Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion 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

Abstract

Even though considerable attention has been given to semantic orientation of words 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 we create a high-quality, moderate-sized emotion lexicon using Mechanical Turk. In addition to questions about emotions evoked by 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 perform an extensive analysis of the annotations to better understand the distribution of emotions evoked by terms of different parts of speech. We identify which emotions tend to be evoked simultaneously by the same term and show that certain emotions indeed go hand in hand.

1 Introduction

When analyzing text, automatically detecting emotions such as joy, sadness, fear, anger, and surprise is useful for a number of purposes, including identifying blogs that express specific emotions towards the topic of interest, identifying what emotion a newspaper headline is trying to evoke, and devising automatic dialogue systems that respond appropriately to different emotional states of the user. Often different emotions are 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. Therefore an **emotion lexicon**—a list of emotions and words that are indicative of each emotion—is likely to be useful in identifying emotions in text.

Words 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, but the emotion lexicon will be a useful component for any sophisticated emotion detecting algorithm. The lexicon will also be useful for evaluating automatic methods that identify the emotions evoked by 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) for six basic emotions and the General Inquirer (GI) (Stone et al., 1966), which categorizes words into a number of categories, including positive and negative semantic orientation.

Amazon has an online service called Mechanical Turk that can be used to obtain a large amount of human annotation in an efficient and inexpensive manner (Snow et al., 2008; Callison-Burch, 2009).[1] 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 moderate-sized English emotion lexicon by manual

[1] https://www.mturk.com/mturk/welcome

annotation through Amazon's Mechanical Turk service. This dataset, which we will call **EmoLex**, is many times as large as the only other known emotion lexicon, WordNet Affect Lexicon. More importantly, the terms in this lexicon are carefully chosen to include some of the most frequent nouns, verbs, adjectives, and adverbs. Beyond unigrams, it has a large number of commonly used bigrams. We also include some words from the General Inquirer and some from WordNet Affect Lexicon, to allow comparison of annotations between the various resources.

We perform an extensive analysis of the annotations to answer several questions that have not been properly addressed so far. For instance, how hard is it for humans to annotate words with the emotions they evoke? What percentage of commonly used terms, in each part of speech, evoke an emotion? Are emotions more commonly evoked by nouns, verbs, adjectives, or adverbs? Is there a correlation between the semantic orientation of a word and the emotion it evokes? Which emotions tend to go together; that is, which emotions are evoked simultaneously by the same term? This work is intended to be a pilot study before we create a much larger emotion lexicon with tens of thousands of terms.

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). Complex emotions can be viewed as combinations of these basic emotions.

2 Related work

WordNet Affect Lexicon (Strapparava and Valitutti, 2004) has a few hundred words annotated with the emotions they evoke.[2] 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 General Inquirer (Stone et al., 1966) has 11,788 words labeled with 182 categories of word tags, including positive and negative semantic orientation.[3] 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.

Work in emotion detection can be roughly classified into that which looks for specific emotion denoting words (Elliott, 1992), that which determines tendency of terms to co-occur with seed words whose emotions are known (Read, 2004), that which uses hand-coded rules (Neviarouskaya et al., 2009), and that which uses machine learning and a number of emotion features, including emotion denoting words (Alm et al., 2005).

Much of this recent work focuses on six emotions studied by Ekman (1992). These emotions— joy, sadness, anger, fear, disgust, and surprise— are a subset of the eight proposed in Plutchik (1980). We focus on the Plutchik emotions because the emotions can be naturally paired into opposites—joy–sadness, anger–fear, trust–disgust, and anticipation–surprise. Natural symmetry apart, we believe that prior work on automatically computing word–pair antonymy (Lin et al., 2003; Mohammad et al., 2008; Lobanova et al., 2010) can now be leveraged in automatic emotion detection.

3 Emotion annotation

In the subsections below we present the challenges in obtaining high-quality emotion annotation, how we address those challenges, how we select the target terms, and the questionnaire we created for the annotators.

3.1 Key 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 on 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

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

[3] http://www.wjh.harvard.edu/∼inquirer

word sense to the annotator. Long definitions will take time to read and limit the number of annotations we can obtain for the same amount of resources. Further, we do not want to bias the annotator towards an emotion through the definition. We want the users to annotate a word only if they are already familiar with it and know its meanings. And lastly, we must ensure that malicious and erroneous annotations are rejected.

3.2 Our solution

In order to overcome the challenges described above, before asking the annotators questions about what emotions are evoked by a target term, we first present them with a word choice problem pertaining to the target. They are provided with four different words and asked which word is closest in meaning to the target. 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. 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 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). Published thesauri, such as *Roget's* and *Macquarie*, divide the vocabulary into about a thousand categories, which may be interpreted as coarse senses. If a word has more than one sense, then it can be found in more than one thesaurus category. Each category also has a head word which best captures the meaning of the category.

Most of the target terms chosen for annotation are restricted to those that are listed in exactly one thesaurus 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.

Only a small number of the words in the WordNet Affect Lexicon are listed in exactly one thesaurus category (have one sense), and so we included target terms that occurred in two thesaurus categories as well. For these questions, we listed head words from both the senses (categories) as two of the alternatives (probability of a random choice being correct is 50%). Depending on the alternative chosen, we can thus determine the sense for which the subsequent emotion responses are provided by the annotator.

3.3 Target terms

In order to generate an emotion lexicon, we first identify a list of words and phrases for which we want human annotations. We chose the *Macquarie Thesaurus* as our source pool for unigrams and bigrams. Any other published dictionary would have worked well too. However, 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 of unigrams and bigrams we chose those that occur frequently in the Google n-gram corpus (Brants and Franz, 2006). Specifically we chose the 200 most frequent n-grams in the following categories: noun unigrams, noun bigrams, verb unigrams, verb bigrams, adverb unigrams, adverb bigrams, adjective unigrams, adjective bigrams, words in the General Inquirer that are marked as having a negative semantic orientation, words in General Inquirer that are marked as having a positive semantic orientation. When selecting these sets, we ignored terms that occurred in more than one *Macquarie Thesaurus* category. Lastly, we chose all words from each of the six emotion categories in the WordNet Affect Lexicon that had at most two senses in the thesaurus (occurred in at most two thesaurus categories). The first and second column of Table 1 list the various sets of target terms as well as the number of terms in each set for which annotations were requested. **EmoLex**$_{\textbf{Uni}}$ stands for all the unigrams taken from the thesaurus. **EmoLex**$_{\textbf{Bi}}$ refers to all the bigrams. **EmoLex**$_{\textbf{GI}}$ are all the words taken from the General Inquirer. **EmoLex**$_{\textbf{WAL}}$ are all the words taken from the WordNet Affect Lexicon.

3.4 Mechanical Turk HITs

An entity submitting a task to Mechanical Turk is called the **requester**. A requester first breaks the task into small independently solvable units called **HITs (Human Intelligence Tasks)** and uploads them on the Mechanical Turk website. The requester specifies the compensation that will be paid for solving each HIT. The people who provide responses to these HITs are called **Turkers**. The requester also specifies the number of different Turkers that are to annotate each HIT. 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 Table 1. Each HIT has a set of questions, all of which are to be answered by the same person. We requested five different assignments for each HIT (each HIT is to be annotated by five different Turkers). Different HITS may be attempted by different Turkers, and a Turker may attempt as many HITs as they wish. Below is an example HIT for the target word “startle”.

Title: Emotions evoked by words
Reward per HIT: $0.04
Directions: Return HIT if you are not familiar with the prompt word.

Prompt word: **startle**

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

- automobile
- shake
- honesty
- entertain

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

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

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

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

4. How much does the word *startle* evoke or produce the emotion joy (for example, *happy* and *fun* may strongly evoke joy)?

- *startle* does not evoke joy
- *startle* weakly evokes joy
- *startle* moderately evokes joy
- *startle* strongly evokes joy

[Questions 5 to 11 are similar to 4, except that joy is replaced with one of the other seven emotions: sadness (*failure* and *heart-break*); fear (*horror* and *scary*); anger (*rage* and *shouting*); trust (*faith* and *integrity*); disgust (*gross* and *cruelty*); surprise (*startle* and *sudden*); anticipation (*expect* and *eager*).]

Before going live, the survey was approved by the ethics committee at the National Research Council Canada.

	# of terms		**Annotns.**
EmoLex	**Initial**	**Master**	**per word**
EmoLex$_{\text{Uni}}$:			
adjectives	200	196	4.7
adverbs	200	192	4.7
nouns	200	187	4.6
verbs	200	197	4.7
EmoLex$_{\text{Bi}}$:			
adjectives	200	182	4.7
adverbs	187	171	4.7
nouns	200	193	4.7
verbs	200	186	4.7
EmoLex$_{\text{GI}}$:			
negatives in GI	200	196	4.7
positives in GI	200	194	4.8
EmoLex$_{\text{WAL}}$:			
anger terms in WAL	107	84	4.8
disgust terms in WAL	25	25	4.8
fear terms in WAL	58	58	4.8
joy terms in WAL	109	92	4.8
sadness terms in WAL	86	73	4.7
surprise terms in WAL	39	38	4.7
Union	**2176**	**2081**	**4.75**

Table 1: 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.

4 Annotation analysis

The first set of emotion annotations on Mechanical Turk were completed in about nine days. The Turkers spent a minute on average to answer the questions in a HIT. This resulted in an hourly pay of slightly more than $2.

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 500 of the 10,880 assignments (2,176 × 5) included at least one unanswered question. These assignments were discarded and rejected. More than 85% of the remaining assignments had the correct answer for the word choice question. This was a welcome result showing that, largely, the annotations were done in a responsible 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, HITs for words not familiar to the annotator were attempted. We discarded and rejected *all* assignments by such annotators (not just the assignments for which they got the word choice question wrong).

HITs pertaining to all the discarded assignments were uploaded for a second time on Mechanical Turk and the validation process was repeated. After the second round, we had three or more valid assignments for 2081 of the 2176 target terms. We will refer to this set of assignments as the **master set**. We create the emotion lexicon from this master set containing 9892 assignments from about 1000 Turkers who attempted 1 to 450 assignments each. About 100 of them provided 20 or more assignments each (more than 7000 assignments in all). The master set has, on average, about 4.75 assignments for each of the 2081 target terms. (See Table 1 for more details.)

4.1 Emotions evoked by words

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 2 lists the percent of 2081 target terms assigned a majority class of no, weak, moderate, and strong emotion. For example, it tells us that 7.6% of the target terms strongly evoke joy. The table also presents an average of the numbers in each column (micro average). Observe that the percentages for individual emotions do not vary greatly from the average. The last row lists the percent of target terms that 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 30.1% of the target terms strongly evoke at least one of the eight basic emotions.

	Intensity			
Emotion	no	weak	moderate	strong
anger	78.8	9.4	6.2	5.4
anticipation	71.4	13.6	9.4	5.3
disgust	82.6	8.8	4.9	3.5
fear	76.5	11.3	7.3	4.7
joy	72.6	9.6	10.0	7.6
sadness	76.0	12.4	5.8	5.6
surprise	84.8	7.9	4.1	3.0
trust	73.3	12.0	9.8	4.7
micro average	**77.0**	**10.6**	**7.2**	**5.0**
any emotion	**17.9**	**23.4**	**28.3**	**30.1**

Table 2: Percent of 2081 terms assigned a majority class of no, weak, moderate, and strong emotion.

Emotion	**% of terms**
anger	15.4
anticipation	20.9
disgust	11.0
fear	14.5
joy	21.9
sadness	14.4
surprise	9.8
trust	20.6
micro average	**16.1**
any emotion	**67.9**

Table 3: Percent of 2081 target terms that are evocative.

Even though we asked Turkers to annotate emotions at four levels of intensity, practical NLP applications often require only two levels—evoking particular emotion (**evocative**) or not (**non-evocative**). 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-evocative bin, all moderate- and strong-intensity assignments in the evocative bin, and then choosing the bin with the majority assignments. Table 3 gives percent of target terms considered to be

EmoLex	**anger**	**anticipation**	**disgust**	**fear**	**joy**	**sadness**	**surprise**	**trust**	**any**
EmoLex$_{\text{Uni}}$:									
adjectives	12	21	8	11	30	13	10	19	72
adverbs	12	16	7	8	21	6	11	25	65
nouns	4	21	2	9	16	3	3	21	47
verbs	12	21	7	11	15	12	11	17	56
EmoLex$_{\text{Bi}}$:									
adjectives	12	24	8	10	26	14	7	18	64
adverbs	3	26	1	5	15	4	8	25	54
nouns	9	30	6	12	15	6	2	24	56
verbs	8	34	2	5	29	6	9	28	67
EmoLex$_{\text{GI}}$:									
negatives in GI	45	5	34	35	1	37	11	2	78
positives in GI	0	23	0	0	48	0	6	47	77
EmoLex$_{\text{WAL}}$:									
anger terms in WAL	**90**	2	54	41	0	32	2	0	91
disgust terms in WAL	40	4	**92**	36	0	20	8	0	96
fear terms in WAL	25	17	31	**79**	0	36	34	0	87
joy terms in WAL	3	32	3	1	**89**	1	18	38	95
sadness terms in WAL	17	0	9	15	0	**93**	1	1	94
surprise terms in WAL	7	23	0	21	52	10	**76**	7	86

Table 4: Percent of terms, in each target set, that are evocative. Highest individual emotion scores for EmoLex$_{\text{WAL}}$ are shown bold. Observe that WAL fear terms are marked most as fear evocative, joy terms as joy evocative, and so on.

evocative. The last row in the table gives the percentage of terms evocative of some emotion (any of the eight). Table 4 shows how many terms in each category are evocative of the different emotions.

4.1.1 Analysis and discussion

Table 4 shows that a sizable percent of nouns, verbs, adjectives, and adverbs are evocative. Adverbs and adjectives are some of the most emotion inspiring terms and this is not surprising considering that they are used to qualify a noun or a verb. Anticipation, trust, and joy come through as the most common emotions evoked by terms of all four parts of speech.

The **EmoLex$_{\text{WAL}}$** 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 anger evocative, joy terms as joy evocative, and so on. The **EmoLex$_{\text{WAL}}$** rows also indicate which emotions get confused for which, or which emotions tend to be evoked simultaneously by a term. Observe that anger terms tend also to be evocative of disgust. Similarly, fear and sadness go together, as do joy, trust, and anticipation.

The **EmoLex$_{\text{GI}}$** rows rightly show that words marked as negative in the General Inquirer, mostly evoke negative emotions (anger, fear, disgust, and sadness). Observe that the percentages for trust and joy are much lower. On the other hand, positive words evoke anticipation, joy, and trust.

4.1.2 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. Observe that for more than 50% of the terms, at least four annotators agree with each other. Table 5 presents these agreement values. Since many NLP systems may rely only on two intensity values (evocative or non-evocative), we also calculate agreement at that level (Table 6). Observe that for more than 50% of the terms, all five annotators agree with each other, and for more than 80% of the terms, at least four annotators agree. This shows a high degree of agreement on emotion annotations despite no real control over the educational background and qualifications of the annotators.

	Majority class size			
Emotion	two	three	four	five
anger	13.1	25.6	27.4	33.7
anticipation	31.6	35.2	20.7	12.3
disgust	14.0	21.6	29.0	35.1
fear	15.0	29.9	28.6	26.2
joy	17.6	26.4	23.0	32.7
sadness	14.2	24.6	28.1	32.8
surprise	17.0	29.3	32.3	21.2
trust	22.4	27.8	22.4	27.2
micro average	**18.1**	**27.6**	**26.4**	**27.7**

Table 5: Agreement at four intensity levels for emotion (no, weak, moderate, and strong): Percent of 2081 terms for which the majority class size was 2, 3, 4, and 5.

	Majority class size		
Emotion	three	four	five
anger	15.0	25.9	58.9
anticipation	32.3	33.7	33.8
disgust	12.8	24.6	62.4
fear	14.9	25.6	59.4
joy	18.4	27.0	54.5
sadness	13.6	22.0	64.2
surprise	17.5	31.4	50.9
trust	23.9	29.3	46.6
micro average	**18.6**	**27.4**	**53.8**

Table 6: Agreement at two intensity levels for emotion (evocative and non-evocative): Percent of 2081 terms for which the majority class size was 3, 4, and 5.

4.2 Semantic orientation of words

We consolidate the semantic orientation (polarity) annotations in a manner identical to the process for emotion annotations. Table 7 lists the percent of 2081 target terms assigned a majority class of no, weak, moderate, and strong semantic orientation. For example, it tells us that 16% of the target terms are strongly negative. The last row in the table lists the percent of target terms that have some semantic orientation (positive or negative) at the various intensity levels. Observe that 35% of the target terms are strongly evaluative (positively or negatively).

Just as in the case for emotions, practical NLP applications often require only two levels of semantic orientation—having particular semantic orientation or not (**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 8 gives percent of target terms considered to be evaluative. The last row in the table gives the percentage of terms evaluative with respect to some semantic orientation (positive or negative). Table 9 shows how many terms in each category are positively and negatively evaluative.

	Intensity			
Polarity	no	weak	moderate	strong
negative	60.8	10.8	12.3	16.0
positive	48.3	11.7	20.7	19.0
micro average	**54.6**	**11.3**	**16.5**	**17.5**
any polarity	**14.7**	**17.4**	**32.7**	**35.0**

Table 7: Percent of 2081 terms assigned a majority class of no, weak, moderate, and strong polarity.

Polarity	**% of terms**
negative	31.3
positive	45.5
micro average	**38.4**
any polarity	**76.1**

Table 8: Percent of 2081 target terms that are evaluative.

4.2.1 Analysis and discussion

Observe in Table 9 that, across the board, a sizable number of terms are evaluative with respect to some semantic orientation. Interestingly 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 semantic orientation of noun concepts is positive, and that usually it takes a negative adjective to make the phrase negative.

The **EmoLex$_{\mathbf{GI}}$** rows in the two tables show that words marked as having a negative semantic orientation in the General Inquirer are mostly marked as negative by the Turkers. And similarly, the positives in GI are annotated as positive. Again, this is confirmation that the quality of annotation obtained is high. The **EmoLex$_{\mathbf{WAL}}$** rows show that anger, disgust, fear, and sadness terms tend not to have a positive semantic orientation 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.

4.2.2 Agreement

In order to analyze how often the annotators agreed with each other, for each term–emotion pair, we cal-

EmoLex	**negative**	**positive**	**any**
EmoLex$_{\textbf{Uni}}$:			
adjectives	33	55	87
adverbs	29	54	82
nouns	6	44	51
verbs	22	41	62
EmoLex$_{\textbf{Bi}}$:			
adjectives	30	48	78
adverbs	10	52	61
nouns	13	49	61
verbs	12	57	68
EmoLex$_{\textbf{GI}}$:			
negatives in GI	**90**	2	92
positives in GI	2	**91**	91
EmoLex$_{\textbf{WAL}}$:			
anger terms in WAL	96	0	96
disgust terms in WAL	96	0	96
fear terms in WAL	87	3	89
joy terms in WAL	4	92	96
sadness terms in WAL	90	1	91
surprise terms in WAL	23	57	81

Table 9: Percent of terms, in each target set, that are evaluative. The highest individual polarity EmoLex$_{\text{GI}}$ row scores 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.

	Majority class size			
Polarity	two	three	four	five
negative	11.8	28.7	29.4	29.8
positive	21.2	30.7	19.0	28.8
micro average	**16.5**	**29.7**	**24.2**	**29.3**

Table 10: Agreement at four intensity levels for polarity (no, weak, moderate, and strong): Percent of 2081 terms for which the majority class size was 2, 3, 4, and 5.

	Majority class size		
Polarity	three	four	five
negative	11.8	21.2	66.9
positive	23.1	26.3	50.5
micro average	**17.5**	**23.8**	**58.7**

Table 11: Agreement at two intensity levels for polarity (evaluative and non-evaluative): Percent of 2081 terms for which the majority class size was 3, 4, and 5.

culated 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 10 presents these agreement values. Observe that for more than 50% of the terms, at least four annotators agree with each other. Table 11 gives agreement values at the two-intensity level. Observe that for more than 50% of the terms, all five annotators agree with each other, and for more than 80% of the terms, at least four annotators agree.

5 Conclusions

We showed how Mechanical Turk can be used to create a high-quality, moderate-sized, emotion lexicon for a very small cost (less than US$500). 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. We compared a subset of our lexicon with existing gold standard data to show that the annotations obtained are indeed of high quality. A detailed analysis of the lexicon revealed insights into how prevalent emotion bearing terms are among common unigrams and bigrams. We also identified which emotions tend to be evoked simultaneously by the same term. The lexicon is available for free download.[4]

Since this pilot experiment with about 2000 target terms was successful, we will now obtain emotion annotations for tens of thousands of English terms. We will use the emotion lexicon to identify emotional tone of larger units of text, such as newspaper headlines and blog posts. We will also use it 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.

Acknowledgments

This research was funded by the National research Council Canada (NRC). Thanks to Diana Inkpen and Diman Ghazi for early discussions on emotion. Thanks to Joel Martin for encouragement and support. 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 1000 anonymous people who answered the emotion survey with diligence and care.

[4] http://www.purl.org/net/emolex

References

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 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.

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

Thorsten Brants and Alex Franz. 2006. Web 1t 5-gram version 1. *Linguistic Data Consortium*.

Chris Callison-Burch. 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.

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

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

Dekang Lin, Shaojun Zhao, Lijuan Qin, and Ming Zhou. 2003. Identifying synonyms among distributionally similar words. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI-03)*, pages 1492–1493, Acapulco, Mexico.

A. Lobanova, T. van der Kleij, and J. Spenader. 2010. Defining antonymy: A corpus-based study of opposites by lexico-syntactic patterns. *International Journal of Lexicography (in press)*, 23:19–53.

Saif Mohammad, Bonnie Dorr, and Codie Dunn. 2008. Computing word-pair antonymy. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2008)*, pages 982–991, Waikiki, Hawaii.

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

Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka. 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.

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

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

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 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.

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

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-2004)*, pages 1083–1086, Lisbon, Portugal.

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