Best Practices in the Creation and Use of Emotion Lexicons

Saif M. Mohammad
National Research Council Canada
saif.mohammad@nrc-cnrc.gc.ca

Abstract

Words play a central role in how we express ourselves. Lexicons of word–emotion associations are widely used in research and real-world applications for sentiment analysis, tracking emotions associated with products and policies, studying health disorders, tracking emotional arcs of stories, and so on. However, inappropriate and incorrect use of these lexicons can lead to not just sub-optimal results, but also inferences that are directly harmful to people. This paper brings together ideas from Affective Computing and AI Ethics to present, some of the practical and ethical considerations involved in the creation and use of emotion lexicons — *best practices*. The goal is to provide a comprehensive set of relevant considerations, so that readers (especially those new to work with emotions) can find relevant information in one place. We hope this work will facilitate more thoughtfulness when one is deciding on what emotions to work on, how to create an emotion lexicon, how to use an emotion lexicon, how to draw meaningful inferences, and how to judge success.

1 Introduction

Words often convey affect (emotions, sentiment, feelings, and attitudes); either explicitly through their core meaning (denotation) or implicitly through connotation. For example, *dejected* denotes sadness. On the other hand, *failure* simply connotes sadness. Either through denotation or connotation, both words are associated with sadness. A compilation of such associations is referred to as a *word–affect association lexicon* (aka *emotion lexicon*).[1] An entry in a lexicon usually includes a word, an emotion category or affect dimension (e.g., joy, fear, valence, arousal, etc.), and a score indicating association (or strength of association).

Examples of emotion lexicons include the General Inquirer (Stone et al., 1966), ANEW (Nielsen, 2011; Bradley and Lang, 1999), LIWC (Pennebaker et al., 2001), Pittsburgh Subjectivity Lexicon (Wilson et al., 2005), *NRC Emotion Lexicon* (Mohammad and Turney, 2010, 2013), and the *NRC Valence, Arousal, and Dominance (VAD) Lexicon* (Mohammad, 2018). These were all created by manual annotation (either by experts or crowdsourced). There also exist lexicons that were generated automatically from large text corpora using statistical and/or machine learning algorithms; e.g., WordNet Affect (Strapparava et al., 2004), SentiWordNet (SWN) (Baccianella et al., 2010).

Emotion lexicons have a wide range of applications in commerce, public health, and research (in NLP, Psychology, Social Sciences, Digital Humanities, etc.). Some notable examples include: tracking brand and product perception via social media posts, tracking support for controversial issues and policies, tracking buy-in for non-pharmaceutical health measures such as social distancing during a pandemic, literary analysis, and developing more natural dialogue systems. The lexicons can be used on their own or in support of neural machine learning (ML) algorithms for emotion recognition.

Lexicon-based emotion analyses are especially popular in real-world applications and research outside of computer science because they are interpretable, have a low carbon footprint, and do not require significant programming expertise. Further, since outputs of ML models are highly dependent on training data, use of a model often requires retraining, and there may not exists labeled data from the target domain Further, Teodorescu and Mohammad (2022) show that when determining broad trends (emotion arcs) and aggregating information from hundreds (if not more) instances for every time step, simple lexicon-based methods are extremely accurate (correlations above 0.95 with ground truth arcs).

[1] This includes *sentiment lexicons* that capture valence (association with the positive–negative dimension) and other lexica that capture affect-related phenomena.

However, inappropriate and incorrect use of these lexicons, can lead to not just sub-optimal results, but also inferences that are directly harmful. For example, using lexicons to infer emotions from limited amount of data to make judgments about refugee applications, to make judgments about which groups of people are shown certain advertisements and which groups are not, marking businesses owned by some groups of people as less liked than that of others, etc.

Emotions are deeply personal, private, and complex. Even the best natural language systems largely only employ pattern matching based on huge amounts of historical data, and thus often do not really understand what the user is trying to convey, let alone how they are feeling. In fact, some recent commercial and governmental uses of emotion recognition have garnered considerable criticism, including: infringing on one's privacy, exploiting vulnerable sub-populations, and even allegations of pseudo-science (Mohammad, 2022b; Wakefield, 2021; ARTICLE19, 2021; Woensel and Nevil, 2019).

This paper brings together ideas from Affective Computing and AI Ethics to present, in one place, some of the practical and ethical considerations involved in the creation and use of emotion lexicons — *best practices*.[2] We hope this work will facilitate more thoughtfulness when one is deciding on what emotions to work on, how to create an emotion lexicon, how to use an emotion lexicon, and how to judge success. Additional benefits of such a document include:

1. Presents the trade-offs of relevant choices so that stakeholders can make informed decisions appropriate for their context.
2. Has citations and pointers; acts as a jumping off point for further reading.
3. Helps engage the various stakeholders of an emotion task with each other. Helps stakeholders challenge assumptions made by researchers and developers.
4. Helps develop harm mitigation strategies.
5. Acts as a useful introductory document on emotion lexicons (complements survey articles).

Note that even though this article is focused on emotion lexicons, many of the ethical considerations apply broadly to natural language lexicons/resources in general. Also, see Mohammad (2022b) for a broader discussion on the ethical considerations associated with automatic emotion recognition (AER).

This work is in the same spirit as other recent innovations in exercising responsible research such as datasheets for datasets (Gebru et al., 2018), model cards for systems (Mitchell et al., 2019), and ethics sheets for AI tasks (Mohammad, 2022a). However, unlike datasheets and model cards which are designed for individual datasets and systems and that are published after the work is done, the goal of this work is to provide a more general-purpose relevant resource, accessible at the very beginning of one's project. Also, unlike an ethics sheet for a automatic emotion recognition that may cover all kinds of ethical considerations associated with the task of interest, this document has a focus on the creation of emotion lexicons and their use in AI tasks.

Ethics considerations are not about objective metrics or simple checklists. They involve engaging with issues that impact stake holders, especially those that are already disadvantaged. Thus, a big component of this work is to raise awareness of relevant issues, to underscore how often there are no easy solutions, and that meaningful change requires painstaking, slow, and deliberate engagement with the stakeholders. Additionally, such documents are useful for those that are impacted to question and challenge assumptions made by unfair decisions of automated systems.

2 Best Practices

Below we present various best practices (practical and ethical considerations) pertaining to 22 aspects of emotion lexicon creation and use. The 22 aspects are grouped under the coarser categories pertaining to a lexicon's life cycle: A. Lexicon Design, B. Annotation, C. Entries in the Lexicon, and D. Applying the Lexicon. Note that while many considerations are presented from the perspective of lexicon creation, they are also relevant to the users of a lexicon — knowing what decisions were made during the creation of a lexicon help one to assess appropriateness of using the lexicon.

The goal is to provide a comprehensive set of relevant considerations, so that readers (especially those new to research or new to work with emotions) can find the information in one place. Thus,

[2] This paper is a reframed and expanded avatar of an earlier datasheet paper for emotion lexicons (Mohammad, 2020).

we include both the considerations that are especially specific to emotions, as well as others that apply more broadly (even if they are somewhat well known). Also, the points listed below are not meant to be the final word, but rather jumping off points for further thought and discussion.

2.1 Overview

An overview of the 22 aspects is presented below; followed by the detailed descriptions.

A. LEXICON DESIGN

1. Purpose or Objective
2. Emotion Category or Dimension
3. Word Senses and Dominant Sense Priors
4. Discrete or Continuous Value Labels

B. ANNOTATION

5. Questionnaire
6. Comparative Annotations
7. Annotators
8. Quality Control

C. ENTRIES IN THE LEXICON

9. Annotation Aggregation
10. Relative (not Absolute)
11. Coverage
12. Not Immutable
13. Perceptions (not "truth")
14. Socio-Cultural Biases
15. Inappropriate Biases
16. Errors

16. Mechanism to Report and Fix Errors

D. APPLYING THE LEXICON

18. Fit of the Lexicon to One's Data
19. Rescaling the Lexicon for One's Task
20. Metrics & Features Drawn from the Lexicon
21. Removing Neutral Words
22. Inferences

2.2 Detailed Descriptions

A. LEXICON DESIGN

#1. Purpose or Objective: Consider and document the objective(s) of building the emotion lexicon. There can be more than one objective. The objectives guide various design choices involved in the creation of the lexicon. See Selbst et al. (2019) for common pitfalls in designing and framing sociotechnical systems; and Mohammad (2022b) for common pitfalls in designing and framing automatic emotion recognition tasks. Users of emotion lexicons can study the purpose of each lexicon to determine which is most suitable for their use case.

Broadly speaking, the objectives tend to be around the study of word–emotion associations (exploring various research questions at the intersection of language an emotions) and aiding automatic emotion detection from utterances. However, individual projects often have specific goals, for example, to study specific phenomenon such as loneliness and empathy, to study inappropriate biases, to detect what emotions people perceive from utterances, to study how automatic systems should perceive the emotions in utterances, how automatic systems should use words to convey emotions, etc. It is important to recognize that some of these objectives are very related, but they have important differences. For example, while a general-purpose emotion lexicon will capture a number of benign associations, it will also capture inappropriate societal biases. If one wants to use a lexicon in a text generation system, then they should either use a lexicon designed specifically for that purpose, or address the biases in a general purpose lexicon, before using it.

Work using emotion lexicons should not claim that using it one can determine one's emotional state from their utterance. At best, recognition systems (whether they use emotion lexicons or not) capture what one is trying to convey or what is perceived by the listener/viewer; and even there, given the complexity of human expression, they are often inaccurate. Several studies have shown that it is difficult to fully measure psychological states of people (Stark, 2018; Barrett, 2017b).

In contrast, statistical analyses with features drawn from emotion lexicons can be used to accurately determine broad trends in the emotional state of a population over time (Teodorescu and Mohammad, 2022). Here, inferences are drawn at aggregate level from much larger amounts of data. Studies on public health, such as those on loneliness (Guntuku et al., 2019; Kiritchenko et al., 2020), depression (De Choudhury et al., 2013; Resnik et al., 2015), suicidality prediction (MacAvaney et al., 2021), bipolar disorder (Karam et al., 2014), stress (Eichstaedt et al., 2015), emotions during a pandemic (Vishnubhotla and Mohammad, 2022), and general well-being (Schwartz et al., 2013) fall in this category. Here too, however, it is best to be cautious in making claims about mental state, and use emotion recognition as one source of evidence amongst many (and involve expertise from public health and psychology).

#2. Emotion Category or Dimension: A key decision in the creation of an emotion lexicon is which conceptualization of emotion to use and which facet of emotion to capture. For example, should it capture emotion categories such as joy, sadness, fear, optimism, etc., or will it capture dimensions such as valence, arousal, and dominance. Psychologists and neuro-scientists have identified several theories of emotion that can inform the choice of categories and dimensions, including: the Basic Emotions Theory (BET) (Ekman, 1992; Ekman and Davidson, 1994), the Dimensional Theory (Osgood et al., 1957; Russell, 1980; Russell and Mehrabian, 1977; Russell, 2003), Cognitive Appraisal Theory (Scherer, 1999; Lazarus, 1991), and the Theory of Constructed Emotions (Barrett, 2017b).

Since ML approaches rely on human-annotated data (which can be hard to obtain in large quantities), emotion recognition research has often gravitated to the Basic Emotions Theory, as that work allows one to focus on a small number of emotions. This attraction has been even stronger in the vision research because of BET's suggested mapping between facial expressions and emotions. However, many of the tenets of BET, such as the universality of some emotions and their fixed mapping to facial expressions, stand discredited or are in question (Barrett, 2017a; Barrett et al., 2019).

Carefully consider which emotion formulation you wish to capture in your lexicon, or is appropriate for your task/project. For example, one may choose to work with the dimensional model or the model of constructed emotions if the goal is to infer behavioural or health outcome predictions. Despite criticisms of BET, it makes sense for some NLP work to focus on *categorical emotions* such as joy, sadness, guilt, pride, fear, etc. (including what some refer to as basic emotions) because people often talk about their emotions in terms of these concepts. Many human languages have words for these concepts (even if our individual mental representations for these concepts vary to some extent) (Wierzbicka, 1999). However, note that work on categorical emotions by itself is not an endorsement of the BET. Do not refer to some emotions as basic emotions, unless you mean to convey your belief in the BET. Careless endorsement of theories can lead to the perpetuation of ideas that are actively harmful (such as suggesting we can determine internal state from outward appearance—physiognomy).

#3. Word Senses and Dominant Sense Priors: Words when used in different senses and contexts may be associated with different emotions. The entries in the emotion lexicons are mostly indicative of the emotions associated with the predominant senses of the words. This is usually not too problematic because most words have a highly dominant main sense (which occurs much more frequently than the other senses). In specialized domains, some terms might have a different dominant sense than in general usage. Entries in the lexicon for such terms should be appropriately updated or removed. However, if the goal of the project is to create a lexicon for a specialized domain, then one should guide the annotation process accordingly.

#4. Discrete or Continuous Value Labels: Many emotion lexicons have discrete binary labels for words (positive–negative, joy–no joy, fear–no fear, and so on). Lexicons such as ANEW and the NRC VAD Lexicon have real-valued scores between 0 and 1, -1 and 1, 0 to 5, 0 to 100, etc. Real-valued scores allows one to make finer distinctions in the degree of emotion. They allow one to determine the intensity of emotion. Binary-labeled lexicons are used primarily to determine density of emotion word usage; for example, to explore whether there is a higher percentage of tweets with loneliness words during the Covid-19 pandemic, than in the years before the pandemic. Determine which type of lexicon is more aligned with your objectives.

B. ANNOTATION

#5. Questionnaire: Arguably the most crucial aspect in the creation of an emotion lexicon is the questionnaire. What is asked and how it is asked determines the outcome. Below are key recommendations in the design of questionnaires:

a. Where appropriate, break the task/question into simpler sub-tasks/sub-questions.

b. It is better to have separate tasks for different questions and emotion dimensions. Asking for responses about more than one emotion dimension requires the annotator to switch contexts and leads to more cognitive load.

c. Keep the instructions clear and easy to follow.

d. Examples are more important than definitions. People tend to learn faster and better through examples. It is still good to include simple definitions of relevant concepts.

e. Refer to the theories for emotions work in psychology on to how to collect emotional information from respondents. Especially useful are the terms used to define emotion dimensions: e.g., as per the dimensional model of emotions (Russell, 1980) *arousal* is defined as the active–sluggish dimension, in the stereotype content model of social perception (Cuddy et al., 2008), *warmth* is defined as the trustworthiness, friendliness, kindness dimension. These words should be used when eliciting annotation responses.

f. Keep the instructions brief. This is respectful of annotator time, and one can only keep track of a limited number of instructions at a time.

g. Explain the purpose of the annotation task. This is respectful of annotators. People have a right to know (in appropriate detail) what research they are contributing their time for. This may also lead to more engaged annotators.

h. Include an optional comment box that gives annotators a way to provide feedback, raise issues, and to be heard.

i. Make the questionnaire and instructions freely available. This helps others to build on your work. It allows users to see exactly how the questions were phrased, and thus how to interpret the resulting emotion lexicon.

See also other data curation and questionnaire development tips from non-NLP fields such as psychology (Aguinis et al., 2021).

#6. Comparative Annotations: Real-valued scores provide fine-grained emotion information; however, it is difficult for humans to provide direct scores at this granularity. A popular approach to obtain real-valued scores is by providing the annotators with numeric rating scales.[3] These scales have numbers (usually 1 to 5 or 1 to 7) and the annotator has to select which number is most indicative of the degree of association with the property of interest for the given word; given that the lowest number on the scale indicates least association and the highest number indicates the most association.[4] The scores for an item from multiple annotators is averaged to obtain a real-valued score that is assigned to the word–emotion pair.

A common problem of annotation by rating scales is inconsistencies in annotations among different annotators. One annotator might assign a score of 87 to one word, while another annotator may assign a score of 81 to the same word. It is also common that the same annotator might assign different scores to the same word, if asked to annotate again after a period of time. Further, annotators often have a bias towards selecting scores in the middle of the scale, known as *scale region bias* (Presser and Schuman, 1996; Baumgartner and Steenkamp, 2001).

Paired Comparisons (Thurstone, 1927; David, 1963) is a comparative annotation method, where respondents are presented with pairs of items and asked which item has more of the property of interest (for example, which is more positive). The annotations can then be converted into a ranking of items by the property of interest, and one can even obtain real-valued scores indicating the degree to which an item is associated with the property of interest. The paired comparison method does not suffer from the problems discussed above for the rating scale, but it requires a large number of annotations—order N^2, where N is the number of items to be annotated.

Best–worst scaling (BWS) (Louviere, 1991) is a form of comparative annotation, like paired comparison, but it requires much fewer annotations. Annotators are given n items (an n-tuple, where $n > 1$ and commonly $n = 4$).[5] They are asked which item is the *best* (highest in terms of the property of interest) and which is the *worst* (least in terms of the property of interest). When working on 4-tuples, best–worst annotations are particularly efficient because each best and worst annotation will reveal the order of five of the six item pairs (e.g., for a 4-tuple with items *w, x, y,* and *z*, if *w* is the best, and *z* is the worst, then $w > x$, $w > y$, $w > z$, $x > z$, and $y > z$). Real-valued scores of association between the items and the property of interest can be determined using simple arithmetic on the number of times an item was chosen best and number of times it was chosen worst (Orme, 2009; Flynn and Marley, 2014). It has been empirically shown that three annotations each for $2N$ 4-tuples is sufficient for obtaining reliable scores

[3] https://www.questionpro.com/blog/rating-scale/

[4] It is good practice to anchor the numeric values with labels such as maximum/moderate/low association.

[5] At its limit, when $n = 2$, best–worst scaling reduces to a *paired comparison* (Thurstone, 1927; David, 1963); However, then a much larger set of tuples need to be annotated (closer to N^2).

(where N is the number of items) (Louviere, 1991; Kiritchenko and Mohammad, 2016). Kiritchenko and Mohammad (2016; 2017) showed through empirical experiments on emotion lexicons that BWS produces more reliable and more discriminating scores than those obtained using rating scales.

Within the NLP community, BWS has been used for creating datasets for relational similarity (Jurgens et al., 2012), word-sense disambiguation (Jurgens, 2013), word–sentiment intensity (Kiritchenko and Mohammad, 2016), sentence–sentence semantic relatedness (Abdalla et al., 2023), etc.

#7. Annotators: Who is recruited to annotate the data also impacts the lexicon that is generated.

a. *Experts or Crowd:* If a task has clear correct and wrong answers and knowing the answers requires some training/qualifications, then one can employ domain experts to annotate the data. However, emotion annotations largely do not fall in this category. People are the best judges of their emotions and how they use words to communicate them. If the goal is to determine how people use language or we want to know how people perceive words, phrases, and sentences then we might want to employ a large number of annotators (crowdsourcing). Note that this is also a scenario where there can be more than one appropriate answer.

b. *Diversity:* Emotion lexicons are a function of their annotators. Consider who all should be represented in the annotator pool, and actively recruit people from under-represented groups. Seek appropriate demographic information (respectfully and ethically). Document annotator demographics at an aggregate level.

c. *Informed Consent, Privacy, and Potential for Harms:* Provide a clear and easy-to-understand description of what the task will involve, potential risks, and what information will be collected, before obtaining consent from the annotators. Note that if the terms included for annotation or the chosen dimension of annotation is particularly negative, then there may be significant risk of adversely impacting the annotator's mental health. In such cases, suitable avenues for recourse must be provided.

d. *Remuneration:* Determine fair compensation for the task. Inform the annotators of the pay and the time commitment expected.

e. *Miscellaneous:* There are several other ethical considerations also involved with such work such as: worker invisibility, lack of learning trajectory, humans-as-a-service paradigm, worker well-being, and worker rights (Dolmaya, 2011; Fort et al., 2011; Standing and Standing, 2018; Irani and Silberman, 2013).

f. *Ethics Approval:* Obtain approval of the project and annotation plan from your institution's research ethics board before conducting the annotation. The ethics boards are also a great source of feedback for improving the ethical standards of the annotation process. If unsure whether some work requires ethics approval, reach out to the ethics board. Many institutions provide expedited review in cases of low risk.

Document these considerations so that the users can judge suitability of the lexicon for their work.

#8. Quality Control: Good quality control strategies can make a large difference for any scenario of annotations, but are especially important when the annotations are done via crowdsourcing. Quality control strategies can be of three kinds:

Type 1: applied before data annotation begins

Type 2: applied during data annotation, and

Type 3: applied after data annotation.

It is recommended to apply measures of all three kinds. Examples of Type 1 include: careful questionnaire design and setting up training or qualification annotations to screen annotators.

A particularly powerful example of a Type 2 measure is to intersperse the instances with small number of hidden gold instances ($\sim$5%) — instances for which the appropriate label(s) are predetermined (by, say, the authors). If a crowd worker responds with an answer not already marked as appropriate, then they are immediately notified, the annotation is discarded. If an annotator's accuracy on the gold questions falls below a pre-chosen threshold (say, 80%), then they are refused further annotation, and all of their annotations are discarded. This way the gold instances serve as a mechanism to avoid malicious annotations, as well as a way to further train the annotators. This also avoids scenarios where an annotator provides responses to a large number of questions, only to later learn that they misinterpreted something, rendering all of their annotations useless. The use of gold questions was popularized by the crowdsourcing platform CrowdFlower (now, Figure8).

Examples of Type 3 quality control measures include: removal of responses from people who answer questions too quickly, or whose responses are more than two standard deviations away from the responses of others. There also exist approaches that identify which annotators to trust using machine learning algorithms (Raykar and Yu, 2012; Hovy et al., 2013).

C. ENTRIES IN THE LEXICON

#9. Annotation Aggregation: Each instance in a lexicon (usually a word) is often annotated by a number of annotators. Standard practice in aggregating the responses from multiple annotators is to take the most frequent response. However, it should be noted that sometimes other responses are also appropriate. Further, different socio-cultural groups can perceive language differently, and taking the majority vote can have the effect of only considering the perceptions of the majority group. When these views are crystallized in the form of a lexicon, it can lead to the false perception that the norms so captured are “standard" or “correct", whereas other associations are “non-standard" or “incorrect". Thus, it is worth explicitly disavowing that view and stating that the lexicon simply captures the perceptions of the majority group among the annotators. Thus, it is recommended to also make available disaggregated annotations (annotations in their raw form – without aggregation). Note that it is also problematic to consider all annotator responses as valid because sometimes annotators make mistakes, and some may have inappropriate biases (see #15).

#10. Relative (not Absolute): The absolute values of the association scores themselves usually have no meaning. The scores help order the words relative to each other. For example, a term with a high valence score is associated with more positiveness than a term with with a lower score.

#11. Coverage: Some lexicons have a few hundred terms, and some have tens of thousands of terms. However, even the largest lexicons do not include all the terms in a language. Mostly, they include entries for the canonical forms (lemmas), but some also include morphological variants. The high-coverage lexicons, such as the NRC Emotion Lexicon, have tens of thousands of terms. However, when using the lexicons in specialized domains, one may find that a number of common terms in the domain are not listed in the lexicons.

#12. Not Immutable: The associations do not indicate an inherent unchangeable attribute. Emotion associations can change with time, but these lexicon entries are largely fixed. They pertain to the time they are created or the time associated with the corpus from which they are created.

#13. Perceptions (not “truth”): Emotion lexicons largely capture how speakers of a language perceive the emotion associations of words. As mentioned in the previous bullet, this can change with time. Further, it can also be different for different people. Mohammad and Turney (2013) found that when the annotators are asked to judge emotion associations in terms of ‘how speakers of a language perceive the word’, the results have lower variance than when asked ‘the emotions evoked in the annotator’. Consider your objective when deciding which of the two framings (or some other) is more appropriate for your use case.

#14. Socio-Cultural Biases: Since the emotion lexicons have been created by people (directly through crowdsourcing or indirectly through the texts written by people) they capture various human biases. These biases may be systematically different for different socio-cultural groups. Document who produced the data (people from which countries, what is the gender distribution, age distribution, etc.) in the paper describing the dataset or in the associated datasheet. An advantage of crowdsourcing is that the annotations are from a wider pool of annotators; however, crowd annotators are systematically different from, and not representative of, the general population.

#15. Inappropriate Biases: Some of the human biases that have percolated into the lexicons may be rather inappropriate. For example, entries with low valence scores for certain demographic groups or social categories. Studying such biases in the lexicon can be useful to show and address some of the historical inequities that have plagued humankind. Nonetheless, when these lexicons are used in specific tasks, care must be taken to remove such entries from the lexicons where necessary.

#16. Errors: Even though the researchers take several measures to ensure high-quality and reliable data annotation (e.g., multiple annotators, clear and concise questionnaires, framing tasks as comparative annotations, interspersed check questions, etc.), human-error can never be fully eliminated in large-scale annotations. Expect a small number of clearly wrong entries. Automatically generated lexicons also can have erroneous entries. They are often built on the assumption that the tendency of a word to co-occur with emotion-associated seed terms is proportional to its association with that emotion. However, in any corpus, there will always be some amount of chance high co-occurrences that are not accurate reflections of the true associations.

#17. Mechanism to Report and Fix Errors: Provide a mechanism for users to report issues and errors. Fix errors and where appropriate issue warnings for how some types of entries can be mis-interpreted or misused. Periodically assess whether certain types of entries need to be proactively checked. For example, there has been growing recognition that emotion associations associated with identity groups are particularly sensitive, affected by historical bias, and so one must be careful in how they interpret the associations captured in lexicons.

D. APPLYING THE LEXICON

#18. Examining the Fit of the Lexicon: Manually examine the emotion associations of the most frequent terms in your data. Remove entries from the lexicon that are not suitable (due to mismatch of sense, inappropriate human bias, etc.).

#19. Rescaling the Lexicon for One's Task: Depending on your specific use case, you may choose to re-scale the scores from 0 to 1, -1 to 1, 1 to 10, etc. Note that if using the lexicon entries as features in machine learning experiments, the scale (0 to 1 or -1 to 1) can make a difference—e.g. if the score is used as a weight for features.

#20 Metrics and Features Drawn from the Lexicon: For text analysis, one can calculate various metrics such as the percentage of emotion words (when the lexicons provides a list of words associated with a category) or average emotion intensity (for real-valued associations). When determining the scores, a further choice is how to handle words that are not in the lexicon. Two common approaches include: 1. Treat words that are not in the lexicon as neutral; 2. Ignore these words in the calculation of the scores. The latter approach does not make assumptions of neutrality, and is not impacted by the number of such out of lexicon words in a piece of text. See Teodorescu and Mohammad (2022) for a systematic analysis of the impact of various lexicon features on the quality of emotion arcs generated with them.

#21. Creating Subsets of the Lexicon: Sometimes it is better to use a subset of the emotion lexicon, rather than the whole lexicon.

Removing Neutral Words: One can use the whole lexicon to calculate metrics such as average valence of the words in a text; however, one can also choose to disregard terms with close to 0 valence scores. when calculating the same metric. Removal of such neutral terms from the analysis will show greater variations in the average scores when comparing across different sets of data of interest or across time. For example, when looking at the average tweet happiness over time of day, using full or neutral-removed lexicon is expected to get roughly similar curves, but the neutral-removed lexicon will show a greater amplitude (divergence of scores from the peaks to troughs). (Dodds et al., 2011) describes this as turning up the magnifier knob in a microscope. Note, however, that just having larger score differences between the target and control does not mean that the emotion word usage is substantially different or significant; and conversely, just because the score difference for a metric is small in value does not mean that the differences in emotion word usages are not substantial. (More on this in #22).

Removing Low-Association Words: Use of low-association terms from a lexicon may not be beneficial for some downstream applications. These entries may also include a greater percentage of annotation errors. See Teodorescu and Mohammad (2022) for experiments on multiple datasets and multiple emotion dimensions that examine usefulness of removing low-association terms from a lexicon when generating emotion arcs.

Removing Highly Polysemous and Certain Domain Words: For some applications, it is beneficial to discard highly ambiguous words. Entries for highly ambiguous words are more likely to include emotion associations for a sense that is not common in one's data. As stated in #3, it is also recommended to remove entries not appropriate for the target domain; e.g., the word *harry* has a negative meaning, but it should not be used when analyzing text where a person has the name *Harry*.

#22. Inferences: When drawing inferences from texts using counts of emotion words:

a. It is more appropriate to make claims about emotion word usage rather than emotions of the speakers. For example, *'the use of anger words grew by 20%'* rather than *'anger grew by 20%'*. A marked increase in anger words is likely an indication that anger increased, but there is no evidence that anger increased by 20%. Further, it is important to understand the emotion metrics and to interpret them accordingly. For example, many off-the-shelf tools provide a "sentiment score" for the input textual instances, without providing adequate details about what this score means. As discussed in #21, the scores themselves can have large or small values, and just knowing that the score difference between a target and control is large (or small) is not enough to draw meaningful inference. On the other hand, grounded metrics that tie the score to attributes such as percentage of positive words tend to be less open to misinterpretation.

b. Comparative analysis is your friend. Often, emotion word counts on their own are not useful. For example, *'the use of anger words grew by 20% when compared to [data from last year, data from a different person, etc.]'* is more useful than saying *'on average, 5 anger words were used in every 100 words'*.

c. Lexicon features (or any other automatically drawn features) are *not* well suited to draw meaningful emotional inferences from individual utterances. Human language and behaviour are highly variable and complex. However, with careful design, they can be useful to draw inferences about broad trends at an aggregate level (Teodorescu and Mohammad, 2022).

d. Inferences drawn from large amounts of text are more reliable than those drawn from small amounts of text. Teodorescu and Mohammad (2022) show that this is the single most important feature in determining the fidelity of the predicted emotion trends with the true emotion trends, among a host of features they explored. For many emotion dimensions and dataset domains, it is advisable to determine aggregate emotion scores using at least 100 instances. For example, if there are at least 100 tweets per day about a product of interest, the average valence scores of all the words in the tweets every day is expected to produce a fairly accurate valence arc (x-axis is day, y-axis is average valence score for the corresponding day).

3 Limitations

This paper does not present a new NLP model or dataset. Thus, there are no corresponding limitations to discuss. However, the paper itself can be viewed as a document discussing limitations of existing approaches to do sentiment and emotion analysis using emotion lexica. The 22 best practises presented in the paper discuss approaches to engage with and counter these limitations.

While this document was a result of engaging a larger community through blog posts, talks, and discussions, we had relatively low access to developers of commercial sentiment analysis systems. Thus the list presented here may have missed some important considerations. We encourage readers and impacted stakeholders to challenge the assumptions latent in the document, and identify new ethical considerations not included here or not gaining adequate attention in the research community.

4 Concluding Remarks

Emotion lexicons are simple yet powerful tools to analyze text. However, use of the lexicons (even for tasks that it is suited for) can lead to inappropriate bias. Applying a lexicon to any new data should only be done after first investigating its suitability, and requires careful analysis to minimize unintentional harm. In this paper, we presented 22 best practises that include considerations that can help mitigate such unwanted outcomes, as well as strategies to make the best use of emotion lexicons towards drawing meaningful and accurate inferences. The best practises are organized as per a lexicon's life cycle: A. Lexicon Design, B. Annotation, C. Entries in the Lexicon, and D. Applying the Lexicon. We also provide pointers to relevant literature to explore the best practises in more detail. It should be noted that these practises are not meant to be the final word, but rather jumping off points for further thought, discussion, and additional measures towards the responsible use of emotion lexicons.

Acknowledgments

Many thanks to Emiel van Miltenburg, Annika Schoene, Mallory Feldman, Tara Small, Roman Klinger, and Peter Turney for thoughtful comments and discussions.

References

Mohamed Abdalla, Krishnapriya Vishnubhotla, and Saif M. Mohammad. 2023. What makes sentences semantically related: A textual relatedness dataset and empirical study. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Dubrovnik, Croatia. Association for Computational Linguistics.

Herman Aguinis, N. Sharon Hill, and James R. Bailey. 2021. Best practices in data collection and preparation: Recommendations for reviewers, editors, and authors. *Organizational Research Methods*, 24(4):678–693.

ARTICLE19. 2021. Emotional entanglement: China's emotion recognition market and its implications for human rights. https://www.article19.org/wp-content/uploads/2021/01/ER-Tech-China-Report.pdf.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Proceeding of the 7th International Conference on Language Resources and Evaluation*, volume 10 of *LREC '10*, pages 2200–2204.

Lisa Feldman Barrett. 2017a. *How emotions are made: The secret life of the brain*. Houghton Mifflin Harcourt.

Lisa Feldman Barrett. 2017b. The theory of constructed emotion: an active inference account of interoception and categorization. *Social cognitive and affective neuroscience*, 12(1):1–23.

Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, Aleix M Martinez, and Seth D Pollak. 2019. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological science in the public interest*, 20(1):1–68.

Hans Baumgartner and Jan-Benedict E.M. Steenkamp. 2001. Response Styles in Marketing Research: A Cross-National Investigation. *Journal of Marketing Research*, 38(2):143–156.

Margaret M Bradley and Peter J Lang. 1999. Affective norms for English words (ANEW): Instruction manual and affective ratings. Technical report, The Center for Research in Psychophysiology, University of Florida.

Amy JC Cuddy, Susan T Fiske, and Peter Glick. 2008. Warmth and competence as universal dimensions of social perception: The stereotype content model and the bias map. *Advances in experimental social psychology*, 40:61–149.

Herbert Aron David. 1963. *The method of paired comparisons*. Hafner Publishing Company, New York.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *Seventh international AAAI conference on weblogs and social media*, pages 128–137.

Peter Sheridan Dodds, Kameron Decker Harris, Isabel M Kloumann, Catherine A Bliss, and Christopher M Danforth. 2011. Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. *PloS one*, 6(12):e26752.

Julie McDonough Dolmaya. 2011. The ethics of crowdsourcing. *Linguistica Antverpiensia, New Series–Themes in Translation Studies*, (10).

Johannes C Eichstaedt, Hansen Andrew Schwartz, Margaret L Kern, Gregory Park, Darwin R Labarthe, Raina M Merchant, Sneha Jha, Megha Agrawal, Lukasz A Dziurzynski, Maarten Sap, et al. 2015. Psychological language on Twitter predicts county-level heart disease mortality. *Psychological science*, 26(2):159–169.

Paul Ekman. 1992. Are there basic emotions? *Psychological Review*, 99(3):550–553.

Paul Ed Ekman and Richard J Davidson. 1994. *The nature of emotion: Fundamental questions.* Oxford University Press.

T. N. Flynn and A. A. J. Marley. 2014. Best-worst scaling: theory and methods. In Stephane Hess and Andrew Daly, editors, *Handbook of Choice Modelling*, pages 178–201. Edward Elgar Publishing.

Karën Fort et al. 2011. Amazon Mechanical Turk: Gold mine or coal mine? *Computational Linguistics*, 37(2):413–420.

Timnit Gebru, Jamie H. Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, H. Wallach, Hal Daumé, and Kate Crawford. 2018. Datasheets for datasets. In *Proceedings of the conference on Fairness, Accountability, and Transparency in Machine Learning*, Stockholm, Sweden.

Sharath Chandra Guntuku, Rachelle Schneider, Arthur Pelullo, Jami Young, Vivien Wong, Lyle Ungar, Daniel Polsky, Kevin G Volpp, and Raina Merchant. 2019. Studying expressions of loneliness in individuals using Twitter: an observational study. *BMJ open*, 9(11):e030355.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130, Atlanta, Georgia. Association for Computational Linguistics.

Lilly C Irani and M Six Silberman. 2013. Turkopticon: Interrupting worker invisibility in Amazon Mechanical Turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 611–620.

David Jurgens. 2013. Embracing ambiguity: A comparison of annotation methodologies for crowdsourcing word sense labels. In *NAACL*.

David Jurgens, Saif M. Mohammad, Peter Turney, and Keith Holyoak. 2012. Semeval-2012 task 2: Measuring degrees of relational similarity. In *Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval)*, pages 356–364, Montréal, Canada.

Zahi N Karam, Emily Mower Provost, Satinder Singh, Jennifer Montgomery, Christopher Archer, Gloria Harrington, and Melvin G Mcinnis. 2014. Ecologically valid long-term mood monitoring of individuals with bipolar disorder using speech. In *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4858–4862. IEEE.

Svetlana Kiritchenko, Will Hipson, Robert Coplan, and Saif M. Mohammad. 2020. SOLO: A corpus of tweets for examining the state of being alone. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1567–1577, Marseille, France.

Svetlana Kiritchenko and Saif Mohammad. 2017. Best-Worst Scaling More Reliable than Rating Scales: A Case Study on Sentiment Intensity Annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 465–470.

Svetlana Kiritchenko and Saif M. Mohammad. 2016. Capturing reliable fine-grained sentiment associations by crowdsourcing and best–worst scaling. In *Proceedings of The 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, San Diego, California.

Richard S Lazarus. 1991. Progress on a cognitive-motivational-relational theory of emotion. *American psychologist*, 46(8):819.

Jordan J. Louviere. 1991. Best-worst scaling: A model for the largest difference judgments. Working Paper.

Sean MacAvaney, Anjali Mittu, Glen Coppersmith, Jeff Leintz, and Philip Resnik. 2021. Community-level research on suicidality prediction in a secure environment: Overview of the CLPsych 2021 shared task. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 70–80, Online. Association for Computational Linguistics.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229.

Saif Mohammad. 2022a. Ethics sheets for AI tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8368–8379, Dublin, Ireland. Association for Computational Linguistics.

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

Saif M. Mohammad. 2020. Practical and ethical considerations in the effective use of emotion and sentiment lexicons.

Saif M. Mohammad. 2022b. Ethics sheet for automatic emotion recognition and sentiment analysis. *Computational Linguistics*, 48(2):239–278.

Saif M. Mohammad and Peter D. Turney. 2010. 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*, LA, California.

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

Finn Årup Nielsen. 2011. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. In *Proceedings of the ESWC Workshop on 'Making Sense of Microposts': Big things come in small packages*, pages 93–98, Heraklion, Crete.

Bryan Orme. 2009. Maxdiff analysis: Simple counting, individual-level logit, and HB. Sawtooth Software, Inc.

Charles Egerton Osgood, George J Suci, and Percy H Tannenbaum. 1957. *The measurement of meaning*. 47. University of Illinois press.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

Stanley Presser and Howard Schuman. 1996. *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context*. SAGE Publications, Inc.

Vikas C Raykar and Shipeng Yu. 2012. Eliminating spammers and ranking annotators for crowdsourced labeling tasks. *The Journal of Machine Learning Research*, 13(1):491–518.

Philip Resnik, William Armstrong, Leonardo Claudino, Thang Nguyen, Viet-An Nguyen, and Jordan Boyd-Graber. 2015. Beyond LDA: Exploring supervised topic modeling for depression-related language in Twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology:*

From Linguistic Signal to Clinical Reality, pages 99–107, Denver, Colorado.

James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

James A Russell. 2003. Core affect and the psychological construction of emotion. *Psychological review*, 110(1):145.

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

Klaus R Scherer. 1999. *Appraisal theory*. John Wiley & Sons Ltd.

Hansen Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Richard E Lucas, Megha Agrawal, Gregory J Park, Shrinidhi K Lakshmikanth, Sneha Jha, Martin EP Seligman, et al. 2013. Characterizing geographic variation in well-being using tweets. In *Seventh International AAAI Conference on Weblogs and Social Media*, pages 583–591.

Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 59–68.

Susan Standing and Craig Standing. 2018. The ethical use of crowdsourcing. *Business Ethics: A European Review*, 27(1):72–80.

Luke Stark. 2018. Algorithmic psychometrics and the scalable subject. *Social Studies of Science*, 48(2):204–231.

Philip Stone, Dexter Dunphy, Marshall Smith, and Daniel M. Ogilvie. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. The MIT Press.

Carlo Strapparava, Alessandro Valitutti, et al. 2004. Wordnet affect: an affective extension of wordnet. In *Lrec*, volume 4, page 40. Lisbon.

Daniela Teodorescu and Saif M. Mohammad. 2022. Evaluating automatically generated emotion arcs: A case for simple methods using emotion lexicons. arXiv.

Louis L. Thurstone. 1927. A law of comparative judgment. *Psychological review*, 34(4):273.

Krishnapriya Vishnubhotla and Saif M. Mohammad. 2022. Tweet emotion dynamics: Emotion word usage in tweets from us and canada. In *Proceedings of the Thirteenth International Conference on Language Resources and Evaluation (LREC 2022)*, Marseille, France.

Jane Wakefield. 2021. AI emotion-detection software tested on Uyghurs. BBC. `https://www.bbc.com/news/technology-57101248`.

Anna Wierzbicka. 1999. *Emotions across languages and cultures: Diversity and universals*. Cambridge university press.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, pages 347–354.

Lieve Van Woensel and Nissy Nevil. 2019. What if your emotions were tracked to spy on you? European Parliamentary Research Service, PE 634.415. `https://www.europarl.europa.eu/RegData/etudes/ATAG/2019/634415/EPRS_ATA(2019)634415_EN.pdf`.