Practical and Ethical Considerations in the Effective use of Emotion and Sentiment Lexicons

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

Abstract

Lexicons of word–emotion associations are widely used in research and real-world applications. As part of my research, I have created several such lexicons (e.g., the *NRC Emotion Lexicon*). This paper outlines some practical and ethical considerations involved in the effective use of these lexical resources.

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*, aka *sentiment lexicon*). 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). The lexicons have a wide range of applications in commerce, public health, and research (not just in Natural Language Processing but also in 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.

As part of my research, I have created and shared a number of word–emotion association lexicons (Mohammad et al., 2009; Mohammad and Turney, 2010; Mohammad, 2011b, 2012; Mohammad and Turney, 2013; Mohammad et al., 2013; Mohammad and Kiritchenko, 2015; Kiritchenko et al., 2014; Salameh et al., 2015; Kiritchenko and Mohammad, 2016a,b; Kiritchenko et al., 2016; Mohammad, 2018a,b).[1] Some of these lexicons such as the *NRC Emotion Lexicon* (Mohammad and Turney, 2010, 2013) and the *NRC Valence, Arousal, and Dominance (VAD) Lexicon* (Mohammad, 2018a) were manually curated by crowdsourcing the task for tens of thousands of English words.[2] Some lexicons were generated automatically from large text corpora using simple statistical algorithms. Below I list some of the practical and ethical considerations involved in the effective use of the lexicons. Many of these considerations likely apply to any emotion lexicon (and not just to the ones I have created); for example, the General Inquirer (Stone et al., 1966), ANEW (Nielsen, 2011; Bradley and Lang, 1999), LIWC (Pennebaker et al., 2001), MPQA (Wiebe et al., 2005), and the lexicon by Warriner et al. (2013).

2. CONSIDERATIONS

a. 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

---

[1]Homepage for Emotion lexicons: http://saifmohammad.com/WebPages/lexicons.html

[2]All of the manual annotation studies were approved by the NRC Research Ethics Board (NRC-REB). Protocol numbers can be found on corresponding project pages. REB review seeks to ensure that research projects involving humans as participants meet Canadian standards of ethics.

for the canonical forms (lemmas), but some also include morphological variants. The high-coverage lexicons, such as the NRC Emotion Lexicon and the NRC VAD Lexicon have most common English words. 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.

b. 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.

c. Associations/Connotations (not Denotations): A word that denotes an emotion is also associated with that emotion, but a word that is associated with an emotion does not necessarily denote that emotion. For example, *party* is associated with joy, but it does not mean (denote) joy. The lexicons capture emotion *associations*. Some have referred to such associations as *connotations or implicit emotions*.

d. 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.

e. Perceptions of emotion associations (not "right" or "correct" associations): A useful question to ask before annotating language data is whether we are looking for "correct" answers/labels or we want the annotators to tell us about how speakers of a language currently perceive the language data? For example, do I want somebody with academic training and expertise to annotate the data, or do I want to simply ask a large number of speakers of a language to get a sense of how the language data is perceived. For the lexicons that I have created, such as the NRC Emotion Lexicon and the NRC VAD Lexicon, our goal was explicitly to determine how speakers of a language perceive the emotion associations of words. Thus, the lexicons were created by crowdsourcing. We also specifically asked the annotators for 'how speakers of a language perceive the word' as apposed to 'the emotions evoked in the annotator'. The former led to less variance in responses (higher inter-annotator agreement) (Mohammad and Turney, 2013).

f. 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. See papers describing individual lexicons to determine who produced the data (people from which countries, what is the gender distribution, age distribution, etc.). An advantage of crowdsourcing is that we obtain annotations from a wider pool of annotators; however, crowd annotators are systematically different from, and not representative of, the general population.

g. Limitations of Aggregation by Majority Vote: Each instance in a lexicon (word) is usually 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, as noted above, 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 crystalized 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. I have made available the full set of disaggregated annotations, where possible. This can facilitate further work on teasing out different appropriate associations.[3] 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 next bullet).

---

[3]In case of the NRC Emotion Lexicon, the original annotations are at word-sense level. We created the word-level lexicon by keeping all majority-voted emotion associations for each of the word's senses.

h. 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. In some instances, it can be tricky to determine whether the biases are appropriate or inappropriate. Capturing the inappropriate 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 ensure that inappropriate biases are not amplified or perpetuated. If required, remove entries from the lexicons where necessary.

i. Source 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.

j. Errors in Translation: Language resources are much more common in some languages (such as English) than in most other languages. Thus often automatic translations of English resources are provided. For example, the automatic translations of the NRC Emotion Lexicon and the VAD Lexicon are provided in over 100 languages. However, automatic translations can have errors; and the number of errors can vary depending on the language pair. Also, as noted earlier, there can be cultural differences in the emotion associations of a concept. That said, several studies have shown that most emotion associations are fairly consistent across many language pairs (Redondo et al., 2007; Moors et al., 2013; Schmidtke et al., 2014; Mohammad et al., 2016; Sianipar et al., 2016; Yu et al., 2016; Stadthagen-Gonzalez et al., 2017).

k. Relative (not Absolute): The absolute values of the association scores themselves 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 valence score. Further, lexicons such as the VAD Lexicon that are created by comparative annotations (asking annotators to provide relative ordering among a small number of items at a time) do not claim that the mid-point valence score separates positive words from negative words. As stated earlier, the only claim is that terms with higher scores are associated with more (valence/arousal/dominance/etc.) than terms with lower scores.

3. PRO-TIPS

Here are some tips for effective and appropriate use of emotion lexicons:

1. 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.).

2. Depending on your specific use case, you may choose to re-scale the scores from 0 to 1 to -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. in terms of how much weight should be assigned to terms with scores close to 0.

3. 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).

4. When the lexicons provide real-valued association scores (say between 0 and 1 for valence), one can calculate average valence of the words in the target text OR separate scores indicating average score of high-valence (positive) words and average score of low-valence (negative) words. You can use '< 0.33' or '> 0.67' as rough thresholds to determine low- and high-valence words.[4]

---

[4]Other thresholds such as '< 0.2' and '> 0.8', or '< 0.5' and '> 0.5' are also suitable. One may even determine these threshold by manually examining the lexicon entries.

5. If you are 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%.

   (b) Comparative analysis is your friend. Often, emotion word counts on their own are not very 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) Inferences drawn from lager amounts of text are often more reliable than those drawn from small amounts of text. For example, *'the use of anger words grew by 20%'* is informative when determined from hundreds, thousands, tens of thousands, or more instances. Do not draw inferences about a single sentence or utterance from the emotion associations of its constituent words.

See these papers for examples of how some of these lexicons can be used (Mohammad, 2011a; Fraser et al., 2019; Hipson and Mohammad, 2020; Mendelsohn et al., 2020). See proceedings of resent shared tasks on emotions such as SemEval-2018 Task 1: Affect in Tweets (Mohammad et al., 2018) for how emotion lexicons are used in combination with training data and the latest machine learning techniques for emotion recognition.[5] You may also be interested in the applications compiled in this blog post: Ten Years of the NRC Word-Emotion Association Lexicon.[6]

4. CONCLUDING REMARKS

Emotion lexicons can be 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. I listed some considerations above that can help mitigate such unwanted outcomes. However, these are not meant to be comprehensive, but rather a jumping off point for further thought. The author welcomes feedback; including additional points to consider and include in this document. See also the papers associated with the lexicons for information about how the lexicons were created, their intended use, and their provenance. There is detailed information about each dataset in the research paper introducing it (often spanning 4 to 8 pages of the article). See Mohammad (2020) (Section 10) for a brief discussion of the broad societal impacts of emotion and sentiment analysis.

References

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.

Kathleen C. Fraser, Frauke Zeller, David Harris Smith, Saif Mohammad, and Frank Rudzicz. 2019. How do we feel when a robot dies? emotions expressed on Twitter before and after hitchBOT's destruction. In *Proceedings of the Tenth Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 62–71, Minneapolis, USA.

Will Hipson and Saif M. Mohammad. 2020. PoKi: A large dataset of poems by children. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1578–1589, Marseille, France.

Svetlana Kiritchenko and Saif M. Mohammad. 2016a. The effect of negators, modals, and degree adverbs on sentiment composition. In *Proceedings of the Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA)*.

Svetlana Kiritchenko and Saif M. Mohammad. 2016b. Happy accident: A sentiment composition lexicon for opposing polarity phrases. In *Proceedings of 10th edition of the the Language Resources and Evaluation Conference (LREC)*, Portorož, Slovenia.

---

[5] https://competitions.codalab.org/competitions/17751

[6] https://medium.com/@nlpscholar/ten-years-of-the-nrc-word-emotion-association-lexicon-eaa47a8dd03e

Svetlana Kiritchenko, Saif M. Mohammad, and Mohammad Salameh. 2016. Semeval-2016 task 7: Determining sentiment intensity of English and Arabic phrases. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '16, San Diego, California.

Svetlana Kiritchenko, Xiaodan Zhu, and Saif M. Mohammad. 2014. Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, 50:723–762.

Julia Mendelsohn, Yulia Tsvetkov, and Dan Jurafsky. 2020. A framework for the computational linguistic analysis of dehumanization. *arXiv preprint arXiv:2003.03014*.

Saif Mohammad. 2011a. From once upon a time to happily ever after: Tracking emotions in novels and fairy tales. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 105–114, Portland, OR, USA. Association for Computational Linguistics.

Saif M. Mohammad. 2011b. Even the abstract have colour: Consensus in word–colour associations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Portland, OR, USA.

Saif M. Mohammad. 2012. #Emotional tweets. In *Proceedings of the Joint Conference on Lexical and Computational Semantics*, pages 246–255, Montréal, Canada.

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

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

Saif M. Mohammad. 2020. Sentiment analysis: Detecting valence, emotions, and other affectual states from text. ArXiv:cs.CL/2005.11882.

Saif M. Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 Task 1: Affect in tweets. In *Proceedings of International Workshop on Semantic Evaluation (SemEval-2018)*, New Orleans, LA, USA.

Saif M. Mohammad, Cody Dunne, and Bonnie Dorr. 2009. Generating high-coverage semantic orientation lexicons from overtly marked words and a thesaurus. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP-2009)*, pages 599–608, Singapore.

Saif M. Mohammad and Svetlana Kiritchenko. 2015. Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence*, 31(2):301–326.

Saif M. Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proceedings of the International Workshop on Semantic Evaluation*, Atlanta, Georgia, USA.

Saif M. Mohammad, Mohammad Salameh, and Svetlana Kiritchenko. 2016. How translation alters sentiment. *Journal of Artificial Intelligence Research*, 55:95–130.

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.

Agnes Moors, Jan De Houwer, Dirk Hermans, Sabine Wanmaker, Kevin Van Schie, Anne-Laura Van Harmelen, Maarten De Schryver, Jeffrey De Winne, and Marc Brysbaert. 2013. Norms of valence, arousal, dominance, and age of acquisition for 4,300 Dutch words. *Behavior research methods*, 45(1):169–177.

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.

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.

Jaime Redondo, Isabel Fraga, Isabel Padrón, and Montserrat Comesaña. 2007. The Spanish adaptation of ANEW (affective norms for English words). *Behavior research methods*, 39(3):600–605.

Mohammad Salameh, Saif M Mohammad, and Svetlana Kiritchenko. 2015. Sentiment after translation: A case-study on Arabic social media posts. In *Proceedings of the North American Chapter of Association of Computational Linguistics*, Denver, Colorado.

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

Agnes Sianipar, Pieter van Groenestijn, and Ton Dijkstra. 2016. Affective meaning, concreteness, and subjective frequency norms for indonesian words. *Frontiers in psychology*, 7:1907.

Hans Stadthagen-Gonzalez, Constance Imbault, Miguel A Pérez Sánchez, and Marc Brysbaert. 2017. Norms of valence and arousal for 14,031 Spanish words. *Behavior research methods*, 49(1):111–123.

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

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

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2-3):165–210.

Liang-Chih Yu, Lung-Hao Lee, Shuai Hao, Jin Wang, Yunchao He, Jun Hu, K Robert Lai, and Xuejie Zhang. 2016. Building Chinese affective resources in valence-arousal dimensions. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 540–545.