Crowdsourcing a Word–Emotion Association Lexicon

Saif M. Mohammad and Peter D. Turney

Institute for Information Technology, National Research Council Canada.
Ottawa, Ontario, Canada, K1A 0R6
{saif.mohammad,peter.turney}@nrc-cnrc.gc.ca

Even though considerable attention has been given to the polarity of words (positive and negative) and the creation of large polarity lexicons, research in emotion analysis has had to rely on limited and small emotion lexicons. In this paper we show how the combined strength and wisdom of the crowds can be used to generate a large, high-quality, word–emotion and word–polarity association lexicon quickly and inexpensively. We enumerate the challenges in emotion annotation in a crowdsourcing scenario and propose solutions to address them. Most notably, in addition to questions about emotions associated with terms, we show how the inclusion of a word choice question can discourage malicious data entry, help identify instances where the annotator may not be familiar with the target term (allowing us to reject such annotations), and help obtain annotations at sense level (rather than at word level). We conducted experiments on how to formulate the emotion-annotation questions, and show that asking if a term is *associated* with an emotion leads to markedly higher inter-annotator agreement than that obtained by asking if a term *evokes* an emotion.

Key words: Emotions, affect, polarity, semantic orientation, crowdsourcing, Mechanical Turk, emotion lexicon, polarity lexicon, word–emotion associations, sentiment analysis.

1. INTRODUCTION

We call upon computers and algorithms to assist us in sifting through enormous amounts of data and also to understand the content—for example, "What is being said about a certain target entity?" (Common target entities include a company, product, policy, person, and country.) Lately, we are going further, and also asking questions such as: "Is something good or bad being said about the target entity?" and "Is the speaker happy with, angry at, or fearful of the target?". This is the area of *sentiment analysis*, which involves determining the opinions and private states (beliefs, feelings, and speculations) of the speaker towards a target entity (Wiebe, 1994). Sentiment analysis has a number of applications, for example in managing customer relations, where an automated system may transfer an angry, agitated caller to a higher-level manager. An increasing number of companies want to automatically track the response to their product (especially when there are new releases and updates) on blogs, forums, social networking sites such as Twitter and Facebook, and the World Wide Web in general. (More applications listed in Section 2.) Thus, over the last decade, there has been considerable work in sentiment analysis, and especially in determining whether a word, phrase, or document has a *positive polarity*, that is, it is expressing a favorable sentiment towards an entity, or whether it has a *negative polarity*, that is, it is expressing an unfavorable sentiment towards an entity (Lehrer, 1974; Turney and Littman, 2003; Pang and Lee, 2008). (This sense of *polarity* is also referred to as *semantic orientation* and *valence* in the literature.) However, much research remains to be done on the problem of automatic analysis of *emotions* in text.

Emotions are often expressed through different facial expressions (Aristotle, 1913; Russell, 1994). Different emotions are also expressed through different words. For example, *delightful* and *yummy* indicate the emotion of joy, *gloomy* and *cry* are indicative of sadness, *shout* and *boiling* are indicative of anger, and so on. In this paper, we are interested in how emotions manifest themselves in language through words.[1] We describe an annotation project aimed at creating a large lexicon of term–emotion associations. A term is either a word or a phrase. Each entry in this lexicon includes a term, an emotion, and a measure of how strongly the term is associated with the emotion. Instead of

[1]This paper expands on work first published in Mohammad and Turney (2010).

providing definitions for the different emotions, we give the annotators examples of words associated with different emotions and rely on their intuition of what different emotions mean and how language is used to express emotion.

Terms may evoke different emotions in different contexts, and the emotion evoked by a phrase or a sentence is not simply the sum of emotions conveyed by the words in it. However, the emotion lexicon can be a useful component for a sophisticated emotion detection algorithm required for many of the applications described in the next section. The term–emotion association lexicon will also be useful for evaluating automatic methods that identify the emotions associated with a word. Such algorithms may then be used to automatically generate emotion lexicons in languages where no such lexicons exist. As of now, high-quality, high-coverage, emotion lexicons do not exist for any language, although there are a few limited-coverage lexicons for a handful of languages, for example, the WordNet Affect Lexicon (WAL) (Strapparava and Valitutti, 2004), the General Inquirer (GI) (Stone *et al.*, 1966), and the Affective Norms for English Words (ANEW) (Bradley and Lang, 1999).

The lack of emotion resources can be attributed to high cost and considerable manual effort required of the human annotators in a traditional setting where hand-picked experts are hired to do all the annotation. However, lately a new model has evolved to do large amounts of work quickly and inexpensively. *Crowdsourcing* is the act of breaking down work into many small independent units and distributing them to a large number of people, usually over the web. Howe and Robinson (2006), who coined the term, define it as follows:[2]

> *The act of a company or institution taking a function once performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an open call. This can take the form of peer-production (when the job is performed collaboratively), but is also often undertaken by sole individuals. The crucial prerequisite is the use of the open call format and the large network of potential laborers.*

Some well-known crowdsourcing projects include Wikipedia, Threadless, iStockphoto, InnoCentive, Netflix Prize, and Amazon's Mechanical Turk.[3]

Mechanical Turk is an online crowdsourcing platform that is especially suited for tasks that can be done over the Internet through a computer or a mobile device. It is already being used to obtain human annotation on various linguistic tasks (Snow *et al.*, 2008; Callison-Burch, 2009). However, one must define the task carefully to obtain annotations of high quality. Several checks must be placed to ensure that random and erroneous annotations are discouraged, rejected, and re-annotated.

In this paper, we show how we compiled a large English term–emotion association lexicon by manual annotation through Amazon's Mechanical Turk service. This dataset, which we call *EmoLex*, is an order of magnitude larger than the WordNet Affect Lexicon. We focus on the emotions of joy, sadness, anger, fear, trust, disgust, surprise, and anticipation—argued by many to be the basic and prototypical emotions (Plutchik, 1980). The terms in EmoLex are carefully chosen to include some of the most frequent English nouns, verbs, adjectives, and adverbs. In addition to unigrams, EmoLex has many commonly used bigrams as well. We also include words from the General Inquirer and the WordNet Affect Lexicon to allow comparison of annotations between the various resources. We perform extensive analysis of the annotations to answer several questions, including the following:

1. How hard is it for humans to annotate words with their associated emotions?
2. How can emotion-annotation questions be phrased to make them accessible and clear to the average English speaker?
3. Do small differences in how the questions are asked result in significant annotation differences?
4. Are emotions more commonly evoked by nouns, verbs, adjectives, or adverbs? How common are emotion terms among the various parts of speech?
5. How much do people agree on the association of a given emotion with a given word?
6. Is there a correlation between the polarity of a word and the emotion associated with it?
7. Which emotions tend to go together; that is, which emotions are associated with the same terms?

[2] http://crowdsourcing.typepad.com/cs/2006/06

[3] Wikipedia: http://en.wikipedia.org, Threadless: http://www.threadless.com, iStockphoto: http://www.istockphoto.com, InnoCentive: http://www.innocentive.com, Netflix prize: http://www.netflixprize.com, Mechanical Turk: https://www.mturk.com/mturk/welcome

Our lexicon now has close to 10,000 terms and ongoing work will make it even larger (we are aiming for about 40,000 terms).

2. APPLICATIONS

The automatic recognition of emotions is useful for a number of tasks, including the following:

1. Managing customer relations by taking appropriate actions depending on the customer's emotional state (for example, dissatisfaction, satisfaction, sadness, trust, anticipation, or anger) (Bougie *et al.*, 2003).
2. Tracking sentiment towards politicians, movies, products, countries, and other target entities (Pang and Lee, 2008; Mohammad and Yang, 2011).
3. Developing sophisticated search algorithms that distinguish between different emotions associated with a product (Knautz *et al.*, 2010). For example, customers may search for banks, mutual funds, or stocks that people trust. Aid organizations may search for events and stories that are generating empathy, and highlight them in their fund-raising campaigns. Further, systems that are not emotion-discerning may fall prey to abuse. For example, it was recently discovered that an online vendor deliberately mistreated his customers because the negative online reviews translated to higher rankings on Google searches.[4]
4. Creating dialogue systems that respond appropriately to different emotional states of the user; for example, in emotion-aware games (Velásquez, 1997; Ravaja *et al.*, 2006).
5. Developing intelligent tutoring systems that manage the emotional state of the learner for more effective learning. There is some support for the hypothesis that students learn better and faster when they are in a positive emotional state (Litman and Forbes-Riley, 2004).
6. Determining risk of repeat attempts by analyzing suicide notes (Osgood and Walker, 1959; Matykiewicz *et al.*, 2009; Pestian *et al.*, 2008).[5]
7. Understanding how genders communicate through work-place and personal email (Mohammad and Yang, 2011).
8. Assisting in writing e-mails, documents, and other text to convey the desired emotion (and avoiding misinterpretation) (Liu *et al.*, 2003).
9. Depicting the flow of emotions in novels and other books (Boucouvalas, 2002; Mohammad, 2011b).
10. Identifying what emotion a newspaper headline is trying to evoke (Bellegarda, 2010).
11. Re-ranking and categorizing information/answers in online question–answer forums (Adamic *et al.*, 2008). For example, highly emotional responses may be ranked lower.
12. Detecting how people use emotion-bearing-words and metaphors to persuade and coerce others (for example, in propaganda) (Kŏvecses, 2003).
13. Developing more natural text-to-speech systems (Francisco and Gervás, 2006; Bellegarda, 2010).
14. Developing assistive robots that are sensitive to human emotions (Breazeal and Brooks, 2004; Hollinger *et al.*, 2006). For example, the robotics group in Carnegie Melon University is interested in building an emotion-aware physiotherapy coach robot.

Since we do not have space to fully explain all of these applications, we select one (the first application from the list: managing customer relations) to develop in more detail as an illustration of the value of emotion-aware systems. Davenport *et al.* (2001) define *customer relationship management (CRM)* systems as:

> *All the tools, technologies and procedures to manage, improve or facilitate sales, support and related interactions with customers, prospects, and business partners throughout the enterprise.*

Central to this process is keeping the customer satisfied. A number of studies have looked at dissatisfaction and anger and shown how they can lead to complaints to company representatives,

[4] http://www.pcworld.com/article/212223/google_algorithm_will_punish_bad_businesses.html

[5] The 2011 Informatics for Integrating Biology and the Bedside (i2b2) challenge by the National Center for Biomedical Computing is on detecting emotions in suicide notes.

litigations against the company in courts, *negative word of mouth*, and other outcomes that are detrimental to company goals (Maute and Forrester, 1993; Richins, 1987; Singh, 1988). Richins (1984) defines *negative word of mouth* as:

> *Interpersonal communication among consumers concerning a marketing organization or product which denigrates the object of the communication.*

Anger, as indicated earlier, is clearly an emotion, and so is dissatisfaction (Ortony *et al.*, 1988; Scherer, 1984; Shaver *et al.*, 1987; Weiner, 1985). Even though the two are somewhat correlated (Folkes *et al.*, 1987), Bougie *et al.* (2003) show through experiments and case studies that dissatisfaction and anger are distinct emotions, leading to distinct actions by the consumer. Like Weiner (1985), they argue that dissatisfaction is an "outcome-dependent emotion", that is, it is a reaction to an undesirable outcome of a transaction, and that it instigates the customer to determine the reason for the undesirable outcome. If customers establish that it was their own fault, then this may evoke an emotion of guilt or shame. If the situation was beyond anybody's control, then it may evoke sadness. However, if they feel that it was the fault of the service provider, then there is a tendency to become angry. Thus, dissatisfaction is usually a precursor to anger (also supported by Scherer (1982); Weiner (1985)), but may often instead lead to other emotions such as sadness, guilt, and shame, too. Bougie *et al.* (2003) also show that dissatisfaction does not have a correlation with complaints and negative word of mouth, when the data is controlled for anger. On the other hand, anger has a strong correlation with complaining and negative word of mouth, even when satisfaction is controlled for (Díaz and Ruz, 2002; Dubé and Maute, 1996).

Consider a scenario in which a company has automated systems on the phone and on the web to manage high-volume calls. Basic queries and simple complaints are handled automatically, but non-trivial ones are forwarded to a team of qualified call handlers. It is usual for a large number of customer interactions to have negative polarity terms because, after all, people often contact a company because they are dissatisfied with a certain outcome. However, if the system is able to detect that a certain caller is angry (and thus, if not placated, is likely to engage in negative word of mouth about the company or the product), then it can immediately transfer the call to a qualified higher-level human call handler.

Apart from keeping the customers satisfied, companies are also interested in developing a large base of loyal customers. Customers loyal to a company buy more products, spend more money, and also spread positive word of mouth (Harris and Goode, 2004). Oliver (1997), Dabholkar *et al.* (2000), Harris and Goode (2004), and others give evidence that central to attaining loyal customers is the amount of trust they have in the company. Trust is especially important in on-line services where it has been shown that consumers buy more and return more often to shop when they trust a company (Shankar *et al.*, 2002; Reichheld and Schefter, 2000; Stewart, 2003).

Thus it is in the interest of the company to heed the consumers, not just when they call, but also during online transactions and when they write about the company in their blogs, tweets, consumer forums, and review websites so that they can immediately know whether the customers are happy with, dissatisfied with, losing trust in, or angry with their product or a particular feature of the product. This way they can take corrective action when necessary, and accentuate the most positively evocative features. Further, an emotion-aware system can discover instances of high trust and use them as sales opportunities (for example, offering a related product or service for purchase).

3. EMOTIONS

Emotions are pervasive among humans, and many are innate. Some argue that even across cultures that have no contact with each other, facial expressions for basic human emotions are identical (Ekman and Friesen, 2003; Ekman, 2005). However, other studies argue that there may be some universalities, but language and culture play an important role in shaping our emotions and also in how they manifest themselves in facial expression (Elfenbein and Ambady, 1994; Russell, 1994). There is some contention on whether animals have emotions, but there are studies, especially for higher mammals, canines, felines, and even some fish, arguing in favor of the proposition (Masson, 1996; Guo *et al.*, 2007). Some of the earliest work is by Charles Darwin in his book *The Expressions of the Emotions in Man and Animals* (Darwin, 1872). Studies by evolutionary biologists

and psychologists show that emotions have evolved to improve the reproductive fitness for a species, as they are triggers for behavior with high survival value. For example, fear inspires fight-or-flight response. The more complex brains of primates and humans are capable of experiencing not just the basic emotions such as fear and joy, but also more complex and nuanced emotions such as optimism and shame. Similar to emotions, other phenomena such as *mood* also pertain to the evaluation of one's well-being and are together referred to as *affect* (Scherer, 1984; Gross, 1998; Steunebrink, 2010). Unlike emotion, mood is not towards a specific thing, but more diffuse, and it lasts for longer durations (Nowlis and Nowlis, 2001; Gross, 1998; Steunebrink, 2010).

Psychologists have proposed a number of theories that classify human emotions into taxonomies. As mentioned earlier, some emotions are considered basic, whereas others are considered complex. Some psychologists have classified emotions into those that we can sense and perceive (*instinctual*), and those that that we arrive at after some thinking and reasoning (*cognitive*) (Zajonc, 1984). However, others do not agree with such a distinction and argue that emotions do not precede cognition (Lazarus, 1984, 2000). Plutchik (1985) argues that this debate may not be resolvable because it does not lend itself to empirical proof and that the problem is a matter of definition. There is a high correlation between the basic and instinctual emotions, as well as between complex and cognitive emotions. Many of the basic emotions are also instinctual.

A number of theories have been proposed on which emotions are basic (Ekman, 1992; Plutchik, 1962; Parrot, 2001; James, 1884). See Ortony and Turner (1990) for a detailed review of many of these models. Ekman (1992) argues that there are six basic emotions: joy, sadness, anger, fear, disgust, and surprise. Plutchik (1962, 1980, 1994) proposes a theory with eight basic emotions. These include Ekman's six as well as trust and anticipation. Plutchik organizes the emotions in a wheel (Figure 1). The radius indicates intensity—the closer to the center, the higher the intensity. Plutchik argues that the eight basic emotions form four opposing pairs, joy–sadness, anger–fear, trust–disgust, and anticipation–surprise. This emotion opposition is displayed in Figure 1 by the spatial opposition of these pairs. The figure also shows certain emotions, called *primary dyads*, in the white spaces between the basic emotions, which he argues can be thought of as combinations of the adjoining emotions. However it should be noted that emotions in general do not have clear boundaries and do not always occur in isolation.

Since annotating words with hundreds of emotions is expensive for us and difficult for annotators, we decided to annotate words with Plutchik's eight basic emotions. We do not claim that Plutchik's eight emotions are more fundamental than other categorizations; however, we adopted them for annotation purposes because: (a) like some of the other choices of basic emotions, this choice too is well-founded in psychological, physiological, and empirical research, (b) unlike some other choices, for example that of Ekman, it is not composed of mostly negative emotions, (c) it is a superset of the emotions proposed by some others (for example, it is a superset of Ekman's six basic emotions), and (d) in our future work, we will conduct new annotation experiments to empirically verify whether certain pairs of these emotions are indeed in opposition or not, and whether the primary dyads can indeed be thought of as combinations of the adjacent basic emotions.

4. RELATED WORK

Over the past decade, there has been a large amount of work on sentiment analysis that focuses on positive and negative polarity. Pang and Lee (2008) provide an excellent summary. Here we focus on the relatively small amount of work on generating emotion lexicons and on computational analysis of the emotional content of text.

The WordNet Affect Lexicon (WAL) (Strapparava and Valitutti, 2004) has a few hundred words annotated with the emotions they evoke.[6] It was created by manually identifying the emotions of a few seed words and then marking all their WordNet synonyms as having the same emotion. The words in WAL are annotated for a number of emotion and affect categories, but its creators also provided a subset corresponding to the six Ekman emotions. In our Mechanical Turk experiments, we re-annotate hundreds of words from the Ekman subset of WAL to determine how much the

[6]http://wndomains.fbk.eu/wnaffect.html

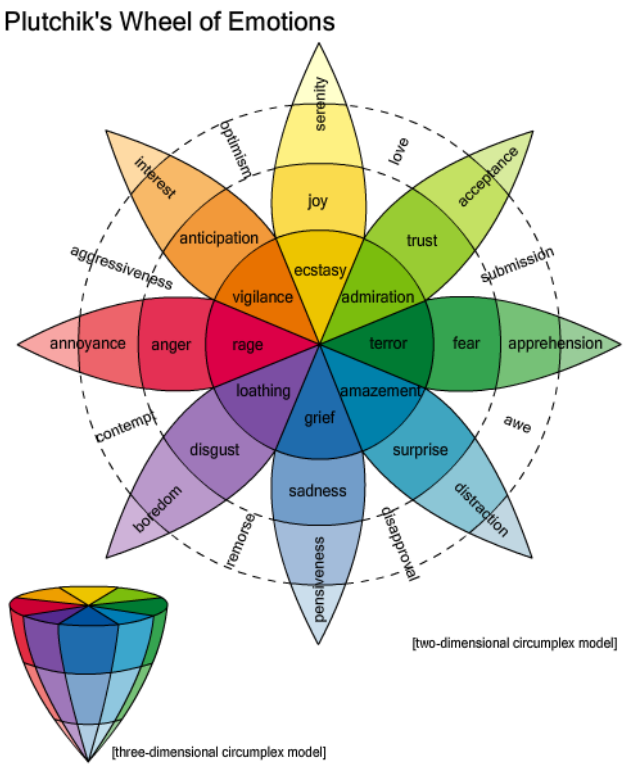

FIGURE 1. Plutchik's wheel of emotions. Similar emotions are placed next to each other. Contrasting emotions are placed diametrically opposite to each other. Radius indicates intensity. White spaces in between the basic emotions represent primary dyads—complex emotions that are combinations of adjacent basic emotions. (The image file is taken from Wikimedia Commons.)

emotion annotations obtained from untrained volunteers matches that obtained from the original hand-picked judges (Section 10). General Inquirer (GI) (Stone *et al.*, 1966) has 11,788 words labeled with 182 categories of word tags, including positive and negative semantic orientation.[7] It also has certain other affect categories, such as pleasure, arousal, feeling, and pain, but these have not been exploited to a significant degree by the natural language processing community. In our Mechanical Turk experiments, we re-annotate thousands of words from GI to determine how much the polarity annotations obtained from untrained volunteers matches that obtained from the original hand-picked judges (Section 11). Affective Norms for English Words (ANEW) has pleasure (happy–unhappy), arousal (excited–calm), and dominance (controlled–in control) ratings for 1034 words.[8]

Automatic systems for analyzing emotional content of text follow many different approaches: a number of these systems look for specific emotion denoting words (Elliott, 1992), some determine the tendency of terms to co-occur with seed words whose emotions are known (Read, 2004), some use hand-coded rules (Neviarouskaya *et al.*, 2009, 2010), and some use machine learning and a number of emotion features, including emotion denoting words (Alm *et al.*, 2005; Aman and Szpakowicz, 2007). Recent work by Bellegarda (2010) uses sophisticated dimension reduction techniques (variations of latent semantic analysis), to automatically identify emotion terms, and obtains marked improvements in classifying newspaper headlines into different emotion categories. Goyal *et al.* (2010) move away from classifying sentences from the writer's perspective, towards attributing mental states to entities

[7] http://www.wjh.harvard.edu/∼inquirer
[8] http://csea.phhp.ufl.edu/media/anewmessage.html

mentioned in the text. Their work deals with polarity, but work on attributing emotions to entities mentioned in text is, similarly, a promising area of future work.

Much recent work focuses on six emotions studied by Ekman (1992) and Sautera *et al.* (2010). These emotions—joy, sadness, anger, fear, disgust, and surprise—are a subset of the eight proposed in Plutchik (1980). There is less work on complex emotions, for example, work by Pearl and Steyvers (2010) that focuses on politeness, rudeness, embarrassment, formality, persuasion, deception, confidence, and disbelief. They developed a game-based annotation project for these emotions. Francisco and Gervás (2006) marked sentences in fairy tales with tags for pleasantness, activation, and dominance, using lexicons of words associated with the three categories.

Emotion analysis can be applied to all kinds of text, but certain domains and modes of communication tend have more overt expressions of emotions than others. Neviarouskaya *et al.* (2010), Genereux and Evans (2006), and Mihalcea and Liu (2006) analyzed web-logs. Alm *et al.* (2005) and Francisco and Gervás (2006) worked on fairy tales. Boucouvalas (2002) and John *et al.* (2006) explored emotions in novels. Zhe and Boucouvalas (2002), Holzman and Pottenger (2003), and Ma *et al.* (2005) annotated chat messages for emotions. Liu *et al.* (2003) worked on email data.

There has also been some interesting work in visualizing emotions, for example that of Subasic and Huettner (2001), Kalra and Karahalios (2005), and Rashid *et al.* (2006). Mohammad (2011a) describes work on identifying colours associated with emotion words.

5. TARGET TERMS

In order to generate a word–emotion association lexicon, we first identify a list of words and phrases for which we want human annotations. We chose the *Macquarie Thesaurus* as our source for unigrams and bigrams (Bernard, 1986).[9] The categories in the thesaurus act as coarse senses of the words. (A word listed in two categories is taken to have two senses.) Any other published dictionary would have worked well too. Apart from over 57,000 commonly used English word types, the *Macquarie Thesaurus* also has entries for more than 40,000 commonly used phrases. From this list we chose those terms that occurred frequently in the Google n-gram corpus (Brants and Franz, 2006). Specifically we chose the 200 most frequent unigrams and 200 most frequent bigrams from four parts of speech: nouns, verbs, adverbs, and adjectives. When selecting these sets, we ignored terms that occurred in more than one *Macquarie Thesaurus* category. (There were only 187 adverb bigrams that matched these criteria. All other sets had 200 terms each.) We chose all words from the Ekman subset of the WordNet Affect Lexicon that had at most two senses (terms listed in at most two thesaurus categories)—640 word–sense pairs in all. We included all terms in the General Inquirer that were not too ambiguous (had at most three senses)—8132 word–sense pairs in all. (We started the annotation on monosemous terms, and gradually included more ambiguous terms as we became confident that the quality of annotations was acceptable.) Some of these terms occur in more than one set. The union of the three sets (Google n-gram terms, WAL terms, and GI terms) has 10,170 term–sense pairs. Table 1 lists the various sets of target terms as well as the number of terms in each set for which annotations were requested. EmoLex-Uni stands for all the unigrams taken from the thesaurus. EmoLex-Bi refers to all the bigrams taken from the thesaurus. EmoLex-GI are all the words taken from the General Inquirer. EmoLex-WAL are all the words taken from the WordNet Affect Lexicon.

6. MECHANICAL TURK

We used Amazon's Mechanical Turk service as a platform to obtain large-scale emotion annotations. An entity submitting a task to Mechanical Turk is called the *requester*. The requester breaks the task into small independently solvable units called *HITs (Human Intelligence Tasks)* and uploads them on the Mechanical Turk website. The requester specifies (1) some key words relevant to the task to help interested people find the HITs on Amazon's website, (2) the compensation that will

[9] http://www.macquarieonline.com.au/thesaurus.html

TABLE 1. Break down of the target terms for which emotion annotations were requested.

EmoLex	# of terms	% of the Union
EmoLex-Uni:		
Unigrams from Macquarie Thesaurus		
adjectives	200	2.0%
adverbs	200	2.0%
nouns	200	2.0%
verbs	200	2.0%
EmoLex-Bi:		
Bigrams from Macquarie Thesaurus		
adjectives	200	2.0%
adverbs	187	1.8%
nouns	200	2.0%
verbs	200	2.0%
EmoLex-GI:		
Terms from General Inquirer		
negative terms	2119	20.8%
neutral terms	4226	41.6%
positive terms	1787	17.6%
EmoLex-WAL:		
Terms from WordNet Affect Lexicon		
anger terms	165	1.6%
disgust terms	37	0.4%
fear terms	100	1.0%
joy terms	165	1.6%
sadness terms	120	1.2%
surprise terms	53	0.5%
Union	**10170**	**100%**

be paid for solving each HIT, and (3) the number of different annotators that are to solve each HIT. The people who provide responses to these HITs are called *Turkers*. Turkers usually search for tasks by entering key words representative of the tasks they are interested in and often also by specifying the minimum compensation per HIT they are willing to work for. The annotation provided by a Turker for a HIT is called an *assignment*.

We created Mechanical Turk HITs for each of the terms specified in Section 5. Each HIT has a set of questions, all of which are to be answered by the same person. (A complete example HIT with directions and all questions is shown in Section 8 ahead.) We requested annotations from five different Turkers for each HIT. (A Turker cannot attempt multiple assignments for the same term.) Different HITS may be attempted by different Turkers, and a Turker may attempt as many HITs as they wish.

7. ISSUES WITH CROWDSOURCING AND EMOTION ANNOTATION

7.1. Key issues in crowdsourcing

Even though there are a number of benefits to using Mechanical Turk, such as low cost, less organizational overhead, and quick turn around time, there are also some inherent challenges. First and foremost is quality control. The task and compensation may attract cheaters (who may input random information) and even malicious annotators (who may deliberately enter incorrect information). We have no control over the educational background of a Turker, and we cannot expect the average Turker to read and follow complex and detailed directions. However, this may not necessarily be a disadvantage of crowdsourcing. We believe that clear, brief, and simple instructions

produce accurate annotations and higher inter-annotator agreements. Another challenge is finding enough Turkers interested in doing the task. If the task does not require any special skills, then more Turkers will do the task. The number of Turkers and the number of annotations they provide is also dependent on how interesting they find the task and how attractive they find the compensation.

7.2. Finer points of emotion annotation

Native and fluent speakers of a language are good at identifying emotions associated with words. Therefore we do not require the annotators to have any special skills other than that they be native or fluent speakers of English. However, emotion annotation, especially in a crowdsource setting, has some important challenges.

Words used in different senses can evoke different emotions. For example, the word *shout* evokes a different emotion when used in the context of admonishment than when used in "*Give me a shout if you need any help.*" Getting human annotations for word senses is made complicated by decisions about which sense-inventory to use and what level of granularity the senses must have. On the one hand, we do not want to choose a fine-grained sense-inventory because then the number of word–sense combinations will become too large and difficult to easily distinguish, and on the other hand we do not want to work only at the word level because, when used in different senses, a word may evoke different emotions.

Yet another challenge is how best to convey a word sense to the annotator. Including long definitions will mean that the annotators have to spend more time reading the question, and because their compensation is roughly proportional to the amount of time they spend on the task, the number of annotations we can obtain for a given budget is impacted. Further, we want the users to annotate a word only if they are already familiar with it and know its meanings. Definitions are good at conveying the core meaning of a word but they are not so effective in conveying the subtle emotional connotations. Therefore we wanted to discourage Turkers from annotating for words they are not familiar with. Lastly, we must ensure that malicious and erroneous annotations are discarded.

8. OUR APPROACH

In order to overcome the challenges described above, before asking the annotators questions about which emotions are associated with a target term, we first present them with a word choice problem. They are provided with four different words and asked which word is closest in meaning to the target. Three of the four options are irrelevant distractors. The remaining option is a synonym for one of the senses of the target word. This single question serves many purposes. Through this question we convey the word sense for which annotations are to be provided, without actually providing annotators with long definitions. That is, the correct choice guides the Turkers to the intended sense of the target. Further, if an annotator is not familiar with the target word and still attempts to answer questions pertaining to the target, or is randomly clicking options in our questionnaire, then there is a 75% chance that they will get the answer to this question wrong, and we can discard all responses pertaining to this target term by the annotator (that is, we also discard answers to the emotion questions provided by the annotator for this target term).

We generated these word choice problems automatically using the *Macquarie Thesaurus* (Bernard, 1986). As mentioned earlier in Section 5, published thesauri, such as *Roget's* and *Macquarie*, divide the vocabulary into about a thousand categories, which may be interpreted as coarse senses. Each category has a head word that best captures the meaning of the category. The word choice question for a target term is automatically generated by selecting the following four alternatives (choices): the head word of the thesaurus category pertaining to the target term (the correct answer); and three other head words of randomly selected categories (the distractors). The four alternatives are presented to the annotator in random order. We generated a separate HIT (and a separate word choice question) for every sense of the target. We created Mechanical Turk HITs for each of the terms (n-gram–sense pairs) specified in Table 1. Each HIT has a set of questions, all of which are to be answered by the same person. As mentioned before, we requested five independent assignments (annotations) for each HIT.

The phrasing of questions in any survey can have a significant impact on the results. With our

questions we hoped to be clear and brief, so that different annotators do not misinterpret what was being asked of them. In order to determine the more suitable way to formulate the questions, we performed two separate annotations on a smaller pilot set of 2100 terms. One, in which we asked if a word is *associated* with a certain emotion, and another independent set of annotations where we asked whether a word *evokes* a certain emotion. We found that the annotators agreed with each other much more in the *associated* case than in the *evokes* case. (Details are in Section 10.3 ahead.) Therefore all subsequent annotations were done with *associated*. All results, except those presented in Section 10.3, are for the *associated* annotations.

Below is a complete example HIT for the target word *startle*. Note that all questions are multiple-choice questions, and the Turkers could select exactly one option for each question. The survey was approved by the ethics committee at the National Research Council Canada.

Title: Emotions associated with words
Keywords: emotion, English, sentiment, word association, word meaning
Reward per HIT: $0.04

Directions:

1. This survey will be used to better understand emotions. Your input is much appreciated.
2. If any of the questions in a HIT are unanswered, then the assignment is no longer useful to us and we will be unable to pay for the assignment.
3. Please return/skip HIT if you do not know the meaning of the word.
4. Attempt HITS only if you are a native speaker of English, or very fluent in English.
5. Certain "check questions" will be used to make sure the annotation is responsible and reasonable. Assignments that fail these tests will be rejected. If an annotator fails too many of these check questions, then it will be assumed that the annotator is not following instructions 3 and/or 4 above, and ALL of the annotator's assignments will be rejected.
6. We hate to reject assignments, but we must at times, to be fair to those who answer the survey with diligence and responsibility. In the past we have approved completed assignments by more than 95% of the Turkers. If you are unsure about your answers and this is the first time that you are answering an emotion survey posted by us, then we recommend that you NOT do a huge number of HITs right away. Once your initial HITS are approved, you gain confidence in your answers and in us.
7. We will approve HITs about once a week. Expected date all the assignments will be approved: April 14, 2010.
8. Confidentiality notice: Your responses are confidential. Any publications based on these responses will not include your specific responses, but rather aggregate information from many individuals. We will not ask any information that can be used to identify who you are.
9. Word meanings: Some words have more than one meaning, and the different meanings may be associated with different emotions. For each HIT, Question 1 (Q1) will guide you to the intended meaning. You may encounter multiple HITs for the same target term, but they will correspond to different meanings of the target word, and they will have different guiding questions.

Prompt word: *startle*

Q1. Which word is closest in meaning (most related) to *startle*?

- *automobile*
- *shake*
- *honesty*
- *entertain*

Q2. How positive (good, praising) is the word *startle*?

- *startle* is not positive
- *startle* is weakly positive
- *startle* is moderately positive
- *startle* is strongly positive

Q3. How negative (bad, criticizing) is the word *startle*?

- *startle* is not negative
- *startle* is weakly negative
- *startle* is moderately negative
- *startle* is strongly negative

Q4. How much is *startle* associated with the emotion joy? (For example, *happy* and *fun* are strongly associated with joy.)

- *startle* is not associated with joy
- *startle* is weakly associated with joy
- *startle* is moderately associated with joy
- *startle* is strongly associated with joy

Q5. How much is *startle* associated with the emotion sadness? (For example, *failure* and *heartbreak* are strongly associated with sadness.)

- *startle* is not associated with sadness
- *startle* is weakly associated with sadness
- *startle* is moderately associated with sadness
- *startle* is strongly associated with sadness

Q6. How much is *startle* associated with the emotion fear? (For example, *horror* and *scary* are strongly associated with fear.)

- Similar choices as in 4 and 5 above

Q7. How much is *startle* associated with the emotion anger? (For example, *rage* and *shouting* are strongly associated with anger.)

- Similar choices as in 4 and 5 above

Q8. How much is *startle* associated with the emotion trust? (For example, *faith* and *integrity* are strongly associated with trust.)

- Similar choices as in 4 and 5 above

Q9. How much is *startle* associated with the emotion disgust? (For example, *gross* and *cruelty* are strongly associated with disgust.)

- Similar choices as in 4 and 5 above

Q10. How much is *startle* associated with the emotion surprise? (For example, *startle* and *sudden* are strongly associated with surprise.)

- Similar choices as in 4 and 5 above

Q11. How much is *startle* associated with the emotion anticipation? (For example, *expect* and *eager* are strongly associated with anticipation.)

- Similar choices as in 4 and 5 above

Q12. Is *startle* an emotion? (For example: *love* is an emotion; *shark* is associated with fear (an emotion), but *shark* is not an emotion.)

- No, *startle* is not an emotion
- Yes, *startle* is an emotion

9. ANNOTATION STATISTICS AND POST-PROCESSING

We conducted annotations in two batches, starting first with a pilot set of about 2100 terms, which was annotated in about a week. The second batch of about 8000 terms (HITs) was annotated in about two weeks. Notice that the amount of time taken is not linearly proportional to the number of HITs. We speculate that as one builds a history of tasks and payment, more Turkers do subsequent tasks. Also, if there are a large number of HITs, then probably more people find it worth the effort to understand and become comfortable at doing the task. Each HIT had a compensation of $0.04 (4 cents) and the Turkers spent about a minute on average to answer the questions in a HIT. This resulted in an hourly pay of about $2.40.

Once the assignments were collected, we used automatic scripts to validate the annotations. Some assignments were discarded because they failed certain tests (described below). A subset of the discarded assignments were officially *rejected* (the Turkers were not paid for these assignments) because instructions were not followed. About 2,666 of the 50,850 (10,170 × 5) assignments included at least one unanswered question. These assignments were discarded and rejected. Even though distractors for Q1 were chosen at random, every now and then a distractor may come too close to the meaning of the target term, resulting in a bad word choice question. For 1045 terms, three or more annotators gave an answer different from the one generated automatically from the thesaurus. These questions were marked as bad questions and discarded. All corresponding assignments (5,225 in total) were discarded. Turkers were paid in full for these assignments regardless of their answer to Q1.

More than 95% of the remaining assignments had the correct answer for the word choice question. This was a welcome result, showing that most of the annotations were done in an appropriate manner. We discarded all assignments that had the wrong answer for the word choice question. If an annotator obtained an overall score that is less than 66.67% on the word choice questions (that is, got more than one out of three wrong), then we assumed that, contrary to instructions, the annotator attempted to answer HITs for words that were unfamiliar. We discarded and rejected *all* assignments by such annotators (not merely the assignments for which they got the word choice question wrong).

For each of the annotators, we calculated the maximum likelihood probability with which the annotator agrees with the majority on the emotion questions. We calculated the mean of these probabilities and the standard deviation. Consistent with standard practices in identifying outliers, we discarded annotations by Turkers who were more than two standard deviations away from the mean (annotations by 111 Turkers).

After this post-processing, 8,883 of the initial 10,170 terms remained, each with three or more valid assignments. We will refer to this set of assignments as the *master set*. We created the word–emotion association lexicon from this master set, containing 38,726 assignments from about 2,216 Turkers who attempted 1 to 2,000 assignments each. About 300 of them provided 20 or more assignments each (more than 33,000 assignments in all). The master set has, on average, about 4.4 assignments for each of the 8,883 target terms. (See Table 2 for more details.) The total cost of the annotation was about US$2,100. This includes fees that Amazon charges (about 13% of the amount paid to the Turkers) as well as the cost for the dual annotation of the pilot set with both *evokes* and *associated*. [10]

10. ANALYSIS OF EMOTION ANNOTATIONS

The different emotion annotations for a target term were consolidated by determining the *majority class* of emotion intensities. For a given term–emotion pair, the majority class is that intensity level that is chosen most often by the Turkers to represent the degree of emotion evoked by the word. Ties are broken by choosing the stronger intensity level. Table 3 lists the percentage of 8,883 target terms assigned a majority class of no, weak, moderate, and strong emotion. For example, it tells us that 5% of the target terms strongly evoke joy. The table also presents averages of the numbers in each column (micro-averages). The last row lists the percentage of target terms that

[10] We will upload HITs of discarded assignments on Mechanical Turk for another round of annotations.

Table 2. Break down of target terms into various categories. Initial refers to terms chosen for annotation. Master refers to terms for which three or more valid assignments were obtained using Mechanical Turk. MQ stands for Macquarie Thesaurus, GI for General Inquirer, and WAL for WordNet Affect Lexicon.

	# of terms		Annotations
EmoLex	**Initial**	**Master**	**per word**
EmoLex-Uni:			
Unigrams from Macquarie Thesaurus			
adjectives	200	190	4.4
adverbs	200	187	4.5
nouns	200	178	4.5
verbs	200	195	4.4
EmoLex-Bi:			
Bigrams from Macquarie Thesaurus			
adjectives	200	162	4.4
adverbs	187	171	4.3
nouns	200	185	4.5
verbs	200	178	4.4
EmoLex-GI:			
Terms from General Inquirer			
negative terms	2119	1837	4.4
neutral terms	4226	3653	4.4
positive terms	1787	1541	4.4
EmoLex-WAL:			
Terms from WordNet Affect Lexicon			
anger terms	165	160	4.5
disgust terms	37	34	4.4
fear terms	100	89	4.4
joy terms	165	149	4.5
sadness terms	120	112	4.5
surprise terms	53	51	4.4
Union	**10170**	**8883**	**4.45**

Table 3. Percentage of terms with majority class of no, weak, moderate, and strong emotion.

	Intensity			
Emotion	no	weak	moderate	strong
anger	81.6	8.5	5.1	4.5
anticipation	84.2	8.9	4.2	2.4
disgust	84.6	8.3	3.8	3.1
fear	79.6	10.3	5.6	4.3
joy	79.5	8.9	6.4	5.0
sadness	80.9	10.0	4.8	4.2
surprise	89.5	6.6	2.2	1.4
trust	81.9	7.9	5.9	4.1
micro-average	**82.7**	**8.7**	**4.8**	**3.6**
any emotion	**35.6**	**21.2**	**20.5**	**22.5**

evoke some emotion (any of the eight) at the various intensity levels. We calculated this using the intensity level of the strongest emotion expressed by each target. Observe that 22.5% of the target terms strongly evoke at least one of the eight basic emotions.

TABLE 4. Percentage of terms, in each target set, that are emotive. Highest individual emotion scores for EmoLex-WAL are shown in bold. The last column, *any*, shows the percentage of terms associated with at least one of the eight emotions. Observe that WAL fear terms are marked most as associate with fear, joy terms as associated with joy, and so on.

	anger	**anticipn.**	**disgust**	**fear**	**joy**	**sadness**	**surprise**	**trust**	**any**
EmoLex	13	12	10	14	16	12	6	16	54
EmoLex-Uni:									
Unigrams from Macquarie Thesaurus									
adjectives	14	14	10	13	29	14	10	15	68
adverb	13	20	8	10	23	11	7	23	67
noun	7	18	3	7	16	6	3	24	46
verb	11	21	5	16	14	11	7	15	52
EmoLex-Bi:									
Bigrams from Macquarie Thesaurus									
adjectives	12	25	8	14	30	15	8	16	66
adverbs	6	23	1	7	19	3	9	29	54
nouns	9	23	6	14	20	9	7	29	58
verbs	8	25	5	7	21	6	3	27	60
EmoLex-GI:									
Terms from General Inquirer									
negative terms	36	4	29	34	0	33	8	2	67
neutral terms	4	11	3	8	10	4	5	13	36
positive terms	1	13	0	2	40	1	4	33	62
EmoLex-WAL:									
Terms from WordNet Affect Lexicon									
anger terms	**83**	1	53	18	0	16	0	0	90
disgust terms	44	0	**94**	14	0	2	0	0	94
fear terms	17	17	19	**74**	1	20	15	3	89
joy terms	2	14	0	2	**78**	2	7	28	91
sadness terms	9	0	13	13	0	**94**	0	0	96
surprise terms	2	6	4	8	42	6	66	6	**88**

Even though we asked Turkers to annotate emotions at four levels of intensity, practical NLP applications often require only two levels—associated with a given emotion (we will refer to these terms as being *emotive*) or not associated with the emotion (we will refer to these terms as being *non-emotive*). For each target term–emotion pair, we convert the four-level annotations into two-level annotations by placing all no- and weak-intensity assignments in the non-emotive bin, all moderate- and strong-intensity assignments in the emotive bin, and then choosing the bin with the majority assignments. Table 4 shows the percentage of terms associated with the different emotions. The last column, *any*, shows the percentage of terms associated with at least one of the eight emotions.

Analysis of Q12 revealed that 9.3% of the 8,883 target terms (826 terms) were considered not merely to be associated with certain emotions, but also to refer directly to emotions.

10.1. Discussion

Table 4 shows that a sizable percentage of nouns, verbs, adjectives, and adverbs are emotive. Trust (16%), and joy (16%) are the most common emotions associated with terms. Among the four parts of speech, adjectives (68%) and adverbs (67%) are most often associated with emotions and this is not surprising considering that they are used to qualify nouns and verbs, respectively. Nouns are more commonly associated with trust (16%), whereas adjectives are more commonly associated with joy (29%).

The EmoLex-WAL rows are particularly interesting because they serve to determine how much the Turker annotations match annotations in the Wordnet Affect Lexicon (WAL). The most common Turker-determined emotion for each of these rows is marked in bold. Observe that WAL anger terms are mostly marked as associated with anger, joy terms as associated with joy, and so on. Here is the

Table 5. Agreement at four intensity levels of emotion (no, weak, moderate, and strong): Percentage of terms for which the majority class size was 2, 3, 4, and 5. Note that, given five annotators and four levels, the majority class size must be between two and five.

	Majority class size					
Emotion	= two	= three	= four	= five	⩾ three	⩾ four
anger	13.7	21.7	25.7	38.7	86.1	64.4
anticipation	19.2	31.7	28.3	20.7	80.7	49.0
disgust	13.8	20.7	23.8	41.5	86.0	65.3
fear	16.7	27.7	25.6	29.9	83.2	55.5
joy	16.1	24.3	21.9	37.5	83.7	59.4
sadness	14.3	23.8	25.9	35.7	85.4	61.6
surprise	11.8	25.3	32.2	30.6	88.1	62.8
trust	18.8	27.4	27.7	25.9	81.0	53.6
micro-average	**15.6**	**25.3**	**26.4**	**32.6**	**84.3**	**59.0**

complete list of terms that are marked as anger terms in WAL, but were not marked as anger terms by the Turkers: *baffled, exacerbate, gravel, pesky, and pestering.* One can see that indeed many of these terms are not truly associated with anger. We also observed that the Turkers marked some terms as being associated with both anger and joy. The complete list includes: *adjourn, credit card, find out, gloat, spontaneously, and surprised.* One can see how many of these words are indeed associated with both anger and joy. The EmoLex-WAL rows also indicate which emotions tend to be jointly associated to a term. Observe that anger terms tend also to be associated with disgust. Similarly, many joy terms are also associated with trust. The surprise terms in WAL are largely also associated with joy.

The EmoLex-GI rows rightly show that words marked as negative in the General Inquirer are mostly associated with negative emotions (anger, fear, disgust, and sadness). Observe that the percentages for trust and joy are much lower. On the other hand, positive words are associated with anticipation, joy, and trust.

10.2. Agreement

In order to analyze how often the annotators agreed with each other, for each term–emotion pair, we calculated the percentage of times the majority class has size 5 (all Turkers agree), size 4 (all but one agree), size 3, and size 2. Table 5 presents these agreement values. Observe that for almost 60% of the terms, at least four annotators agree with each other (see bottom right corner of Table 5). Since many NLP systems may rely only on two intensity values (emotive or non-emotive), we also calculate agreement at that level (Table 6). For more than 60% of the terms, all five annotators agree with each other, and for almost 85% of the terms, at least four annotators agree (see bottom right corner of Table 6). These agreements are despite the somewhat subjective nature of word–emotion associations, and despite the absence of any control over the educational background of the annotators. We provide agreement values along with each of the termemotion pairs so that downstream applications can selectively use the lexicon.

Cohen's κ (Cohen, 1960) is a widely used measure for inter-annotator agreement. It corrects observed agreement for chance agreement by using the distribution of classes chosen by each of the annotators. However, it is appropriate only when the same judges annotate all the instances (Fleiss, 1971). In Mechanical Turk, annotators are given the freedom to annotate as many terms as they wish, and many annotate only a small number of terms (the long tail of the zipfian distribution). Thus the judges do not annotate all of the instances, and further, one cannot reliably estimate the distribution of classes chosen by each judge when they annotate only a small number of instances. Scott's Π (Scott, 1955) calculates chance agreement by determining the distribution each of the categories (regardless of who the annotator is). This is more appropriate for our data, but it applies only to scenarios with exactly two annotators. Fleiss (1971) proposed a generalization of Scott's Π for when there are more than two annotators, which he called κ even though Fleiss's κ is more like Scott's Π than Cohen's κ. All subsequent mentions of κ in this paper will refer to Fleiss's κ unless

TABLE 6. Agreement at two intensity levels of emotion (emotive and non-emotive): Percentage of terms for which the majority class size was 3, 4, and 5. Note that, given five annotators and two levels, the majority class size must be between three and five.

	Majority class size			
Emotion	= three	= four	= five	⩾ four
anger	13.2	19.4	67.2	86.6
anticipation	18.8	32.6	48.4	81.0
disgust	13.4	18.4	68.1	86.5
fear	15.3	24.8	59.7	84.5
joy	16.2	22.6	61.0	83.6
sadness	12.8	20.2	66.9	87.1
surprise	10.9	22.8	66.2	89.0
trust	20.3	28.8	50.7	79.5
micro-average	**15.1**	**23.7**	**61.0**	**84.7**

TABLE 7. Segments of Fleiss κ values and their interpretations (Landis and Koch, 1977).

Fleiss's κ	**Interpretation**
< 0	poor agreement
0.00 – 0.20	slight agreement
0.21 - 0.40	fair agreement
0.41 - 0.60	moderate agreement
0.61 - 0.80	substantial agreement
0.81 - 1.00	almost perfect agreement

TABLE 8. Agreement at two intensity levels of emotion (emotive and non-emotive): Fleiss's κ, and its interpretation.

Emotion	**Fleiss's κ**	**Interpretation**
anger	0.39	fair agreement
anticipation	0.14	slight agreement
disgust	0.31	fair agreement
fear	0.32	fair agreement
joy	0.36	fair agreement
sadness	0.39	fair agreement
surprise	0.18	slight agreement
trust	0.24	fair agreement
micro-average	**0.29**	**fair agreement**

explicitly stated otherwise. Landis and Koch (1977) provided Table 7 to interpret the κ values. Table 8 lists the κ values for the Mechanical Turk emotion annotations.

The κ values show that for six of the eight emotions the Turkers have fair agreement, and for anticipation and trust there is only slight agreement. The κ values for anger and sadness are the highest. The average κ value for the eight emotions is 0.29, and it implies fair agreement. Below are some reasons why agreement values are much lower than certain other tasks, for example, part of speech tagging:

- The target word is presented out of context. We expect higher agreement if we provided words in particular contexts, but words can occur in innumerable contexts, and annotating too many instances of the same word is costly. By providing the word choice question, we bias the Turker towards a particular sense of the target word, and aim to obtain the prior probability of the word sense's emotion association.

Table 9. *Evokes* versus *associated*: Agreement at two intensity levels of emotion (emotive and non-emotive). Percentage of terms in the pilot set for which the majority class size was 5.

	Majority class size five	
Emotion	evokes	associated
anger	61.6	**68.2**
anticipation	34.8	**49.6**
disgust	65.4	**66.4**
fear	**62.0**	59.4
joy	54.6	**62.3**
sadness	**66.7**	65.3
surprise	54.0	**67.3**
trust	47.3	**49.8**
micro-average	55.8	**61.0**

- Words are associated with emotions to different degrees, and there are no clear classes corresponding to different levels of association. Since we ask people to place term-emotion associations in four specific bins, more people disagree for term–emotion pairs whose degree of association is closer to the boundaries, than for other term–emotion pairs.
- Holsti (1969), Brennan and Prediger (1981), Perreault and Leigh (1989), and others consider the κ values (both Fleiss's and Cohen's) to be conservative, especially when one category is much more prevalent than the other. In our data, the "not associated with emotion" category is much more prevalent than the "associated with emotion" category, so these κ values might be underestimates of the true agreement.

Nonetheless, as mentioned earlier, when using the lexicon in downstream applications, one may employ suitable strategies such as choosing instances that have high agreement scores, averaging information from many words, and using contextual information in addition to information obtained form the lexicon.

10.3. Evokes versus Associated

As alluded to earlier, we performed two separate sets of annotations on the pilot set: one where we asked if a word *evokes* a certain emotion, and another where we asked if a word is *associated* with a certain emotion. Table 9 lists the the percentage of times all five annotators agreed with each other on the classification of a term as emotive, for the two scenarios. Observe that the agreement numbers are markedly higher with *associated* than with *evokes* for anger, anticipation, joy, and surprise. In case of fear and sadness, the agreement is only slightly better with *evokes*, whereas for trust and disgust the agreement is slightly better with *associated*. Overall, *associated* leads to an increase in agreement by more than 5 percentage points over *evokes*. Therefore all subsequent annotations were performed with *associated* only. (All results shown in this paper, except for those in Table 9, are for *associated*.)

We speculate that to answer which emotions are *evoked* by a term, people sometimes bring in their own varied personal experiences, and so we see relatively more disagreement than when we ask what emotions are *associated* with a term. In the latter case, people may be answering what is more widely accepted rather than their own personal perspective. Further investigation on the differences between *evoke* and *associated*, and why there is a marked difference in agreements for some emotions and not so much for others, is left as future work.

11. ANALYSIS OF POLARITY ANNOTATIONS

We consolidate the polarity annotations in the same manner as for emotion annotations. Table 10 lists the percentage of 8,883 target terms assigned a majority class of no, weak, moderate, and strong polarity. It states, for example, that 15.6% of the target terms are strongly negative. The last row

TABLE 10. Percentage of terms given majority class of no, weak, moderate, and strong polarity.

	Intensity			
Polarity	no	weak	moderate	strong
negative	64.3	9.1	10.8	15.6
positive	61.9	9.8	13.7	14.4
polarity average	**63.1**	**9.5**	**12.3**	**15.0**
either polarity	**29.9**	**15.4**	**24.3**	**30.1**

TABLE 11. Percentage of terms, in each target set, that are evaluative. The highest scores for EmoLex-GI positives and negatives are shown bold. Observe that the positive GI terms are marked mostly as positively evaluative and the negative terms are marked mostly as negatively evaluative.

	negative	**positive**	**either**
EmoLex	30	35	65
EmoLex-Uni:			
Unigrams from Macquarie Thesaurus			
adjectives	32	48	79
adverbs	26	55	80
nouns	8	39	46
verbs	26	37	63
EmoLex-Bi:			
Bigrams from Macquarie Thesaurus			
adjectives	30	47	77
adverbs	11	42	52
nouns	14	45	57
verbs	14	48	60
EmoLex-GI:			
Terms from General Inquirer			
negative terms	**83**	1	85
neutral terms	12	30	41
positive terms	2	82	**84**
EmoLex-WAL:			
Terms from WordNet Affect Lexicon			
anger terms	96	1	97
disgust terms	97	0	97
fear terms	85	1	86
joy terms	4	93	97
sadness terms	91	4	95
surprise terms	26	57	80

in the table lists the percentage of target terms that have some polarity (positive or negative) at the various intensity levels. Observe that 30.1% of the target terms are either strongly positive or strongly negative.

Just as in the case for emotions, practical NLP applications often require only two levels of polarity—having particular polarity (*evaluative*) or not (*non-evaluative*). For each target term–emotion pair, we convert the four-level semantic orientation annotations into two-level ones, just as we did for the emotions. Table 11 shows how many terms overall and within each category are positively and negatively evaluative.

Table 12. Agreement at four intensity levels of polarity (no, weak, moderate, and strong): Percentage of terms for which the majority class size was 2, 3, 4, and 5.

	Majority class size					
Polarity	= two	= three	= four	= five	⩾ three	⩾ four
negative	12.8	27.3	27.2	32.5	87.0	59.7
positive	23.5	28.5	18.0	29.8	76.3	47.8
micro-average	**18.2**	**27.9**	**22.6**	**31.2**	**81.7**	**53.8**

Table 13. Agreement at two intensity levels of polarity (evaluative and non-evaluative): Percentage of terms for which the majority class size was 3, 4, and 5.

	Majority class size			
Polarity	three	four	five	⩾ four
negative	11.5	22.3	66.1	88.4
positive	24.2	26.3	49.3	75.6
micro-average	**17.9**	**24.3**	**57.7**	**82.0**

11.1. Discussion

Observe in Table 11 that, across the board, a sizable number of terms are evaluative with respect to some semantic orientation. Unigram nouns have a markedly lower proportion of negative terms, and a much higher proportion of positive terms. It may be argued that the default polarity of noun concepts is neutral or positive, and that usually it takes a negative adjective to make the phrase negative.

The EmoLex-GI rows in the two tables show that words marked as having a negative polarity in the General Inquirer are mostly marked as negative by the Turkers. And similarly, the positives in GI are annotated as positive. Observe that the Turkers mark 12% of the GI neutral terms as negative and 30% of the GI neutral terms as positive. This may be because the boundary between positive and neutral terms is more fuzzy than between negative and neutral terms. The EmoLex-WAL rows show that anger, disgust, fear, and sadness terms tend not to have a positive polarity and are mostly negative. In contrast, and expectedly, the joy terms are positive. The surprise terms are more than twice as likely to be positive than negative.

11.2. Agreement

For each term–polarity pair, we calculated the percentage of times the majority class has size 5 (all Turkers agree), size 4 (all but one agree), size 3, and size 2. Table 12 presents these agreement values. For more than 50% of the terms, at least four annotators agree with each other (see bottom right corner of Table 12). Table 13 gives agreement values at the two-intensity level. For more than 55% of the terms, all five annotators agree with each other, and for more than 80% of the terms, at least four annotators agree (see bottom right corner of Table 13). Table 14 lists the Fleiss κ values for the polarity annotations. They are interpreted based on the segments provided by Landis and Koch (1977) (listed earlier in Table 7). Observe that annotations for negative polarity have markedly higher agreement than annotations for positive polarity. This too may be because of the somewhat more fuzzy boundary between positive and neutral, than between negative and neutral.

12. CONCLUSIONS

Emotion detection and generation have a number of practical applications including managing customer relations, human computer interaction, information retrieval, more natural text-to-speech systems, and in social and literary analysis. However, only a small number of limited-coverage emotion resources exist, and that too only for English. In this paper we show how the combined strength and wisdom of the crowds can be used to generate a large term–emotion association lexicon

TABLE 14. Agreement at two intensity levels of polarity (evaluative and non-evaluative): Fleiss's κ, and its interpretation.

Polarity	**Fleiss's κ**	**Interpretation**
negative	0.62	substantial agreement
positive	0.45	moderate agreement
micro-average	**0.54**	moderate agreement

quickly and inexpensively. This lexicon, EmoLex, has entries for more than 10,000 word–sense pairs. Each entry lists the association of the a word–sense pair with 8 basic emotions. We used Amazon's Mechanical Turk as the crowdsourcing platform.

We outlined various challenges associated with crowdsourcing the creation of an emotion lexicon (many of which apply to other language annotation tasks too), and presented various solutions to address those challenges. Notably, we used automatically generated word choice questions to detect and reject erroneous annotations and to reject all annotations by unqualified Turkers and those who indulge in malicious data entry. The word choice question is also an effective and intuitive way of conveying the sense for which emotion annotations are being requested.

We compared a subset of our lexicon with existing gold standard data to show that the annotations obtained are indeed of high quality. We identified which emotions tend to be evoked simultaneously by the same term, and also how frequent the emotion associations are in high-frequency words. We also compiled a list of 826 terms that are not merely associated with emotions, but also refer directly to emotions. All of the 10,170 terms in the lexicon are also annotated with whether they have a positive, negative, or neutral semantic orientation.

13. FUTURE DIRECTIONS

Our future work includes expanding the coverage of the lexicon even further, creating similar lexicons in other languages, identifying cross-cultural and cross-language differences in emotion associations, and using the lexicon in various emotion detection applications such as those listed in Section 2. Mohammad and Yang (2011) describe some of these efforts, in which we use the *Roget's Thesaurus* as the source of target terms, and create an emotion lexicon with entries for more than 24,000 word–sense pairs (covering about 14,000 unique word-types). We will use this manually created emotion lexicon to evaluate automatically generated lexicons, such as the polarity lexicons by Turney and Littman (2003) and Mohammad *et al.* (2009). We will explore the variance in emotion evoked by near-synonyms, and also how common it is for words with many meanings to evoke different emotions in different senses.

We are interested in further improving the annotation process by applying *Maximum Difference Scaling* (or *MaxDiff*) (Louviere, 1991; Louviere and Finn, 1992). In MaxDiff, instead of asking annotators for a score representing how strongly an item is associated with a certain category, the annotator is presented with four or five items at a time and asked which item is *most* associated with the category and which one the *least*. The approach forces annotators to compare items directly, which leads to better annotations (Louviere and Finn, 1992; Cohen and Associates, 2003), which we hope will translate into higher inter-annotator agreements. Further, if A, B, C, and D are the four items in a set, by asking only the most and least questions, we will know five out of the six inequalities. For example, if A is the maximum, and D is the least, then we know that $A > B, A > C, A > D, B > D, C > D$. This makes the annotations significantly more efficient than just providing pairs of items and asking which is more associated with a category. Hierarchical Bayes estimation can then be used to convert these MaxDiff judgments into scores (from 0 to 10 say) and to rank all the items in order of association with the category.

Many of the challenges associated with polarity analysis have correspondence in emotion analysis too. For example, using context information in addition to prior probability of a word's polarity or emotion association, to determine the true emotional impact of a word in a particular occurrence. Our emotion annotations are at word-sense level, yet accurate word sense disambiguation systems must be employed to make full use of this information. For example, Rentoumi *et al.* (2009) show

that word sense disambiguation improves detection of polarity of sentences. There is also a need for algorithms to identify who is experiencing an emotion, and determine what or who is evoking that emotion. Further, given a sentence or a paragraph, the writer, the reader, and the entities mentioned in the text may all have different emotions associated with them. Yet another challenge is how to handle negation of emotions. For example, *not sad* does not usually mean *happy*, whereas *not happy* can often mean *sad*.

Finally, emotion detection can be used as a tool for social and literary analysis. For example, how have books portrayed different entities over time? Does the co-occurrence of fear words with entities (for example, cigarette, or homosexual, or nuclear energy) reflect the feelings of society as a whole towards these entities? What is the distribution of different emotion words in novels and plays? How has this distribution changed over time, and across different genres? Effective emotion analysis can help identify trends and lead to a better understanding of humanity's changing perception of the world around it.

Acknowledgments

This research was funded by the National Research Council Canada (NRC). We are grateful to the reviewers for their thoughtful comments. Thanks to Joel Martin, Diana Inkpen, and Diman Ghazi for discussions and encouragement. Thanks to Norm Vinson and the Ethics Committee at NRC for examining, guiding, and approving the survey. And last but not least, thanks to the more than 2000 anonymous annotators who answered the emotion survey with diligence and care.

REFERENCES

Adamic, L. A., Zhang, J., Bakshy, E., and Ackerman, M. S. (2008). Knowledge sharing and yahoo answers: everyone knows something. In *Proceeding of the 17th international conference on World Wide Web*, WWW '08, pages 665–674, New York, NY, USA. ACM.

Alm, C. O., Roth, D., and Sproat, R. (2005). Emotions from text: Machine learning for text-based emotion prediction. In *Proceedings of the Joint Conference on Human Language Technology / Empirical Methods in Natural Language Processing (HLT/EMNLP-2005)*, pages 579–586, Vancouver, Canada.

Aman, S. and Szpakowicz, S. (2007). Identifying expressions of emotion in text. In V. Matoušek and P. Mautner, editors, *Text, Speech and Dialogue*, volume 4629 of *Lecture Notes in Computer Science*, pages 196–205. Springer Berlin / Heidelberg.

Aristotle (1913). Physiognomonica. In W. D. Ross, editor, *The Works of Aristotle*, pages 805–813. Oxford, England: Clarendon. Translated by T. Loveday and E. S. Forster.

Bellegarda, J. (2010). Emotion analysis using latent affective folding and embedding. In *Proceedings of the NAACL-HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, Los Angeles, California.

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

Boucouvalas, A. C. (2002). Real time text-to-emotion engine for expressive internet communication. *Emerging Communication: Studies on New Technologies and Practices in Communication*, **5**, 305–318.

Bougie, J. R. G., Pieters, R., and Zeelenberg, M. (2003). Angry customers don't come back, they get back: The experience and behavioral implications of anger and dissatisfaction in services. Open access publications from tilburg university, Tilburg University.

Bradley, M. and Lang, P. (1999). Affective norms for english words (anew): Instruction manual and affective ratings. In *Technical Report, C-1*, The Center for Research in Psychophysiology, University of Florida.

Brants, T. and Franz, A. (2006). Web 1t 5-gram version 1. *Linguistic Data Consortium*.

Breazeal, C. and Brooks, R. (2004). Robot emotions: A functional perspective. In *Who Needs Emotions*. Oxford University Press.

Brennan, R. L. and Prediger, D. J. (1981). Coefficient Kappa: Some Uses, Misuses, and Alternatives. *Educational and Psychological Measurement*, **41**(3), 687–699.

Callison-Burch, C. (2009). Fast, cheap and creative: Evaluating translation quality using amazon's mechanical turk. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2009)*, pages 286–295, Singapore.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, **20**(1), 37–46.

Cohen, S. H. and Associates, S. . (2003). Maximum difference scaling: Improved measures of importance and preference for segmentation. In *Sawtooth Software Conference Proceedings, Sawtooth Software, Inc. 530 W. Fir St*, pages 61–74.

Dabholkar, P. A., Shepherd, C. D., and Thorpe, D. I. (2000). A comprehensive framework for service quality: an investigation of critical conceptual and measurement issues through a longitudinal study. *Journal of Retailing*, **76**(2), 139–173.

Darwin, C. (1872). *The Expressions of the Emotions in Man and Animals*. John Murray.

Davenport, T. H., Harris, J. G., and Kohli, A. K. (2001). How do they know their customers so well? **42**(2), 63–73.

Díaz, A. B. C. and Ruz, F. J. M. (2002). The consumers reaction to delays in service. *International Journal of Service Industry Management*, **13**(2), 118–140.

Dubé, L. and Maute, M. (1996). The antecedents of brand switching, brand loyalty and verbal responses to service failure. *Advances in Services Marketing and Management*, **5**, 127–151.

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, **6**(3), 169–200.

Ekman, P. (2005). *Emotion in the Human Face*. Oxford University Press.

Ekman, P. and Friesen, W. V. (2003). *Unmasking the Face: A Guide to Recognizing Emotions From Facial Expressions*. Malor Books.

Elfenbein, H. A. and Ambady, N. (1994). Is there universal recognition of emotion from facial expression? a review of the cross-cultural studies. *Psychological Bulletin*, **115**, 102–141.

Elliott, C. (1992). *The affective reasoner: A process model of emotions in a multi-agent system*. Ph.D. thesis, Institute for the Learning Sciences, Northwestern University.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, **76**(5), 378–382.

Folkes, V. S., Koletsky, S., and Graham, J. L. (1987). A field study of causal inferences and consumer reaction: The view from the airport. *Journal of Consumer Research: An Interdisciplinary Quarterly*, **13**(4), 534–39.

Francisco, V. and Gervás, P. (2006). Automated mark up of affective information in english texts. In P. Sojka, I. Kopecek, and K. Pala, editors, *Text, Speech and Dialogue*, volume 4188 of *Lecture Notes in Computer Science*, pages 375–382. Springer Berlin / Heidelberg.

Genereux, M. and Evans, R. P. (2006). Distinguishing affective states in weblogs. In *AAAI-2006 Spring Symposium on Computational Approaches to Analysing Weblogs*, pages 27–29, Stanford, California.

Goyal, A., Riloff, E., Daume III, H., and Gilbert, N. (2010). Toward plot units: Automatic affect state analysis. In *Proceedings of the NAACL-HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, Los Angeles, California.

Gross, J. J. (1998). The emerging field of emotion regulation: An integrative review. *Review of General Psychology*, **2**(3), 271–299.

Guo, K., Hall, C., Hall, S., K., M., and Mills, D. (2007). Left gaze bias in human infants, rhesus monkeys, and domestic dogs. *Animal Cognition*, **12**, 409–418.

Harris, L. C. and Goode, M. M. H. (2004). The four levels of loyalty and the pivotal role of trust: a study of online service dynamics. *Journal of Retailing*, **80**(2), 139–158.

Hollinger, G., Georgiev, Y., Manfredi, A., Maxwell, B. A., Pezzementi, Z. A., and Mitchell, B. (2006). Design of a social mobile robot using emotion-based decision mechanisms. In *Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on*, pages 3093–3098.

Holsti, O. R. (1969). *Content analysis for the social sciences and humanities*. Addison-Wesley, Reading, MA.

Holzman, L. E. and Pottenger, W. M. (2003). Classification of emotions in internet chat: An application of machine learning using speech phonemes. Technical report, Leigh University.

Howe, J. and Robinson, M. (2006). Crowdsourcing: A definition. In *Crowdsourcing: Tracking the Rise of the Amateur*. Weblog.

James, W. (1884). What is an emotion? *Mind*, **9**, 188–205.

John, D., Boucouvalas, A. C., and Xu, Z. (2006). Representing emotional momentum within expressive internet communication. In *Proceedings of the 24th IASTED international conference on Internet and multimedia systems and applications*, pages 183–188, Anaheim, CA. ACTA Press.

Kalra, A. and Karahalios, K. (2005). Texttone: Expressing emotion through text. In M. F. Costabile and F. Patern, editors, *Human-Computer Interaction - INTERACT 2005*, volume 3585 of *Lecture Notes in Computer Science*, pages 966–969. Springer Berlin / Heidelberg.

Knautz, K., Siebenlist, T., and Stock, W. G. (2010). Memose: search engine for emotions in multimedia documents. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, pages 791–792, New York, NY. ACM.

Kŏvecses, Z. (2003). *Metaphor and Emotion: Language, Culture, and Body in Human Feeling (Studies in Emotion and Social Interaction)*. Cambridge University Press.

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

Lazarus, R. S. (1984). On the primacy of cognition. *American Psychologist*, **39**(2), 124–129.

Lazarus, R. S. (2000). The cognition-emotion debate: A bit of history. In M. Lewis and J. Haviland-Jones, editors, *Handbook of Cognition and Emotion*, pages 1–20. New York: Guilford Press.

Lehrer, A. (1974). *Semantic fields and lexical structure*. North-Holland, American Elsevier, Amsterdam, NY.

Litman, D. J. and Forbes-Riley, K. (2004). Predicting student emotions in computer-human tutoring dialogues. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Morristown, NJ, USA. Association for Computational Linguistics.

Liu, H., Lieberman, H., and Selker, T. (2003). A model of textual affect sensing using real-world knowledge. In *Proceedings of the 8th international conference on Intelligent user interfaces*, IUI '03, pages 125–132, New York, NY. ACM.

Louviere, J. J. (1991). Best-worst scaling: A model for the largest difference judgments. Technical report, University of Alberta.

Louviere, J. J. and Finn, A. (1992). Determining the appropriate response to evidence of public concern: The case of food safety. *Journal of Public Policy and Marketing*, **11**(2), 12–25.

Ma, C., Prendinger, H., and Ishizuka, M. (2005). Emotion estimation and reasoning based on affective textual interaction. In J. Tao and R. W. Picard, editors, *First International Conference on Affective Computing and Intelligent Interaction (ACII-2005)*, pages 622–628, Beijing, China.

Masson, J. M. (1996). *When Elephants Weep: The Emotional Lives of Animals*. Delta.

Matykiewicz, P., Duch, W., and Pestian, J. P. (2009). Clustering semantic spaces of suicide notes and newsgroups articles. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, BioNLP '09, pages 179–184, Stroudsburg, PA, USA. Association for Computational Linguistics.

Maute, M. F. and Forrester, W. J. (1993). The structure and determinants of consumer complaint intentions and behavior. *Journal of Economic Psychology*, **14**(2), 219–247.

Mihalcea, R. and Liu, H. (2006). A corpus-based approach to finding happiness. In *AAAI-2006 Spring Symposium on Computational Approaches to Analysing Weblogs*, pages 139–144. AAAI Press.

Mohammad, S. M. (2011a). Even the abstract have colour: Consensus in wordcolour associations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Portland, OR, USA.

Mohammad, S. M. (2011b). From once upon a time to happily ever after: Tracking emotions in novels and fairy tales. In *Proceedings of the ACL 2011 Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, Portland, OR, USA.

Mohammad, S. M. and Turney, P. D. (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.

Mohammad, S. M. and Yang, T. W. (2011). Tracking sentiment in mail:

how genders differ on emotional axes. In *Proceedings of the ACL 2011 Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA)*, Portland, OR, USA.

Mohammad, S. M., Dunne, C., and Dorr, B. (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.

Neviarouskaya, A., Prendinger, H., and Ishizuka, M. (2009). Compositionality principle in recognition of fine-grained emotions from text. In *Proceedings of the Proceedings of the Third International Conference on Weblogs and Social Media (ICWSM-09)*, pages 278–281, San Jose, California.

Neviarouskaya, A., Prendinger, H., and Ishizuka, M. (2010). Affect analysis model: novel rule-based approach to affect sensing from text. *Natural Language Engineering*, **17**(Part 1), 95–135.

Nowlis, V. and Nowlis, H. H. (2001). The description and analysis of mood. *Annals of the New York Academy of Sciences*, **65**(4), 345–355.

Oliver, R. L. (1997). *Satisfaction a behavioral perspective on the consumer*. New York: McGraw-Hill.

Ortony, A. and Turner, T. J. (1990). What's basic about basic emotions? *Psychological Review*, **97**, 315–331.

Ortony, A., Clore, G. L., and Collins, A. (1988). *The Cognitive Structure of Emotions*. Cambridge University Press.

Osgood, C. E. and Walker, E. G. (1959). Motivation and language behavior: A content analysis of suicide notes. *Journal of Abnormal and Social Psychology*, **59**(1), 58–67.

Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, **2**(1–2), 1–135.

Parrot, W. (2001). *Emotions in Social Psychology*. Psychology Press.

Pearl, L. and Steyvers, M. (2010). Identifying emotions, intentions, and attitudes in text using a game with a purpose. In *Proceedings of the NAACL-HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, Los Angeles, California.

Perreault, W. D. and Leigh, L. E. (1989). Reliability of nominal data based on qualitative judgments. *Journal of Marketing Research*, **26**, 135–148.

Pestian, J. P., Matykiewicz, P., and Grupp-Phelan, J. (2008). Using natural language processing to classify suicide notes. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, BioNLP '08, pages 96–97, Stroudsburg, PA, USA. Association for Computational Linguistics.

Plutchik, R. (1962). *The Emotions*. New York: Random House.

Plutchik, R. (1980). A general psychoevolutionary theory of emotion. *Emotion: Theory, research, and experience*, **1**(3), 3–33.

Plutchik, R. (1985). On emotion: The chicken-and-egg problem revisited. *Motivation and Emotion*, **9**(2), 197–200.

Plutchik, R. (1994). *The psychology and biology of emotion*. New York: Harper Collins.

Rashid, R., Aitken, J., and Fels, D. (2006). Expressing emotions using animated text captions. In K. Miesenberger, J. Klaus, W. Zagler, and A. Karshmer, editors, *Computers Helping People with Special Needs*, volume 4061 of *Lecture Notes in Computer Science*, pages 24–31. Springer Berlin / Heidelberg.

Ravaja, N., Saari, T., Turpeinen, M., Laarni, J., Salminen, M., and Kivikangas, M. (2006). Spatial presence and emotions during video game playing: Does it matter with whom you play? *Presence: Teleoperators and Virtual Environments*, **15**(4), 381–392.

Read, J. (2004). *Recognising affect in text using pointwise-mutual information*. Ph.D. thesis, Department of Informatics, University of Sussex.

Reichheld, F. F. and Schefter, P. (2000). E-loyalty: your secret weapon on the web. *Harvard Business Review*, pages 105–113.

Rentoumi, V., Giannakopoulos, G., Karkaletsis, V., and Vouros, G. A. (2009). Sentiment analysis of figurative language using a word sense disambiguation approach. In *Proceedings of the International Conference RANLP-2009*, pages 370–375, Borovets, Bulgaria. Association for Computational Linguistics.

Richins, M. (1987). A multivariate analysis of responses to dissatisfaction. *Journal of the Academy of Marketing Science*, **15**, 24–31.

Richins, M. L. (1984). Word of mouth communication as negative information. *Advances in*

Consumer Research, **11**, 697–702.

Russell, J. A. (1994). Is there universal recognition of emotion from facial expression? a review of the cross-cultural studies. *Psychological Bulletin*, **115**, 102–141.

Sautera, D. A., Eisner, F., Ekman, P., and Scott, S. K. (2010). Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. *Proceedings of the National Academy of Sciences*, **107**(6), 2408–2412.

Scherer, K. R. (1982). Emotion as a process: Function, origin and regulation. *Social Science Information*, **21**(4–5), 555–570.

Scherer, K. R. (1984). Emotion as a multicomponent process: a model and some cross-cultural data. *Review of Personality and Social Psychology*, **5**, 37–63.

Scott, W. A. (1955). Reliability of Content Analysis:. *Public Opinion Quarterly*, **19**(3), 321–325.

Shankar, V., Urban, G. L., and Sultan, F. (2002). Online trust: a stakeholder perspective, concepts, implications, and future directions. *The Journal of Strategic Information Systems*, **11**(3–4), 325–344.

Shaver, P., Schwartz, J., Kirson, D., and O'Connor, G. (1987). Emotion knowledge: Further exploration of a prototype approach. *Journal of Personality and Social Psychology*, **52**, 1061–1086.

Singh, J. (1988). Consumer complaint intentions and behavior: Definitional and taxonomical issues. *The Journal of Marketing*, **52**(1), 93–107.

Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. (2008). Cheap and fast - but is it good? Evaluating nonexpert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2008)*, pages 254–263, Waikiki, Hawaii.

Steunebrink, B. R. (2010). *The Logical Structure of Emotions*. Ph.D. thesis, Dutch Research School for Information and Knowledge Systems.

Stewart, K. J. (2003). Trust transfer on the world wide web. *Organization Science*, **14**, 5–17.

Stone, P., Dunphy, D. C., Smith, M. S., Ogilvie, D. M., and associates (1966). *The General Inquirer: A Computer Approach to Content Analysis*. The MIT Press.

Strapparava, C. and Valitutti, A. (2004). Wordnet-Affect: An affective extension of WordNet. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC-2004)*, pages 1083–1086, Lisbon, Portugal.

Subasic, P. and Huettner, A. (2001). Affect analysis of text using fuzzy semantic typing. *IEEE Transaction on Fuzzy Systems*, **9**(4), 483–496.

Turney, P. and Littman, M. (2003). Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, **21**(4), 315–346.

Velásquez, J. D. (1997). Modeling emotions and other motivations in synthetic agents. In *Proceedings of the fourteenth national conference on artificial intelligence and ninth conference on Innovative applications of artificial intelligence*, AAAI'97/IAAI'97, pages 10–15. AAAI Press.

Weiner, B. (1985). An Attributional Theory of Achievement Motivation and Emotion. *Psychological Review*, **92**(4), 548–73.

Wiebe, J. M. (1994). Tracking point of view in narrative. *Computational Linguistics*, **20**(2), 233–287.

Zajonc, R. B. (1984). On the primacy of affect. *American Psychologist*, **39**(2), 117–123.

Zhe, X. and Boucouvalas, A. (2002). *Text-to-Emotion Engine for Real Time Internet CommunicationText-to-Emotion Engine for Real Time Internet Communication*, pages 164–168.