Hell Hath No Fury? Correcting Bias in the NRC Emotion Lexicon

Samira Zad, Joshuan Jimenez, & Mark A. Finlayson
Knight Foundation School of Computing and Information Sciences
Florida International University
11200 SW 8th Street, Miami, FL
{szad001,jjime178,markaf}@fiu.edu

Abstract

There have been several attempts to create an accurate and thorough emotion lexicon in English, which identifies the emotional content of words. Of the several commonly used resources, the NRC emotion lexicon (Mohammad and Turney, 2013b) has received the most attention due to its availability, size, and its choice of Plutchik's expressive 8-class emotion model. In this paper we identify a large number of troubling entries in the NRC lexicon, where words that should in most contexts be emotionally neutral, with no affect (e.g., *lesbian*, *stone*, *mountain*), are associated with emotional labels that are inaccurate, nonsensical, pejorative, or, at best, highly contingent and context-dependent (e.g., *lesbian* labeled as DISGUST and SADNESS, *stone* as ANGER, or *mountain* as ANTICIPATION). We describe a procedure for semi-automatically correcting these problems in the NRC, which includes disambiguating POS categories and aligning NRC entries with other emotion lexicons to infer the accuracy of labels. We demonstrate via an experimental benchmark that the quality of the resources is thus improved. We release the revised resource and our code to enable other researchers to reproduce and build upon results[1].

1 Introduction

Emotion detection is an NLP task that has long been of interest to the field (Hancock et al., 2007; Danisman and Alpkocak, 2008; Agrawal and An, 2012), and is usually conceived as a single- or multi-label classification in which zero (or more) emotion labels are assigned to variously defined semantic or syntactic subdivisions of the text. The importance of this task has only grown as the amount of available affective text has increased: social media, in particular, has made it especially convenient for people around the world to express their feelings and emotions regarding events large and small.

There are generally two ways to express emotions in textual data (Al-Saqqa et al., 2018). First, emotions can be expressed using *emotive* vocabulary: words directly referring to emotional states (*surprise*, *sadness*, *joy*). Second, emotions can be expressed using *affective* vocabulary: words whose emotional content depends on the context, without direct reference to emotional states, for example, interjections (*ow!*, *ouch!*, *ha-ha!*).

An *emotion lexicon* is a specific type of linguistic resource that maps the emotive or affective vocabulary of a language to a fixed set of emotion labels (e.g. Plutchik's eight-emotion model), where each entry in the lexicon associates a word with zero or more emotion labels. Because this information is difficult to find elsewhere, emotion lexicons are often used as one of the key components of affective text mining systems (Yadollahi et al., 2017). However, as is usual with linguistic resources, creating an emotion lexicon is a time-consuming, costly, and sometimes impractical part of the task. The difficulty is only accentuated when one considers the many affective uses of words, in which the emotional content is context dependent. Such context dependency underlines the utility of General-Purpose (context-independent) Emotion Lexicons (GPELs), which captures the mostly fixed emotive content of words, and which can serve as a foundation for more context-dependent systems.

In this paper, we analyze and improve one of the most commonly used GPELs, namely, the NRC lexicon (National Research Council of Canada; also known as the Emolex emotion lexicon Mohammad and Turney, 2013b,a, 2010). The NRC used Macquarie's Thesaurus (Bernard, 1986) as the source for terms, retaining only words that are repeated more than 120,000 times in Google n-gram corpus (Michel et al., 2011). The NRC maps each word to zero or more labels drawn from Plutchik's 8-

[1] https://doi.org/10.34703/gzx1-9v95/PO3YGX

Proceedings of the Fifth Workshop on Online Abuse and Harms, pages 102–113
August 6, 2021.

emotion psychological model (Plutchik, 1980), and provides labels for 14,182 individual words.

While the NRC has been used extensively across the emotion mining literature (Tabak and Evrim, 2016; Abdaoui et al., 2017; Rose et al., 2018; Lee et al., 2019; Ljubešić et al., 2020; Zad et al., 2021), close inspection reveals a large number of incorrect, non-sensical, pejorative, or otherwise troubling entries. While we provide more examples later in the paper, to give a flavor of the problem, the NRC provides emotion labels for many generic nouns (*tree*→ANGER), common verbs (*dance*→TRUST), colors (*white*→ANTICIPATION), places (*mosque*→ANGER), relations (*aunt*→TRUST), and adverbs (*scarcely*→SADNESS). Furthermore, the NRC suffers from significant ambiguity because it does not include part of speech categories for the terms: for example, while *console* implies SADNESS in its most common verb sense (as the NRC indicates), in its most common noun sense means a small side table, which probably should have no emotive content. In our analysis, many of these problematic entries seem to stem from a conflation of *emotive* (context-independent) and *affective* (context-dependent) emotion language use: it is as if, during the annotation of Shakespeare's *Macbeth*, the annotators of the NRC marked *hell*→ANGER and *woman*→ANGER because of the bard's highly contextualized statement "Hell hath no fury like a woman scorned": while it is true that this statement is often cited to support an assertion that women are angry people in general, and such a lexicon entry would help in correct marking of the affective implication of this specific statement in this particular context, it does not generalize to all, or even most, uses of the word *woman*. Therein lies the rub.

We begin the paper with a brief review of psychological models of emotion, available emotion lexicons, and datasets of emotion labeled text (§2). We then discuss in detail the deficiencies of the NRC, giving a variety of problematic examples, and speculating as to how these entries were included (§3). Next we describe a semi-automatic procedure designed to filter out many of these deficiencies (§4), after which we evaluate the effectiveness of the filtering procedure by integrating the corrected version of the NRC into an emotion detection system (§5). We conclude with a list of our contributions (§6).

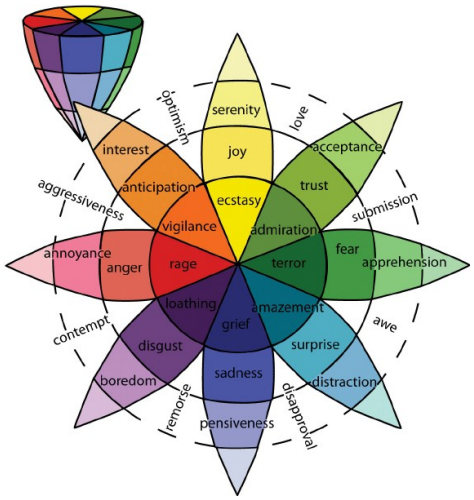

Figure 1: Plutchik's emotions wheel, Plutchik and Conte (1997). Figure taken from (Maupome and Isyutina, 2013), with permission.

2 Literature Review

2.1 Psychological Emotion Model

Emotion detection tasks are fundamentally predicated on a particular conception of what emotions exist. There are three broad classes of psychological theories of emotion: *discrete*, *dimensional*, and *hybrid*. Discrete psychological models represent basic emotions as individual, distinct categories, e.g., Oatley and Johnson-Laird (1987) with five emotions, Ekman (1992); Shaver et al. (1987) with six, Parrott (2001) with six basic emotions in the first level of a tree structure, Panksepp et al. (1998) with seven emotions, and Izard (2007) with ten. Dimensional psychological models, in contrast, propose that emotions are best described as lying in multi-dimensions space, e.g., the models of Russell (1980); Scherer (2005); Cambria (2016) with two dimensions, (Lövheim, 2012) with three, and Ortony et al. (1990); Fontaine et al. (2007); Cambria et al. (2012) with four. Hybrid models combine the two approaches by arranging categorical emotions along various dimensions, e.g., Plutchik's model (1980; 1984; 2001) with eight basic emotion categories (ANGER, FEAR, SADNESS, JOY, DISGUST, TRUST, SURPRISE, ANTICIPATION) arranged in two or three dimensions, as illustrated in Figure 1.

2.2 Emotion Lexicons

Emotion lexicons take a specific emotional theory and associate the labels or values in that theory with specific lexical entries. If the emotion lexi-

Author, Year	Lexicon	Size (words)	Set of Emotions
(Mohammad and Turney, 2010)	NRC	14,182	ANGER, FEAR, ANTICIPATION, TRUST, SURPRISE, SADNESS, JOY, DISGUST
(Mohammad and Turney, 2013b)	NRC hashtag	16,862	ANGER, FEAR, ANTICIPATION, TRUST, SURPRISE, SADNESS, JOY, DISGUST
(Stone et al., 1966)	General Enquirer	11,788	PLEASURE, AROUSAL, FEELING, PAIN
(Strapparava and Valitutti, 2004)	WordNet Affect	289	A HIERARCHY OF EMOTIONS
(Pennebaker et al., 2007)	LIWC	2,300	AFFECTIVE OR NOT, POSITIVE, NEGATIVE, ANXIETY, ANGER, SADNESS

Table 1: Comparison of emotion lexicons

con identifies emotive and affective uses tied to a specific context, then it is referred to as a *domain-specific emotion lexicon* (DSEL). In contrast, an emotion lexicon that seeks to represent the context-independent emotional meaning of words is referred to as a General Purpose Emotion Lexicon (GPEL). There are a variety of GPELs available, which we describe below.

The NRC lexicon (National Research Council of Canada; also known as the Emolex emotion lexicon) is one of the most commonly used lexicons. It comprises 14,182 words labeled according to Plutchik's psychological model. The NRC was created via a crowd-sourcing, and used Roget's Thesaurus as the source for terms (Mohammad and Turney, 2010, 2013a,b). Because we focus on the NRC lexicon in this paper, we discuss it in detail in the next section.

The WordNet Affect Lexicon (WNA or WAL Strapparava and Valitutti, 2004) is an emotion lexicon based on WordNet (Fellbaum, 1998). WNA arranges 289 noun synsets into an emotion hierarchy and associates 1,191 verbs, adverbs, and adjectives synsets to those emotion terms to WordNet.

NRC Hashtag Emotion Lexicon (Mohammad, 2012) comprises 16,862 words, drawn from Twitter hashtags, that are labeled with a strength of association (from 0 to infinity) for each of six emotion classes. It was created automatically by extracting tweets that contains `#joy`, `#sadness`, `#surprise`, `#disgust`, `#fear`, and `#anger`. Mohammad (2012) showed that the NRC Hashtag emotion lexicon provides better performance on Twitter Emotion Corpus than the WordNet-Affect emotion lexicon, but not as good as the original NRC emotion lexicon. Mohammad and Kiritchenko (2015) extended this work by expanding the hashtag word list to 585 emotion words, producing 15,825 labeled entries, with performance on headline data set again better than WNA.

The General Enquirer lexicon, while not specifically designed as an emotion lexicon, comprises 11,788 concepts labeled with 182 category labels that includes certain affect categories (e.g., pleasure, arousal, feeling, and pain) in addition to positive/negative semantic orientation for concepts (Stone et al., 1966).

Linguistic Inquiry and Word Count (LIWC Pennebaker et al., 2001, 2007) is a text analysis program that includes a lexicon comprising 2,300 entries spread across 73 categories, many of which are emotive or have sentiment, including NEGATION, ANGER, ANXIETY, SADNESS, etc.

There are lexicons which are related to emotion, but not themselves emotion lexicons. For example, Staiano and Guerini (2014) described the DepecheMood lexicon, which was an automatically generated, general-purpose, and mood lexicon with 37K terms. It includes eight mood-related labels (*don't care*, *amused*, *annoyed*, *inspired*, *anger*, *sadness*, *fear*, and *joy*) based on Rappler's mood meter (obtained by crawling the `rappler.com` social news network). Kušen et al. (2017) compared the four labels shared between NRC and DepecheMood (anger, sadness, fear, and joy), and showed that NRC had the highest recall. NRC performed better at capturing fear, anger, and joy, and DepecheMood performed better at recognize sadness. Araque et al. (2019) created the extended DepecheMood++ (DM++) for English on Rappler news and Italian on Corriere news (`corriere.it`, an online Italian newspaper).

Table 1 lists the main emotion lexicons in details. As can be seen, the NRC is one of the largest resources and uses one of the more expressive emotion ontologies, hence researchers' preference for it in their work.

2.3 Data Set

Annotated corpora of emotion-laden language go hand-in-hand with emotion lexicons. This is because one of the first tests of the utility of a lexicon is how well a system that uses the lexicon performs on automatic labeling. In general, data annotation is a crucial part of most machine learning research and affects the quality of the work substantially. As is commonly known, in the case of linguistic annotation, manually labeling large amounts of text

is expensive and time consuming; further, in most cases, assigning labels can be subjective and dependent on the personality, emotions, background, and point of view of the annotator; and finally, unbalanced label frequency creates challenges for training various learning algorithms.

There are several text corpora annotated with emotional categorical models (Yadollahi et al., 2017; Sailunaz et al., 2018; Acheampong et al., 2020). For example, the International Survey on Emotion Antecedents and Reactions (ISEAR) corpus Scherer and Wallbott (1994) comprises 7,665 sentences drawn from 3,000 students from 37 countries were asked to report as a sentense or paragraph situations in which they had experienced FEAR, SADNESS, JOY, ANGER, SHAME, GUILT, and DISGUST emotions. ISEAR data set is annotated by authors and labeled by seven emotions (FEAR, SADNESS, JOY, ANGER, SHAME, GUILT, and DISGUST). Similarly, Aman's corpus Aman and Szpakowicz (2007) comprises of 1,466 sentences from blogs and labeled by seven emotions (SADNESS, SURPRISE, ANGER, FEAR, DISGUST, HAPPINESS, and MIXED EMOTIONS). The Semantic Evaluations (SemEval) corpus (Rosenthal et al., 2019) includes 1,250 news headlines and labeled by Ekman's six basic emotions (ANGER, DISGUST, SURPRISE, FEAR, JOY, and SADNESS). These are just three examples of many.

For evaluation we use Alm's fairy tale corpus (Alm, 2008, 2010) which contains 15,302 sentences from 176 children's fairy tales from classic collections by Beatrix Potter, the Brother's Grimm's, and Hans C. Andersen. We chose this corpus because of the ready availability of an emotion detection system (Zad and Finlayson, 2020) that uses this corpus for evaluation. Two annotators marked both the emotion and mood of each sentence in the corpus (i.e., two separate judgements by both annotators, for a total of four labels per sentence), using Ekman's six emotions (JOY, FEAR, SADNESS, SURPRISE, ANGER, and DISGUST). 1,167 sentences in the corpus had "high annotation agreement" which Alm defined as all four labels being the same, and there are a total of 4,627 other sentences which annotators have all labeled them as neutral. One reason to focus on only the high agreement sentences is because the overall Cohen's Kappa for the dataset agreement is a quite poor -0.2086. If we focus only on high agreement, the Cohen's Kappa is perfect. Emotion annotation is notoriously difficult, and very few emotion annotation projects have achieved high agreement. This suggests that most of the approaches to emotion annotation have suffered from lack of conceptual clarity.

3 Problems with the NRC

In our close inspection of the entries in the NRC, we noted three main problems. First, the NRC does not indicate the part of speech of terms labeled with emotion. This obviously causes a great deal of ambiguity as to whether a particular emotion label should apply to a particular use of a word form. Second, the NRC contains numerous incorrect, inaccurate, nonsensical, or pejorative associations, most of which can be ascribed to an apparent conflation of the distinction between emotive and affective emotional language, i.e., ignoring the importance of context for emotional semantics. Third, and finally, there are emotion markings in the lexicon for which we can find no support in Keyword-in-Context (KWIC) databases for any sense; we count these as simple errors.

3.1 Missing Parts of Speech

As Mohammad and Turney (2010) noted, the NRC includes some of the most frequent English nouns, verbs, adjectives, and adverbs. Problematically, however, the NRC does not indicate the part of speech for any entry. For example, the wordform *bombard* is labeled as ANGER|FEAR; however, in WordNet the gloss for the first sense of *bombard* as a noun is "a large shawm[2] ; the bass member of the shawm family". On the other hand, the gloss of the first sense of the verb form of *bombard* is "cast, hurl, or throw repeatedly with some missile", which is more compatible with the emotion ANGER|FEAR. Another example is the word *console*. The NRC marks *console*→SADNESS, but the primary sense of the noun form refers to "a small table fixed to a wall or designed to stand against a wall." Clearly there is no context-independent emotional inflection to this sense. The SADNESS label is more appropriate for the first verb sense "to give moral or emotional strength to", usually to a sad person.

Despite Araque et al. (2019) claims that adding POS tags to lexicons may decrease the performance of emotion detection mechanisms, we observe that lack of POS tagging has caused considerable ambi-

[2] a *shawm* is a type of musical intstrument

guity which negatively affects our emotion detection system performance.

Table 2 lists a small selection of NRC word-form labels that are problematic because of part-of-speech-related ambiguity. Although we did not count the number of NRC entries suffering this particular part-of-speech ambiguity problem, our best guess is that it affects roughly several thousand entries, about a third of the non-neutral portion of the lexicon.

3.2 Context Dependency

In general-purpose emotion lexicons (GPELs), words are generally marked with an emotion (one or more labels) if there is a dominant sense of the word, and it has emotion semantics. In domain-specific emotion lexicons (DSELs), by contrast, assignment of an emotion label is based on the common sense of each term in a specific domain (Bandhakavi et al., 2017). For example, the noun "shot" in a DSEL tailored for *sports*, referring taking a shot at a goal, might be plausibly marked as (*shot*→ANTICIPATION|JOY), while in a medical DSEL, referring to a injection, might be marked as (*shot*→ANTICIPATION|FEAR). Similarly, the adjective "crazy" in sports might be marked according to the sense in the statement "that goal was crazy!" (*crazy*→JOY|SURPRISE) while in the behavioral domain, it might be (*crazy*→DISGUST|FEAR). Table 3 gives a small selection of NRC entries where each label is appropriate only in a limited context, not corresponding to the literal meaning of the word in its dominant sense. The extreme version of this problem can be seen with words like *abundance* which have a multitude of labels that conflict (DISGUST|JOY|TRUST|ANTICIPATION). Overall this is a problem with regards to NRC because it is explicitly presented as a GPEL. In our evaluation of the NRC, while again we did not count exactly how many entries suffered from this issue, we estimate at least 600 or so entries, or 10% of the NRC, fall into this category.

3.3 Simple Errors

The NRC has a large number of terms, and as with any resource of this size there are bound to be minor faults or errors. Since human annotators provided the data needed to create the resource, we can assume that certain terms were given labels that are not appropriate and that some small number of these errors would have escaped notice of any manual error correcting procedures. We define these sorts of errors as those where the provided emotional labels do not make sense in any context supported by Keyword-in-Context (KWIC) indicies (iWeb, 2021; Davies and Kim, 2019). Table 4 lists a small selection of examples of seemingly simple errors in labels, for example *architecture*→TRUST. Some markings, furthermore, might be reflective of relatively obvious biases, which in light of recent work demonstrating the built-in biases of various AI and NLP resources (Bolukbasi et al., 2016; Bender and Friedman, 2018; Mehrabi et al., 2019; Blodgett et al., 2020), it would be good to try to correct for. Examples of the latter case include the entries *fat*→DISGUST|SADNESS, *lesbian*→DISGUST|SADNESS, or *mosque*→ANGER. We estimate that the number of entries affected by simple errors or biases is at least a few hundred, or roughly 5% of the NRC.

3.4 Problems with the NRC Annotation Process

Some aspects of the NRC annotation process go part of the way toward explaining some of the above problems. As discussed by Mohammad and Turney (2013a), the annotation process relied upon approximately 2,000 native and fluent speakers of English who answered a series of questions regarding the emotion terms. The directions were made ambiguous on purpose to minimize biasing the subject's judgements. The concern with this method is that the annotators could have been shown a term that is not familiar to them. This was circumvented by asking the individual to associate the term with a certain word similar in meaning amongst three non-viable options.

After selecting the most similar word, the annotator could continue annotating even when they do not really know the meaning of a word. This could have happened by the annotator quickly looking up the definition online. The annotators were told not to look up the words[3], but there is no guarantee that they did so, and much work has shown that crowdworkers are often unreliable (Ipeirotis et al., 2010; Vuurens et al., 2011).

Another concern with the annotation process was question wording. Questions 4–11 in particular raise specific concerns. These asked, for all combinations of a term X and each of the eight emotions Y, "How much is X associated with the emotion

[3] Annotators were instructed "please skip HIT if you do not know the meaning of the word"

Word	POS	Original NRC Labels	First Sense in WordNet	Corrected Label
awful	RB	ANGER|DISGUST|FEAR|SADNESS	used as a verbal intensifier	NEUTRAL
belt	NN	ANGER|FEAR	endless loop of flexible material between two rotating shafts or pulleys	NEUTRAL
bias	JJ	ANGER	slanting diagonally across the grain of a fabric	NEUTRAL
bloody	RB	ANGER|DISGUST|FEAR|SADNESS	extremely	NEUTRAL
board	VB	ANTICIPATION	get on board of (trains, buses, ships, aircraft, etc.)	NEUTRAL
boil	VB	DISGUST	come to the boiling point and change from a liquid to vapor	NEUTRAL
buffet	NN	ANGER	a piece of furniture that stands at the side of a dining room; has shelves and drawers	NEUTRAL
bully	JJ	ANGER|FEAR	very good	SURPRISE|JOY
cage	NN	SADNESS	an enclosure made or wire or metal bars in which birds or animals can be kept	NEUTRAL
case	NN	FEAR|SADNESS	an occurrence of something	NEUTRAL
collateral	JJ	TRUST	descended from a common ancestor but through different lines	NEUTRAL
console	NN	SADNESS	a small table fixed to a wall or designed to stand against a wall	NEUTRAL
desert	NN	ANGER|DISGUST|FEAR|SADNESS	arid land with little or no vegetation	NEUTRAL
kind	NN	JOY|TRUST	a category of things distinguished by some common characteristic or quality	NEUTRAL
rail	NN	ANTICIPATION|ANGER	a barrier consisting of a horizontal bar and supports	NEUTRAL

Table 2: Examples of NRC terms paired with parts of speech (first two columns) whose emotional labels in NRC are inappropriate. The last column shows the proposed correction.

Term	NRC Labels	Term	NRC Labels
abundance	DISGUST|JOY |TRUST|ANTICIPATION	monk	TRUST
		oblige	TRUST
baby	JOY	recreation	JOY|ANTICIPATION
count	TRUST	remedy	JOY
create	JOY	remove	ANGER|FEAR |SADNESS
explain	TRUST		
fact	TRUST	saint	ANTICIPATION|JOY |TRUST|SURPRISE
fall	SADNESS		
fee	ANGER	save	JOY
fire	FEAR	score	ANTICIPATION|JOY |SURPRISE
gain	JOY|ANTICIPATION		
grow	ANTICIPATION|JOY|TRUST	star	ANTICIPATION|JOY |TRUST
larger	DISGUST|SURPRISE|TRUST		
leader	TRUST	understand	TRUST
mate	TRUST	unnatural	DISGUST|FEAR

Table 3: Examples of context dependency

Term	Labels	Term	Labels
abacus	TRUST	cabinet	TRUST
alb	TRUST	calculation	ANTICIPATION
ambulance	FEAR|TRUST	coyote	FEAR
ammonia	DISGUST	critter	DISGUST
anaconda	DISGUST|FEAR	crypt	FEAR|SADNESS
aphid	DISGUST	fat	DISGUST|SADNESS
archaeology	ANTICIPATION	fee	ANGER
architecture	TRUST	iron	TRUST
assembly	TRUST	lamb	JOY|TRUST
association	TRUST	mill	ANTICIPATION
asymmetry	DISGUST	mountain	ANTICIPATION
atherosclerosis	FEAR|SADNESS	mosque	ANGER
baboon	DISGUST	machine	TRUST
backbone	ANGER|TRUST	organ	ANTICIPATION|JOY
balm	ANTICIPATION|JOY	pine	SADNESS
basketball	ANTICIPATION|JOY	rack	SADNESS
bee	ANGER|FEAR	ravine	FEAR
belt	ANGER|FEAR	ribbon	ANTICIPATION|JOY |ANGER
bier	FEAR|SADNESS		
biopsy	FEAR	rod	TRUST |FEAR
birthplace	ANGER	spine	ANGER
blackness	FEAR|SADNESS	stone	ANGER
bran	DISGUST	title	TRUST
infant	ANTICIPATION |FEAR|JOY |SURPRISE	tree	ANTICIPATION|JOY |DISGUST|TRUST |SURPRISE|ANGER

Table 4: Examples of simple errors.

Y?" Posing this in only the positive formulation potentially biased annotators to find confirmatory evidence. A more balanced procedure would have been to ask annotators to imagine not only how much of emotion Y was associated X, but also how much Y *wasn't* associated with X, prompting them to consider disconfirmatory evidence. Because of this confirmation bias in the collection procedure we posit that many of the terms in the NRC were associated with particular emotions even when those terms do not bring those emotions to mind when mentioned in isolation in normal usage.

Another way of addressing this bias would have been to show words in specific contexts; this avoids the need for an annotator to think up their own evidence to support their label, which may have been limited by the annotators's time, attention, creativity, or knowledge of English usage. Such an approach would no doubt have been costlier, but it perhaps would have produced higher quality labels.

When it came to validating the NRC, the authors compared their crowdsourced labels with labels from the WNA lexicon to see how close the judgements were. In the one earlier paper (Mohammad and Turney, 2013a), when the NRC had 10,000 entries, the authors reported that only 6.5% of the entries could be matched with those in WNA. Later, when the NRC was expanded to 14,182 entries, the authors did not report the percentage overlap. We measured this ourselves, and found the overlap between the full NRC and WNA is 2,328 (16%). This is a concern because this means most of the data could not be independently validated to see how accurate the annotations were, and so a majority were not subject to any rigorous or systematic quality control check.

4 Semi-Automatic Correction of the NRC

The NRC includes 14,182 entries made up of a unigram (single token wordforms) associated with a

Term	Label	Term	Label	Term	Label
arm	NEUTRAL	diversity	NEUTRAL	office	NEUTRAL
buy	NEUTRAL	endpoint	NEUTRAL	road	NEUTRAL
carrier	NEUTRAL	flat	NEUTRAL	weather	NEUTRAL
clothes	NEUTRAL	filter	NEUTRAL	yeast	NEUTRAL

Table 5: Examples of neutral words

selection of Plutchik's emotions eight (SADNESS, JOY, FEAR, ANGER, SURPRISE, TRUST, DISGUST, and ANTICIPATION), NEUTRAL, and two sentiments; as noted, no words had part of speech tags. After removing 9,719 wordforms marked neutral, examples of which are shown in Table 5, 4,463 wordforms remained. In the remainder of the paper we refer to this set as `NRC.orig`. We developed a procedure to semi-automatically correct the problems discussed in prior section. First, we assigned part-of-speech tags to entries. Second, we developed an automatic emotional word test leveraging both the original version of WNA and the larger WordNet resource. Finally, we manually checked all entries for correctness.

4.1 Assigning Part of Speech to NRC words

We began by constructing an expanded list of wordforms in NRC, each associated with a valid part of speech (POS). To determine whether a POS applied to a wordform, we looking up each wordform in WordNet under each of the main open class POS tags—Verb (VB), Adjective (JJ), Noun (NN), and Adverb (RB)—so each wordform could potentially have been associated with up to four POS tags. Every wordform was present in WordNet under at least one POS. If a WordNet sense was found for a POS, we consider that a valid tags for the wordform. After this step, our list contained has 6,166 entries of wordform-POS pairs (4,463 unique wordforms). We call this set `NRC.v1`.

4.2 Emotional Word Test

In the second step, we sought to automatically determine, on the one hand, which wordform-POS pairs likely had an emotional sense (whether emotive or affective), and on the other, pairs for which we had no direct evidence of emotional semantics. To do this, we performed the following comparisons with WNA and WordNet—if any one returned true, the pair was presumed emotional; otherwise, it was marked "unknown".

1. Is the wordform-POS pair labeled as non-neutral in WNA?
2. Is the first sense of the wordform-POS pair have a synonym labeled as non-neutral in WNA?
3. Does the WordNet gloss of the first sense of the wordform-POS pair contain words that are marked as emotional in WNA?
   (a) Find the first sense in WordNet for the wordform-POS pair.
   (b) Tokenize the gloss of the first sense.
   (c) Lemmatize the gloss.
   (d) Check if the lemmas are labeled as non-neutral in WNA.

Tokenization and lemmatization were performed with `nltk` (Loper and Bird, 2002). The above procedure identified 2,328 out of 6,166 pairs as "presumed emotional", leaving 3,838 pairs as "unknown." In the rest of this paper, we will refer to the lexicon of 2,328 pairs "presumed affective" pairs as `NRC.v2`.

4.3 Manual Checking

With NRC entries now organized as to whether or not they are presumed emotional (according to WNA or WordNet), we proceeded to manually check all entries. We used WNA only to remove the emotion label of some NRC wordforms. Since the number of synsets in WNA is 2,328 and the number of wordforms in NRC.v1 is 6,166 there must exist many wordforms that are not associated to WNA synsets and therefore will fail the Emotional Word Test. We did not rely soly on WNA when correcting bias in NRC, as we manually annotated every wordform in `NRC.v1` regardless of its Emotional Word Test result. The first two authors of the paper performed the below checks on all 6,166 entries in `NRC.v1`. We used the Cohen's Kappa metric to assess inter-annotator agreement (Landis and Koch, 1977), which we measured as 0.928, which represents near-perfect agreement. Notably, this emotion annotation task has much higher agreement than the sentence-level annotation emotion tasks discussed in Section 2.3. We suspect that this is the case for at least three reasons. First, focusing on words is an easier because sentences often have complex emotion valence: there might multiple emotions in a sentence. Second, the NRC words that are retained at this stage are clearly emotional, they are selected to be such, and so are less emotionally ambiguous than neutral words: there are no borderline cases. Finally, we defined a clear set of procedures for identifying the emotion, which were

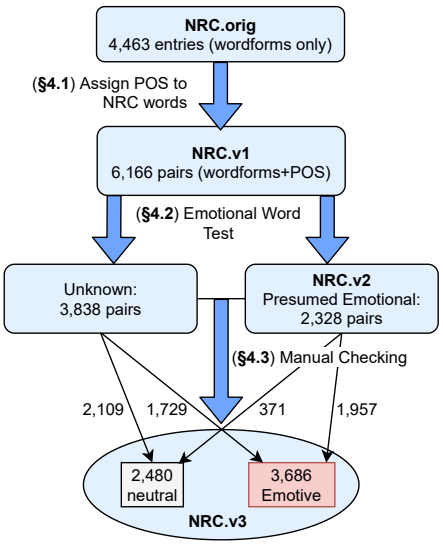

Figure 2: The semi-automatic procedure for correcting the NRC.

developed during several rounds of pilot annotation, following best practice in linguistic annotation.

- **Presumed Emotional:** For each wordform-POS pair, we examined the first sense in WordNet, any labels in WNA, and the labels in `NRC.orig` to determine if they were compatible, focusing on identified emotional words and synonyms. If there were disagreements between the WNA and `NRC.orig` we examined the Keyword-in-Context index for that POS. In cases where it was ambiguous whether `NRC.orig`, WNA, or WordNet was the correct analysis, we defaulted to `NRC.orig`. Out of 2,328 presumed emotional pairs, 1,957 were ultimately kept as having at least one emotion label.

- **Unknown:** Pairs in this group were distinguished from the Presumed Emotional group by the lack of obvious emotional words in the WordNet glosses of the pair or its synonyms. While we examined the WordNet entries for these pairs carefully, we spent more time examining the Keyword-in-Context index to look for emotional senses. Out of 3,838 unknown pairs, ultimately 1,729 were marked as having at least one emotion label.

Figure 2 shows the outline of the process to construct final, corrected version of the NRC, which we refer to as `NRC.v3` in the rest of the paper.

5 Evaluation of the Corrected Resource

In order to compare and evaluate the outcome of the correction procedure, we ran the emotion detection model developed by Zad and Finlayson (2020) using `NRC.v1`, `NRC.v2`, and `NRC.v3` as the emotion lexicon. We chose this model because the code was helpfully provided in full, and the model uses a single emotion lexicon with wordform-POS pairs to drive its emotion detection. In this section, we discuss the details of this comparison.

The emotion detection system of Zad and Finlayson originally used WNA as the emotion lexicon (leveraging wordform+POS pairs), and tested on Alm's fairy tale dataset (Alm, 2008). While the system is convenient as an experimental testbed because the full code is available, Alm's dataset uses only six emotions (ANGER, FEAR, SADNESS, SURPRISE, DISGUST, and JOY), as opposed to Plutchick's eight used by the NRC. This means we needed to trim our NRC versions down to six labels for compatibility (we dropped ANTICIPATION and TRUST). This makes the evaluation of the NRC using this experimental setup at best an approximation for the quality of our procedure. One would imagine that, if we had an experimental testbed that used all eight of Plutchik's emotions, performance would be correspondingly higher.

As described below, we also experimented with reducing the number of labels, following the experimental procedure outlined in Zad and Finlayson (2020). Further, following the same procedure, we conducted our emotion detection comparisons on the subset of Alm's dataset which represented "high agreement", namely, only sentences for which the annotators fully agreed with each other.

5.1 Comparing `NRC.v1`, `NRC.v2`, and `NRC.v3`

Table 6 shows the precision, recall and F_1 measurements of the emotion detection system when substituting the three different versions of the NRC in experimental setup for WNA, using just the six emotions present in Alm's data (dropping all the labels of ANTICIPATION and TRUST). The first three columns result gives a baseline for performance of what is effectively the original NRC in the Zad and Finlayson (2020) experimental setup.

The next two groups show `NRC.v2` and `NRC.v3`, respectively. As can be seen, overall micro-average performance rises from 0.435 for `NRC.v1` to 0.460 for `NRC.v2` and 0.484 for

	NRC.v1			NRC.v2			NRC.v3		
Emotion label	p	r	F_1	p	r	F_1	p	r	F_1
JOY	0.738	0.570	0.643	0.805	0.577	0.672	0.855	0.572	0.686
ANGER	0.359	0.253	0.297	0.347	0.226	0.274	0.432	0.240	0.308
SURPRISE	0.151	0.263	0.192	0.144	0.254	0.184	0.178	0.254	0.209
DISGUST	0.095	0.324	0.147	0.124	0.353	0.183	0.137	0.500	0.215
FEAR	0.407	0.212	0.279	0.589	0.200	0.299	0.535	0.327	0.406
SADNESS	0.632	0.417	0.502	0.661	0.473	0.552	0.717	0.451	0.553
macro-Avg.	0.397	0.340	0.343	0.445	0.347	0.361	0.476	0.391	**0.396**
micro-Avg.	0.466	0.408	0.435	0.510	0.418	0.460	0.545	0.435	**0.484**

Table 6: Result of using different, corrected versions of the NRC to the Zad and Finlayson (2020) emotion detection system on Alm's fairy tales.

	w SURPRISE			w/o SURPRISE			Avg.
	(1) w/ DISGUST	(2) w/o DISGUST	(3) DISGUST+ANGER	(4) w/ DISGUST	(5) w/o DISGUST	(6) DISGUST+ANGER	
NRC.v1	0.343	0.421	0.402	0.421	0.533	0.513	0.439
NRC.v2	0.361	0.439	0.429	0.451	0.573	0.551	0.467
NRC.v3	0.396	0.462	0.463	0.489	0.594	0.583	0.498
NRC.v1	0.435	0.481	0.461	0.545	0.603	0.577	**0.517**
NRC.v2	0.460	0.505	0.491	0.585	0.644	0.622	**0.551**
NRC.v3	0.484	0.520	0.517	0.607	0.655	0.637	**0.570**

Table 7: Comparing the macro-average (top three rows) and micro-average (bottom three rows) F_1-scores of using the three corrected versions of NRC with Zad and Finlayson's emotion detection system on Alm's fairy tales using different emotion label sets.

NRC.v3. This provides solid evidence that our correction procedure improved the quality of the resource.

While one might expect that the recall in Table 6 might strictly go down moving from NRC.v1 to NRC.v3, because we are removing terms, we are in fact correcting labels continuously in these revisions, which results in an improvement in recall and overall performance.

5.2 Varying the Label Sets

Alm's "high agreement" dataset only contains 148 sentences with DISGUST and SURPRISE labels, a highly imbalanced distribution. To investigate the impact of this imbalance on the results, we repeated the emotion detection experiment six times for each of the three version of the NRC, once for each of the reduced label sets shown in Table 7, which also shows how varying the label sets affects the performance of the emotion detection system for different version of the NRC. In all cases our corrected verisons of the NRC improve performance, anywhere from 5.3 to 7 points of F_1.

6 Contributions

We noted three categories of error in the popular NRC emotion lexicon, including a large number of seemingly biased entries. We developed and applied a semi-automatic procedure to generate three different corrected version of the NRC, and showed via experiment that these new versions improved the performance of an existing emotion-lexicon-based emotion detection system. This work shows the utility of careful error checking of lexical resources, especially with attention to correcting for unintended biases. Finally, we release the revised resource and our code to enable other researchers to reproduce and build upon results[4].

Acknowledgments

This material is based upon work supported by the U.S. Department of Homeland Security under Grant Award Number, 2017-ST-062-000002. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security.

References

Amine Abdaoui, Jérôme Azé, Sandra Bringay, and Pascal Poncelet. 2017. Feel: a french expanded emotion lexicon. *Language Resources and Evaluation*, 51(3):833–855.

Francisca Adoma Acheampong, Chen Wenyu, and Henry Nunoo-Mensah. 2020. Text-based emotion detection: Advances, challenges, and opportunities. *Engineering Reports*, 2(7):e12189.

[4] https://doi.org/10.34703/gzx1-9v95/PO3YGX

Ameeta Agrawal and Aijun An. 2012. Unsupervised emotion detection from text using semantic and syntactic relations. In *Proceedings of the 2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, volume 1, pages 346–353, Macau, China.

Samar Al-Saqqa, Heba Abdel-Nabi, and Arafat Awajan. 2018. A survey of textual emotion detection. In *2018 8th International Conference on Computer Science and Information Technology (CSIT)*, pages 136–142, Los Alamitos, CA.

Cecilia Ovesdotter Alm. 2010. Characteristics of high agreement affect annotation in text. In *Proceedings of the 4th Linguistic Annotation Workshop*, pages 118–122, Uppsala, Sweden.

Ebba Cecilia Ovesdotter Alm. 2008. *Affect in *Text and Speech*. Ph.D. thesis, University of Illinois at Urbana-Champaign, Urbana-Champaign, IL.

Saima Aman and Stan Szpakowicz. 2007. Identifying expressions of emotion in text. In *Proceedings of the 10th International Conference on Text, Speech and Dialogue (TSD)*, pages 196–205, Pilsen, Czech Republic.

Oscar Araque, Lorenzo Gatti, Jacopo Staiano, and Marco Guerini. 2019. DepecheMood++: A bilingual emotion lexicon built through simple yet powerful techniques. *IEEE Transactions on Affective Computing*, pages 1–1.

Anil Bandhakavi, Nirmalie Wiratunga, Deepak Padmanabhan, and Stewart Massie. 2017. Lexicon based feature extraction for emotion text classification. *Pattern Recognition Letters*, 93:133–142.

Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

J. Bernard, editor. 1986. *The Macquarie Thesaurus*. Macquarie Library, Sydney, Australia.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050*.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *arXiv preprint arXiv:1607.06520*.

Erik Cambria. 2016. Affective computing and sentiment analysis. *IEEE Intelligent Systems*, 31(2):102–107.

Erik Cambria, Andrew Livingstone, and Amir Hussain. 2012. The hourglass of emotions. In Anna Esposito, Antonietta M. Esposito, Alessandro Vinciarelli, Rüdiger Hoffmann, and Vincent Müller, editors, *Cognitive Behavioural Systems*, pages 144–157. Springer, Berlin. Published as Volume 7403, Lecture Notes in Computer Science (LNCS).

Taner Danisman and Adil Alpkocak. 2008. Feeler: Emotion classification of text using vector space model. In *Proceedings of the AISB 2008 Symposium on Affective Language in Human and Machine*, volume 1, page 53, Aberdeen, Scotland.

Mark Davies and Jong-Bok Kim. 2019. The advantages and challenges of "big data": Insights from the 14 billion word iWeb corpus. *Linguistic Research*, 36(1):1–34.

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

Christiane Fellbaum. 1998. Towards a representation of idioms in wordnet. In *Proceedings of the Workshop on the Use of WordNet in Natural Language Processing Systems*, pages 52–57, Montreal, Canada.

Johnny RJ Fontaine, Klaus R Scherer, Etienne B Roesch, and Phoebe C Ellsworth. 2007. The world of emotions is not two-dimensional. *Psychological Science*, 18(12):1050–1057.

Jeffrey T Hancock, Christopher Landrigan, and Courtney Silver. 2007. Expressing emotion in text-based communication. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 929–932, San Jose, CA.

Panagiotis G Ipeirotis, Foster Provost, and Jing Wang. 2010. Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD workshop on human computation*, pages 64–67, Washington, DC.

iWeb. 2021. The iWeb Corpus. https://www.english-corpora.org/iweb/. Last accessed on April 25, 2021.

Carroll E Izard. 2007. Basic emotions, natural kinds, emotion schemas, and a new paradigm. *Perspectives on Psychological Science*, 2(3):260–280.

Ema Kušen, Giuseppe Cascavilla, Kathrin Figl, Mauro Conti, and Mark Strembeck. 2017. Identifying emotions in social media: Comparison of word-emotion lexicons. In *Proceeedings of the 5th International Conference on Future Internet of Things and Cloud (FiCloud)*, pages 132–137, Prague, Czech Republic.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Young-Jun Lee, Chan-Yong Park, and Ho-Jin Choi. 2019. Word-level emotion embedding based on semi-supervised learning for emotional classification in dialogue. In *Proceedings of the IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 1–4, Kyoto, Japan.

Nikola Ljubešić, Ilia Markov, Darja Fišer, and Walter Daelemans. 2020. The lilah emotion lexicon of croatian, dutch and slovene. In *Proceedings of the 3rd Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 153–157, Barcelona, Spain (Online).

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. *arXiv preprint cs/0205028*.

Hugo Lövheim. 2012. A new three-dimensional model for emotions and monoamine neurotransmitters. *Medical Hypotheses*, 78(2):341–348.

Gerardo Maupome and Olga Isyutina. 2013. Dental students' and faculty members' concepts and emotions associated with a caries risk assessment program. *Journal of Dental Education*, 77(11):1477–1487.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Joseph P Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, et al. 2011. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182.

Saif Mohammad. 2012. # emotional tweets. In *Proceedings of the 1st Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 246–255, Montreal, Canada.

Saif Mohammad and Peter 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*, pages 26–34, Los Angeles, CA.

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 and Peter D Turney. 2013a. Crowdsourcing a word–emotion association lexicon. *Computational intelligence*, 29(3):436–465.

Saif M Mohammad and Peter D Turney. 2013b. Nrc emotion lexicon. *National Research Council, Canada*, 2.

Keith Oatley and Philip N Johnson-Laird. 1987. Towards a cognitive theory of emotions. *Cognition and Emotion*, 1(1):29–50.

Andrew Ortony, Gerald L Clore, and Allan Collins. 1990. *The Cognitive Structure of Emotions*. Cambridge University Press, Cambridge, UK.

Jaak Panksepp, Brian Knutson, and Douglas L Pruitt. 1998. Toward a neuroscience of emotion. In Michael F. Mascolo and Sharon Griffin, editors, *What develops in emotional development?*, pages 53–84. Springer, Berlin, Germany.

W Gerrod Parrott. 2001. *Emotions in Social Psychology: Essential Readings*. Psychology Press, London.

James W Pennebaker, Roger J Booth, and Martha E Francis. 2007. Linguistic inquiry and word count: LIWC [computer software].

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: LIWC 2001 [computer software].

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. In Robert Plutchik, editor, *Theories of Emotion*, pages 3–33. Elsevier, Amsterdam, The Netherlands.

Robert Plutchik. 1984. Emotions and imagery. *Journal of Mental Imagery*, 8:105–111.

Robert Plutchik. 2001. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist*, 89(4):344–350.

Robert Ed Plutchik and Hope R Conte. 1997. *Circumplex Models of Personality and Emotions.* American Psychological Association, Washington, DC.

S Lovelyn Rose, R Venkatesan, Girish Pasupathy, and P Swaradh. 2018. A lexicon-based term weighting scheme for emotion identification of tweets. *International Journal of Data Analysis Techniques and Strategies*, 10(4):369–380.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2019. Semeval-2017 task 4: Sentiment analysis in twitter. *arXiv preprint arXiv:1912.00741*.

James A Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178.

Kashfia Sailunaz, Manmeet Dhaliwal, Jon Rokne, and Reda Alhajj. 2018. Emotion detection from text and speech: a survey. *Social Network Analysis and Mining*, 8(1):1–26.

Klaus R Scherer. 2005. What are emotions? and how can they be measured? *Social Science Information*, 44(4):695–729.

Klaus R Scherer and Harald G Wallbott. 1994. Evidence for universality and cultural variation of differential emotion response patterning. *Journal of Personality and Social Psychology*, 66(2):310.

Phillip Shaver, Judith Schwartz, Donald Kirson, and Cary O'connor. 1987. Emotion knowledge: Further exploration of a prototype approach. *Journal of Personality and Social Psychology*, 52(6):1061–1086.

Jacopo Staiano and Marco Guerini. 2014. DepecheMood: A lexicon for emotion analysis from crowd-annotated news. *arXiv preprint arXiv:1405.1605*.

Philip James Stone, Dexter Colboyd Dunphy, Daniel M Ogilvie, and Marshall S Smith. 1966. *The General Inquirer: a Computer Approach to Content Analysis*. MIT Press, Cambridge, MA.

Carlo Strapparava and Alessandro Valitutti. 2004. Wordnet affect: An affective extension of wordnet. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, pages 1083–1086, Lisbon, Portugal.

Feride Savaroğlu Tabak and Vesile Evrim. 2016. Comparison of emotion lexicons. In *Proceedings of the 13th Interntaional Symposium on Smart Microgrids for Sustainable Energy Sources Enabled by Photonics and IoT Sensors (HONET-ICT)*, pages 154–158, Nicosia, Cyprus.

Jeroen Vuurens, Arjen P de Vries, and Carsten Eickhoff. 2011. How much spam can you take? an analysis of crowdsourcing results to increase accuracy. In *Proceedings of the ACM SIGIR Workshop on Crowdsourcing for Information Retrieval (CIR'11)*, pages 21–26, Beijing, China.

Ali Yadollahi, Ameneh Gholipour Shahraki, and Osmar R Zaiane. 2017. Current state of text sentiment analysis from opinion to emotion mining. *ACM Computing Surveys (CSUR)*, 50(2):1–33.

Samira Zad and Mark Finlayson. 2020. Systematic evaluation of a framework for unsupervised emotion recognition for narrative text. In *Proceedings of the 1st Joint Workshop on Narrative Understanding, Storylines, and Events*, pages 26–37, Online.

Samira Zad, Maryam Heidari, James H Jr Jones, and Ozlem Uzuner. 2021. Emotion detection of textual data: An interdisciplinary survey. In *Proceedings of the IEEE World AI IoT Congress (AIIoT 2021)*, Seattle, WA.