indicating very low agreement [36].

Previous annotation tasks have shown that even with binary or ternary classification schemes, human annotators agree only about 70-80% of the time and the more categories there are, the harder it becomes for annotators to agree [6, 8, 33]. For example, when creating the DENS dataset [28], only 21% of the annotations had consensus between all annotators with 73.5% having to resort to majority agreement, and a further 5.5% could not be agreed upon and were left to expert annotators to be resolved. In [3] all annotators agreed upon only 8.36% of sentences, and for DENS, for the final dataset only 21% of the annotations had consensus between all annotators [28] — and this was after noisy annotations had already been removed.

Among the things that have been shown to influence inter-annotator agreement are: the domain of the data being annotated, the number of labels and categories in the annotation scheme, the training and guidelines of the annotators as well as the intensity of that training, how many annotators there are in total, for what purpose the annotations are, and of course the method used to calculate the inter-annotator agreement [6]. Some of these considerations relate to the mathematical aspect of calculating agreement, including the point about

the number of annotators. Interestingly, Bayerl et al. [6] found that increasing training improved agreement scores, when Mohammad [29, 32] found that minimal guidelines improved agreement scores because over-training annotators led to confusion and apprehension in judgment tasks.

Some emotions are also harder to detect and recognize. [16] show that the emotions of *admiration, approval, annoyance* and *gratitude* had the highest interrater correlations at around 0.6, and *grief, relief, pride, nervousness*, *embarrassment* had the lowest interrater correlations between 0-0.2, with a vast majority of emotions falling in the range of 0.3-0.5. Liu et al. [28] too note that they had difficulties with the categories of *disgust* and *surprise*. In their case, *disgust* was such a small category that it was discarded, and surprise was such a noisy category, that it too was discarded. Alm [2], who had the same observations regarding both *disgust* and *surprise*, speculates that this is because *surprise* is characterized in text by 'direct speech' and 'unexpected observations' that only indicate *surprise*[1] in the context of those surrounding sentences. On the other hand, she found that *fear*[2] was often marked by words directly associated with *fear*. Sentences that contained outright affect words were more likely to have high inter-annotator agreement [2].

Similarly, Demszky et al. [16] found that lexical items that are significantly associated with a particular emotion had also significantly higher interrater correlation, and that the reverse is also true. Their results supports the notion that some emotions are more verbally explicit (*gratitude, admiration, approval*) and other more contextual (*grief, surprise, relief, pride*).

If human annotators find it this hard to agree, it seems unreasonable to expect computers to perform much better. Especially since if computers are trained on human annotated data, these disagreements can easily confuse a computer in the learning process reducing the accuracy of predictions further, and even more so since the performance of the classifier is measured again on human annotated data. Although computers are getting better and better at natural language understanding and machine learning is progressing fast, human annotators are unlikely to ever achieve better agreement rates, and therefore the key to improving machine learning results does not lie solely with improving algorithms, but on improving the reliability of datasets.

Some researchers have adopted different annotation schemes beyond the typical 6-8 categories based on Ekman and Plutchik [19, 43]. As mentioned, GoEmotions is one such dataset [16]. In GoEmotions posts were pre-tagged based on a small annotated dataset that were used to train a classifier that then preassigned emotion labels. This approach was also used to balance the sentiments and emotions in the dataset and of course to guide the annotators. As mentioned in the previous section, despite the carefully curated dataset and meticulous annotation task, the final emotion labels were far from balanced. The accuracies achieved by their BERT [18] model can still be considered good for such a large number of categories (27) at an f1 averaged macro score of 0.46 for all categories,

[1] *surprised* in her annotation scheme
[2] *fearful* in her annotation scheme.

but ranging from 0.00 (grief) to 0.86 (gratitude) for specific emotions. Both the GoEmotions [16] and the XED [41] datasets are also multilabel, meaning that one data point can have multiple labels. This models the real world better than single-label categorization, but introduces additional hurdles for the machine learning algorithm.

Abdul et al. [1] achieved some truly remarkable accuracy scores (f1 of 95.68%) by distilling 665 Twitter hashtags into the full 24 categories of Plutchik [43] (the core emotions and their 3 levels of intensities where e.g. *anger* is the core emotion, *annoyance* its less intense category, and *rage* its more intense category). It is not entirely clear how they have managed such a high classifier performance. The authors emphasize the size of their dataset (1.3M tweets), but it seems more likely that they managed to create some truly disjoint categories in their distantly supervised pre-processing stage.

Öhman et al. [41] tried combining some categories that co-occurred and were confused the most by their model, such as *anger* and *disgust*. This resulted in significant increases in accuracy. The exclusion of the categories of *neutral* and *surprise* also improved the results, as did replacing names and locations with generic tags using NER (Named Entity Recognition).

4 The Future of Emotion Detection in NLP

There are no claims in this paper as to what the "correct" theory of emotions should be, but the theories of emotion that much of the NLP work – and subsequent downstream tasks in digital humanities and computational social sciences – today is based upon, is heavily reliant on emotion theories created for a very different modality than text. I believe the key to improving emotion detection is first and foremost a re-thinking of emotion categorization. There are a few different options here: Ideally, new theories specifically developed for the modality of text could be developed or explored. However, this would require vast interdisciplinary collaboration. A second, less resource-heavy option, could be something akin to the approach taken in [16] where sentences were pre-tagged with emotions based on the classifications results from a classifier trained on a small dataset. Another, similar approach would be to use existing emotion lexicons to pre-tag data. However, this would likely work best for the data points where lexical items are highly correlated with the overall emotion expressed. This has already been shown to be the easiest data for annotators to annotate, so the effects might be marginal.

A larger issue has to do with the overlap of emotion categories. Many categories are by definition closer to each other than others and it is quite impossible to create truly disjoint categories, unless some very relevant categories are excluded. With very few categories, such as simply positive and negative, the range of emotions expressed is not suitable for most downstream tasks. The more categories that are in the annotation scheme, the better the categorization represents true human emotions, but this usually means more overlap between categories

possibly leading to more confusion for the learning algorithm in machine learning.

In [39] I describe annotator experiences from an emotion annotation task. Based on the comments from the annotators, the first step in making the annotation task less repetitive and thus easier for annotators would be to add context to the sentence being annotated. However, although annotating with context is much easier for the annotator, it also renders the annotation heavily context-dependent. Overall, more context is a desirable feature, however, having the same sentence in two different contexts, with that context unavailable to a machine learning model, could lead to a confusing training process. More research into the pros and cons of context-dependent annotations versus context-free annotations is required. In addition to comparing these two approaches, one possible solution would be to expand the sequence to go beyond sentence-level.

There is some evidence that the optimal granularity lies between sentence and document level - social media posts or paragraphs are likely close to optimal here. Tweets in particular are somewhat self-contained and due to platform constraints also quite short and therefore lend themselves well to emotion annotation. Since not all downstream tasks are based on Twitter, unfortunately, Tweets as source data are likely to be quite domain-dependent.

All in all, the process of emotion annotation and detection is very difficult with many hurdles still to be resolved. With this paper, we would like to contribute to solving some of these issues.

References

1. Abdul-Mageed, M., Ungar, L.: 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). vol. 1, pp. 718–728 (2017)
2. Alm, C.O.: Characteristics of high agreement affect annotation in text. In: Proceedings of the fourth linguistic annotation workshop. pp. 118–122 (2010)
3. Alm, C.O., Sproat, R.: Emotional sequencing and development in fairy tales. In: International Conference on Affective Computing and Intelligent Interaction. pp. 668–674. Springer (2005)
4. Andreevskaia, A., Bergler, S.: Clac and clac-nb: Knowledge-based and corpus-based approaches to sentiment tagging. In: Proceedings of the 4th International Workshop on Semantic Evaluations. pp. 117–120. SemEval '07, Association for Computational Linguistics, Stroudsburg, PA, USA (2007)
5. Barrett, L.F., Lewis, M., Haviland-Jones, J.M.: Handbook of emotions. Guilford Publications (2016)
6. Bayerl, P.S., Paul, K.I.: What determines inter-coder agreement in manual annotations? a meta-analytic investigation. Computational Linguistics **37**(4), 699–725 (2011)
7. Bermingham, A., Smeaton, A.F.: A study of inter-annotator agreement for opinion retrieval. In: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval. pp. 784–785 (2009)
8. Boland, K., Wira-Alam, A., Messerschmidt, R.: Creating an annotated corpus for sentiment analysis of german product reviews (2013)

9. Cambria, E., Hussain, A.: Sentic computing. marketing **59**(2), 557–577 (2012)
10. Cambria, E., Livingstone, A., Hussain, A.: The hourglass of emotions. In: Cognitive behavioural systems, pp. 144–157. Springer (2012)
11. Cambria, E., Mazzocco, T., Hussain, A., Eckl, C.: Sentic medoids: Organizing affective common sense knowledge in a multi-dimensional vector space. In: International Symposium on Neural Networks. pp. 601–610. Springer (2011)
12. Cambria, E., Schuller, B., Xia, Y., Havasi, C.: New avenues in opinion mining and sentiment analysis. IEEE Intelligent Systems **28**(2), 15–21 (2013)
13. Campbell, N.: Perception of affect in speech-towards an automatic processing of paralinguistic information in spoken conversation. In: Eighth International Conference on Spoken Language Processing (2004)
14. Cowen, A.S., Keltner, D.: Self-report captures 27 distinct categories of emotion bridged by continuous gradients. Proceedings of the National Academy of Sciences **114**(38), E7900–E7909 (2017)
15. Cowen, A.S., Laukka, P., Elfenbein, H.A., Liu, R., Keltner, D.: The primacy of categories in the recognition of 12 emotions in speech prosody across two cultures. Nature human behaviour **3**, 369 – 382 (2019)
16. Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., Ravi, S.: GoEmotions: A Dataset of Fine-Grained Emotions. arXiv preprint arXiv:2005.00547 (2020), https://identifiers.org/arxiv:2005.00547
17. Devitt, A., Ahmad, K.: Sentiment analysis and the use of extrinsic datasets in evaluation. In: LREC (2008)
18. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (2019). https://doi.org/10.18653/v1/N19-1423, https://www.aclweb.org/anthology/N19-1423
19. Ekman, P.: Universals and cultural differences in facial expressions of emotion. In: Nebraska symposium on motivation. University of Nebraska Press (1971)
20. Ekman, P.: An argument for basic emotions. Cognition & emotion **6**(3-4), 169–200 (1992)
21. Ekman, P., Friesen, W.V.: Constants across cultures in the face and emotion. Journal of personality and social psychology **17**(2), 124 (1971)
22. Greenhill, A., Holmes, K., Lintott, C., Simmons, B., Masters, K., Cox, J., Graham, G.: Playing with science: Gamised aspects of gamification found on the online citizen science project-zooniverse. In: GAMEON'2014. EUROSIS (2014)
23. Hofmann, J., Troiano, E., Sassenberg, K., Klinger, R.: Appraisal theories for emotion classification in text. arXiv preprint arXiv:2003.14155 (2020)
24. Hsueh, P.Y., Melville, P., Sindhwani, V.: Data quality from crowdsourcing: A study of annotation selection criteria. In: Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing. pp. 27–35. HLT '09, Association for Computational Linguistics, Stroudsburg, PA, USA (2009), http://dl.acm.org/citation.cfm?id=1564131.1564137
25. Keltner, D., Sauter, D., Tracy, J., Cowen, A.: Emotional expression: Advances in basic emotion theory. Journal of nonverbal behavior pp. 1–28 (2019)
26. Keltner, D., Tracy, J.L., Sauter, D., Cowen, A.: What basic emotion theory really says for the twenty-first century study of emotion. Journal of nonverbal behavior **43**(2), 195–201 (2019)

27. Kiritchenko, S., Mohammad, S.M.: Examining gender and race bias in two hundred sentiment analysis systems. arXiv preprint arXiv:1805.04508 (2018)
28. Liu, C., Osama, M., de Andrade, A.: DENS: A Dataset for Multi-class Emotion Analysis. ArXiv **abs/1910.11769** (2019)
29. Mohammad, S.: A practical guide to sentiment annotation: Challenges and solutions. In: WASSA@ NAACL-HLT. pp. 174–179 (2016)
30. Mohammad, S., Bravo-Marquez, F., Salameh, M., Kiritchenko, S.: SemEval-2018 Task 1: Affect in Tweets. In: Proceedings of The 12th International Workshop on Semantic Evaluation. pp. 1–17. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018). https://doi.org/10.18653/v1/S18-1001, https://www.aclweb.org/anthology/S18-1001
31. Mohammad, S., Turney, P.: Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In: Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text. pp. 26–34 (2010)
32. Mohammad, S.M., Turney, P.D.: Crowdsourcing a word-emotion association lexicon **29**(3), 436–465 (2013)
33. Mozetič, I., Grčar, M., Smailović, J.: Multilingual Twitter sentiment classification: The role of human annotators. PloS one **11**(5), e0155036 (2016)
34. Munezero, M., Montero, C., Sutinen, E., Pajunen, J.: Are They Different? Affect, Feeling, Emotion, Sentiment, and Opinion Detection in Text. IEEE Transactions on Affective Computing **5**(02), 101–111 (apr 2014). https://doi.org/10.1109/TAFFC.2014.2317187
35. Mäntylä, M.V., Graziotin, D., Kuutila, M.: The evolution of sentiment analysis—a review of research topics, venues, and top cited papers. Computer Science Review **27**, 16 – 32 (2018). https://doi.org/https://doi.org/10.1016/j.cosrev.2017.10.002, http://www.sciencedirect.com/science/article/pii/S1574013717300606
36. Ng, H.T., Lim, C.Y., Foo, S.K.: A case study on inter-annotator agreement for word sense disambiguation. In: SIGLEX99: Standardizing Lexical Resources (1999)
37. Ngai, S.: Ugly feelings, vol. 6. Harvard University Press Cambridge, MA (2005)
38. Oberländer, L.A.M., Reich, K., Klinger, R.: Experiencers, stimuli, or targets: Which semantic roles enable machine learning to infer the emotions? In: Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media. pp. 119–128 (2020)
39. Öhman, E.: Challenges in Annotation: Annotator Experiences from a Crowdsourced Emotion Annotation Task. In: Digital Humanities in the Nordic Countries 2020. CEUR Workshop Proceedings (2020)
40. Öhman, E.: The Language of Emotions: Building and Applying Computational Methods for Emotion Detection for English and Beyond. Ph.D. thesis, Helsingin yliopisto (2021)
41. Öhman, E., Pàmies, M., Kajava, K., Tiedemann, J.: XED: A Multilingual Dataset for Sentiment Analysis and Emotion Detection. In: Proceedings of the 28th International Conference of Computational Linguistics (COLING 2020) (2020)
42. Öhman, E.S., Tiedemann, J., Honkela, T., Kajava, K.: Creating a dataset for multilingual fine-grained emotion-detection using gamification-based annotation. In: Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. Association for Computational Linguistics (2018)
43. Plutchik, R.: A general psychoevolutionary theory of emotion. Theories of emotion **1**, 3–31 (1980)

44. Rossi, R.: Alkukantaisuus ja tunteet — Primitivismi 1900-luvun alun suomalaisessa kirjallisuudessa. No. 1456 in Suomalaisen Kirjallisuuden Seuran toimituksia, Suomalaisen Kirjallisuuden Seura (2020)
45. Scarantino, A.: The philosophy of emotions and its impact on affective science. In: Barrett, L.F., Lewis, M., Haviland-Jones, J.M. (eds.) The handbook of emotions, chap. 1, pp. 3–48. Guilford Publications (2016)
46. Schmidt, T., Burghardt, M., Dennerlein, K.: Sentiment annotation of historic German plays: An empirical study on annotation behavior. Proceedings of 2nd Conference on Language, Data, and Knowledge (2018)
47. Sprugnoli, R., Tonelli, S., Marchetti, A., Moretti, G.: Towards sentiment analysis for historical texts. Digital Scholarship in the Humanities **31**(4), 762–772 (2016)
48. Strapparava, C., Mihalcea, R.: Annotating and identifying emotions in text. In: Intelligent information access, pp. 21–38. Springer (2010)
49. Turney, P.D.: Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. pp. 417–424. Association for Computational Linguistics, Stroudsburg, PA, USA (2002)
50. Wiebe, J., Riloff, E.: Creating subjective and objective sentence classifiers from unannotated texts. In: International Conference on Intelligent Text Processing and Computational Linguistics. pp. 486–497. Springer (2005)