Using Google n-Grams to Expand Word-Emotion Association Lexicon

Jessica Perrie, Aminul Islam, Evangelos Milios, and Vlado Keselj

Dalhousie University, Faculty of Computer Science
Halifax, NS, Canada B3H 4R2
jssc.perrie@gmail.com,
{islam,eem,vlado}@cs.dal.ca

Abstract. We present an approach to automatically generate a word-emotion lexicon based on a smaller human-annotated lexicon. To identify associated feelings of a target word (a word being considered for inclusion in the lexicon), our proposed approach uses the frequencies, counts or unique words around it within the trigrams from the Google n-gram corpus. The approach was tuned using as training lexicon, a subset of the National Research Council of Canada (NRC) word-emotion association lexicon, and applied to generate new lexicons of 18,000 words. We present six different lexicons generated by different ways using the frequencies, counts, or unique words extracted from the n-gram corpus. Finally, we evaluate our approach by testing each generated lexicon against a human-annotated lexicon to classify feelings from affective text, and demonstrate that the larger generated lexicons perform better than the human-annotated one.

1 Introduction

Problem. Users exchange ideas and opinions by writing blogs, product reviews and comments, producing a massive amount of information. Applications for sentiment and emotion analysis that take advantage of this data to automatically find the feelings conveyed by the word choice, can be used, for example, to track feelings towards a product over time [1].

Consider, for example, the words *delightful* and *gloomy*; according to the National Research Council of Canada (NRC) word-emotion association lexicon, *delightful* is associated with uplifting feelings like anticipation, and joy, while *gloomy* is associated with negative feelings like sadness [1].

While there are hundreds of possible emotions to choose from, many studies have used a small subset of basic emotions. Our study uses emotions as defined by Plutchik: anger, anticipation, disgust, fear, joy, sadness, surprise and trust, because annotating hundreds of emotions would be expensive and difficult, while Plutchik's basic set are well-founded in psychological, physiological and empirical research [1]. They are a superset of the Ekman emotions, which are commonly used in emotion studies [2,3], and are not mostly composed of negative emotions [1]. The sentiments (positive and negative) are also included in our study, but are treated exactly like the emotions. In this paper, both sentiments and emotions are referred together as *feelings*.

A. Gelbukh (Ed.): CICLing 2013, Part II, LNCS 7817, pp. 137–148, 2013.

Sentiment and emotion analysis applications have lexicon- or dictionary-based approaches when they use a general word lexicon as a starting point (and then may refine results with more domain- or feature- specific terms) [4]. Word-emotion lexicons, especially ones created by human-annotators, are essential to evaluate automatic approaches, like the one presented in this paper, that identify emotions associated with additional terms [1].

Motivation. We present an automatic approach to generate word-emotion lexicons using a smaller word-emotion lexicon and the Google n-gram corpus. Automatic approaches, like the one proposed, have many advantages over human-annotated or manual approaches. Although manual approaches tend to be more reliable, automatic approaches require less work and avoid human random error [5]. Furthermore, manually created lexicons are noted for having relatively poor coverage of technical and scientific terms that are essential to analyze research papers [5]. Another major limitation is the additional labour needed to translate the lexicon into each new language [5].

The main advantage of automatic methods is in their creation. Automatic construction approaches expand lexicons by following the smaller lexicon's patterns [5]. Additionally, depending on the similarity of languages and assuming the data needed for that approach is available, the automatic construction can also be applied to generate an emotion lexicon in another language, or plot out the evolution of different words over time [6,7]. Therefore, unlike manual methods, a smaller amount of human work is needed.

Given the advantages of word-emotion lexicons and their use in emotion and sentiment analysis, we developed an approach to generate effective word-emotion association lexicons. Each lexicon was built by comparing the data within the Google n-gram corpus and using a training lexicon of seed words, words where the associated feeling is already known. Training sets in our study are subsets of the NRC lexicon. In Section 4, we present three different methods with two variations of finding the feeling associations of target words in novel ways: the frequency of surrounding feeling associated words, the number of times surrounding feeling associated words occur, and the number of times unique surrounding feeling associated words occur. Finally, in Section 5, the lexicons generated by our methods are evaluated against the testing lexicon in a simple feeling classification task.

2 Related Work

In this section we present a description of related work: sentiment or emotion lexicons that were expanded using automatic methods.

In [5], Turney presented an unsupervised learning algorithm to find synonyms by comparing their Pointwise Mutual Information collected by Information Retrieval (PMI-IR) which measures the association between two terms, a target word and a possibly related word, by finding their probabilities of appearing together within the same document [5]. As the definition of "document" became smaller and meant the two words must appear within ten terms of each other (within a 10-gram), it was observed that the results for matching each synonym improved. In our study we used trigrams. Turney also used PMI-IR to classify the sentiment

at document-level of reviews based on the average semantic orientation of their phrases. The orientation for each phrase was found by calculating the mutual information between it and the word "excellent" and "poor" [8]. Most similar to our work, Turney extended this idea further to find the polarity of target words by looking at their statistical association based on their co-occurrence with fourteen positive or negative seed words that kept their polarity no matter the context [9]. To measure co-occurrence, he counted when the target word was within ten words of the polarity word. We extend this idea by only considering appearances within three words and using over 10,000 words as seeds.

For emotion lexicons, automatic approaches have used large corpora from the web. In [10], Yang et al. used weblog corpora and a collocation model to build an emotion lexicon from online articles. Blog data were used because they were timestamped and because blogs can express emotional states of users who may use emoticons to represent their feelings [10]. A training set was used to measure the word's associations with one of forty possible emoticons–each emoticon represented an emotion–by a modified version of Pointwise Mutual Information. This approach had two variations by choosing the top n collocated word-emotion pairs; the first variation had 4,776 entries with 25,000 word sense associations, and the second had 11,243 entries and 50,000 sense pairs [10]. In their comparison of the two lexicons, they observed that the larger one had better performance in classifying emotions.

The use of the NRC word sense lexicon with Google n-grams was briefly touched on in [6] in which is it stated that "[w]ords found in proximity of target entities can be good indicators of emotions associated with the targets." Using Google n-grams frequency data from books scanned up to July 15, 2009, Mohammad placed the n-grams into bins of five years and measured the percentage of different emotion words that appeared in 5-grams with certain target words [6]. This idea is similar to our work, except we expand on it to build a lexicon with emotion and sentiment associations, but do not consider changes of the associations over time, although that is a possible future application.

3 Resources

NRC Word-Emotion Association Lexicon. The NRC word-emotion association lexicon version 0.92 is used in our study to build and test our proposed approach. It contains about 14,200 individual terms and their associations to each of the eight Plutchik basic emotions and two sentiments: anger, anticipation, disgust, fear, joy, negative, positive, sadness, surprise, trust. Each word in the lexicon has ten 2-level values indicating its association for each feeling. For example, a word like *torture* has the values $\langle 1, 1, 1, 1, 0, 1, 0, 1, 0, 0 \rangle$, which indicates there exists an association between *torture* and the feelings anger, anticipation, disgust, fear, negative, and sadness, while there are no associations to feelings of joy, positive, surprise nor trust.

The NRC lexicon was made by dividing the annotation work to a large network of laborers through Mechanical Turk [1]. The NRC lexicon terms were

chosen from the most frequent English nouns, verbs, adjectives and adverbs selected from the *Macquarie Thesaurus* and Google n-gram corpus, and from other emotion lexicons like the General Inquirer and the WordNet Affect Lexicon [1].

Google N-Gram Corpus. The Google Web 1T n-gram corpus, contributed by Google Inc., contains English word n-grams (from uni-grams to 5-grams) and their observed frequencies calculated over one trillion words from web page text collected by Google in January 2006. The text was tokenized following the Penn Treebank tokenization, except that hyphenated words, dates, email addresses and URLs are kept as single tokens. The n-grams themselves must appear at least 40 times to be included in the Google n-gram corpus[1].

In October 2009, Google released the Web 1T 5-gram, 10 European Languages Version 1 [11], consisting of word n-grams and their observed frequency for ten European languages: Czech, Dutch, French, German, Italian, Polish, Portuguese, Romanian, Spanish, and Swedish. Thus, it is possible to use our proposed approach to generate lexicons for these languages as well.

Our study uses only trigrams ($n = 3$) from the Google n-gram corpus. Some examples of trigrams provided by the corpus are *he was a* with a supplied frequency 3,683,417; *hehe was a* with 52; and *he was an* with 563,471.

4 The Proposed Approach

The method to develop our proposed approach is shown at high-level in Figure 1. Actions above the second dashed line are explained in this section; actions below the dashed line are explained in the next section where we evaluate our computed lexicon in a feeling classification method.

Description of Approach. To find the feeling associations from each target word, the approach first searches for that word in the n-gram corpus, finds all the n-grams that contain the target word, and, within each n-gram, finds surrounding words from the training lexicon which we call `assoc_word`. It then generates three vectors of size ten (one value for each of the ten feelings) for each target word: `assoc_freq`, `assoc_counts` and `assoc_unique` as defined in Figure 2. To normalize the results, the totals for each of these sums where a feeling is not associated are also detected, respectively as `assoc_not_freq`, `assoc_not_counts`, `assoc_not_unique`.

Each value in each of the three vectors is normalized by taking it over the sum of itself and its inverse (e.g., normalized `assoc_freq[joy]` = `assoc_freq[joy]` / (`assoc_freq[joy]` + `assoc_not_freq[joy]`) and is farther referred to as a "feeling association strength". In our approach three different methods are used for each of the three normalized vectors. If the feeling association strength of a certain feeling for a target word is higher than a tuned parameter, threshold-1 (as defined by our variations) then that target word is classified as having an association to that feeling. Alternately, if that feeling association strength is below another threshold-2, then that target word is identified as *not* having an

[1] Details can be found at `www.ldc.upenn.edu/Catalog/docs/LDC2006T13/readme.txt`

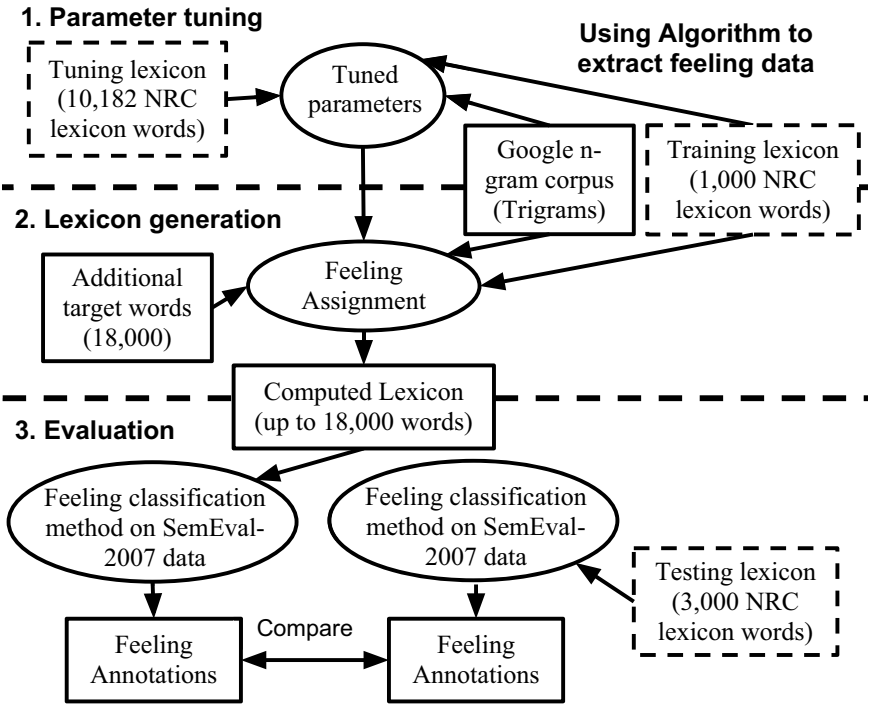

Fig. 1. High-level overview of proposed approach and evaluation. Parameter tuning and lexicon generation are carried out using three different methods in two variations; therefore, six different lexicons are produced. Each lexicon is evaluated against the Testing lexicon by measuring their performance on the SemEval-2007 text using the same classification method: keyword spotting.

association to that feeling. Feeling associations for a target word may also be classified as "unknown" because of a lack of sufficient data.

Our approach can be used with any type of n-grams–bigram, trigram, 4-gram or 5-gram; however, from our experiments, we found that trigrams produced the best results–a greater difference in feeling association strengths between target words with an association and target words without an association. Therefore, we believe that 4-grams and 5-grams are less suitable because they included too much noise in the form of surrounding words that were not indicative of the target word's associated feelings. These results also suggest that bigrams don't contain enough surrounding words to classify the target word.

Method 1: Feeling association strength: normalized `assoc_freq`. This idea follows the idea that frequencies of surrounding words in close proximity to a target word are indicative of its associated feelings [6]. Surrounding words that have a high frequency of occurring with the target word are assumed to share their associated feelings more strongly.

Extracting feeling data using frequencies, counts, and unique words from Google n-grams

```
list_of_ngrams: n-grams containing both assoc_word and target_word
feeling: anger, anticipation, disgust, fear, joy, negative, positive,
 sadness, surprise, trust
Derived over assoc_words with assoc. feeling
 assoc_freq[feeling]:       sum of n-gram frequencies in list_of_ngrams
 assoc_counts[feeling]:     total counts in list_of_ngrams
 assoc_unique[feeling]:     total unique words in list_of_ngrams
Derived over assoc_words without assoc. feeling
 assoc_not_freq[feeling]:   total freq. in list_of_ngrams
 assoc_not_counts[feeling]: total counts in list_of_ngrams
 assoc_not_unique[feeling]: total unique words in list_of_ngrams

for each (ngram_phrase, ngram_freq) in list_of_ngrams
 for each assoc_word in ngram_phrase
  for each feeling
   if (assoc_word is associated to this feeling in training lexicon)
   add ngram_freq to assoc_freq[feeling];
   add 1 to assoc_counts[feeling];
   if (assoc_word wasn't yet encountered in list_of_ngrams)
    add 1 to assoc_unique[feeling];
  else
   add ngram_freq to assoc_not_freq[feeling];
   add 1 to assoc_not_counts[feeling];
   if (assoc_word wasn't yet encountered in list_of_ngrams)
    add 1 to assoc_not_unique[feeling];
```

Fig. 2. Given a target word, use the training lexicon to find the total frequency, the total counts and the total unique words of the feelings (emotions and sentiments) of surrounding words in the Google n-gram corpus.

Method 2: Feeling association strength: normalized ***assoc_counts****.* Method 2 measures the variety of words in different n-grams listed in the trigram corpus.

Method 3: Feeling association strength: normalized ***assoc_unique****.* The idea comes from observing the data, and assuming that if a greater number of different surrounding words convey the same feeling, then that feeling is more strongly associated with the target word.

Validation. The first challenge with our approach was dealing with scarcity of data within the n-gram corpus [5]. Furthermore, this step is needed because we found removing target words with scarce data reduced the number of falsely detected associations in the tuning lexicon. Consider the relatively obscure word *obi*, which, according to the NRC lexicon, has associations with *disgust*, *fear*, and *negative*. Within the trigram corpus, no surrounding words of *obi* with associations to *disgust* are spotted, which incorrectly suggests that *obi* is not associated with *disgust*. Therefore, we need a baseline validation to ensure that

enough data is available before we can classify an associated feeling for better and more precise results.

Thus, for each possible feeling of each target word, if its `assoc_unique` was smaller than $10 \log_{10}$ of the number of words with that feeling in the training lexicon, than that word-feeling association was declared unknown. However, this specific value is arbitrary because while other thresholds work better for some feelings, they do not work better for all and the amount of improvement in each result depended on the feeling type. Future work could be done in identifying this threshold more specifically.

Tuned Parameters. Each method has two variations with bounds based off the two different tuned parameter sets. The first variation's goal is to maximize the number of true values found between the computed lexicons and the human-annotated lexicon. The second variation's goal is to maximize the precision and recall of the lexicons produced when compared to the human-annotated lexicon. For each feeling, it was observed that there was a range where feeling association strengths of the tuning lexicon words with an association, and the words without an association, would overlap. The second variation works by declaring most word-feeling associations with feeling association strengths within this range as being unknown, which produces a smaller number of true values in the computed lexicon.

Variation 1: Threshold: [0.1, 0.1). If the feeling association strength for a certain feeling of a target word is ≥ 0.1, the target word is classified as having an association to that feeling; else, the target word was classified as not associated to that feeling. This value is arbitrary, because other thresholds produce similar results; however, after observing the different tuning lexicon words, most feeling association strengths with an association were over this threshold, while most feeling association strengths without an association were below.

Variation 2: Threshold: (0.05, 0.15). We expand the threshold by 0.05 to reduce the number of falsely classified associations. If the feeling association strength for a certain feeling of a target word is ≥ 0.15, the word is classified with an association to that feeling. If the feeling association strength is ≤ 0.05, then the word is classified with not having an association to that feeling. Finally, if the feeling association strength is between 0.05 and 0.15, the association of the target word to that feeling remains unknown.

Results of Comparing Human-annotated Feelings with Computed Feelings on Tuning Lexicon Words. The results of comparing the tuning lexicon with the computed lexicon built using the same words with each method at each variation are presented in Table 1. We measured the precision (p)–the number of true and detected word-feeling associations over the number of detected word-feeling associations; the recall (r)–the number of true and detected word-feeling associations over the number of true word-feeling associations in the tuning lexicon; the f-measure (f)–an average of the precision and recall as outlined in the first equation in Eq. 1; and the accuracy (a)–a measurement involving detected true associations (TP) and no associations (TN), and the number of falsely

detected associations (FP) and no associations (FN), as shown in the second equation in Eq. 1.

$$f = \frac{2 * p * r}{p + r} \qquad a = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{1}$$

Table 1. Matching of the hand-annotated feelings of the words in the tuning lexicon with computed feelings of the same words based on the training lexicon and Google n-grams for the selected parameters using each method (M). In each variation (Var.), for each feeling, **boldface** values are the highest of their type: f-measure (f) or accuracy (a). The interval next to each variation name indicates the "middle area" where feeling association strengths (as measured by the type of method) are ignored if they fall within this interval, or indicate an association if greater, or no association if lower.

	Var.1: [0.1, 0.1)						**Var.2: (0.05, 0.15)**					
	M1:freq		**M2:count**		**M3:uniq**		**M1:freq**		**M2:count**		**M3:uniq**	
Feeling	f	a	f	a	f	a	f	a	f	a	f	a
anger	0.53	0.84	0.57	0.86	**0.60**	**0.87**	0.64	0.91	0.68	0.93	**0.81**	**0.94**
anticipation	0.20	0.72	0.24	0.77	**0.28**	**0.85**	0.28	0.76	**0.40**	**0.85**	0.33	0.71
disgust	0.39	0.84	0.50	0.88	**0.56**	**0.90**	0.39	0.90	0.46	0.95	**0.67**	**0.98**
fear	0.49	0.78	0.47	0.79	**0.52**	**0.79**	0.50	0.81	0.59	**0.84**	**0.61**	0.71
joy	0.30	0.78	**0.33**	0.81	0.33	**0.88**	0.33	0.83	0.45	0.89	**0.71**	**0.97**
negative	**0.58**	**0.66**	0.58	0.60	0.50	0.40	0.68	**0.67**	**0.70**	0.58	0.65	0.48
positive	**0.33**	**0.32**	0.31	0.21	0.30	0.18	**0.40**	**0.27**	0.36	0.22	0.32	0.19
sadness	0.36	0.78	0.39	0.81	**0.47**	**0.85**	0.50	0.85	0.61	0.91	**0.75**	**0.93**
surprise	**0.20**	0.90	0.15	0.92	0.12	**0.93**	**0.17**	0.94	0.13	**0.96**	0.00	0.95
trust	**0.30**	**0.57**	0.29	0.49	0.28	0.44	0.36	**0.54**	0.45	0.36	**0.48**	0.32

Discussion. In all variations and methods, negative feelings like *anger*, *disgust*, *fear*, *negative* and *sadness* tend to have higher f-measures than positive or neutral feelings like *anticipation*, *joy*, *positive*, *surprise* and *trust*. From observing the data, positive feeling association strengths between the words with associations and words without associations were less different. This result suggests that most word-feeling combinations in trigrams are related to expressing negative emotions [2]. Additionally, the poor results for positive feelings may be because the training lexicon has fewer words with associations to them, and thus, did not have enough positive feeling words to spot. It is also possible that words surrounding positive target words in trigrams don't reflect positive feelings.

With the exception of *trust*, the sentiments *negative* and *positive* have lower values of f-measure and accuracy, suggesting that polarities may act different than emotions, and thus, should be treated differently.

From the results for Variation 1 and Variation 2, Method 3 produces the highest results, which supports its design. Graphs of the feeling association strengths for Method 1 and Method 2 did not have as great a difference between words with an association and words without an association.

Feeling Assignment. To test our methods, we created a lexicon for each method with each variation using the 3,000 target words from the testing lexicon

and 15,000 words commonly used within the Google unigram corpus that were not included in the NRC lexicon. The number of word-feeling associations of each lexicon is presented in Table 2.

Table 2. Summary of the number of words in each of the generated lexicons as classified by the different methods (M) and variations (Var.). Variation 2 has more words with only unknown associations, because it assigns most word-feeling associations with feeling association strength within an overlapping area as unknown. The intervals associated with the variations are explained in Table 1.

	Var.1: [0.1, 0.1)			**Var.2: (0.05, 0.15)**		
Feelings	**M1:freq**	**M2:count**	**M3:uniq**	**M1:freq**	**M2:count**	**M3:uniq**
anger	808	824	794	467	375	208
anticipation	1696	1528	1039	735	318	34
disgust	332	305	198	150	90	16
fear	1545	1524	1692	893	727	454
joy	1227	1194	778	612	424	87
negative	4854	5656	7382	3160	3356	4156
positive	10788	12529	13137	7966	9601	12090
sadness	844	771	628	442	292	100
surprise	143	59	7	54	8	0
trust	4084	4982	5775	2109	1861	1135
only unknown	4391	4391	4391	6378	6149	4988

5 Evaluation

We tested each lexicon against a baseline, the NRC testing lexicon. Because we are only interested in testing the generated lexicon and not the classification method, each lexicon was put through the same naive lexicon-based feeling classifier, keyword spotting, using data from the SemEval-2007 Task 14: Affective Text as shown from Figure 1. In future, we will consider a more accurate lexicon-based method.

SemEval-2007 Task 14: Affective Text data is a collection of news titles (which are often written to provoke readers' emotions) from newspapers and news websites like Google news and CNN [3]. All 1,250 headlines are human-annotated with measures of six emotions–anger, disgust, fear, joy, sadness and surprise–and a sentiment–negative, positive or neutral. The agreement using the Pearson correlation measure among the annotators for each feeling varied, but was lowest for *disgust* and *surprise*. Because emotions *anticipation* and *trust* are not included in this dataset, they are not included in the evaluation.

The human-annotated measurements of feelings are mapped to 1 (meaning there is an association between the feeling and headline) or 0 (meaning there is not an association), in accordance with the coarse-grained evaluation in the SemEval task. In our evaluation, we use all 1250 sentences from this dataset.

Approach. The emotion and sentiment classification method used in this evaluation was keyword spotting as shown in Figure 3.

Keyword spotting procedure to classify feelings in headlines

```
Preprocessing headline (transform to lowercase, remove punctuation, and
 tokenize words)
for each feeling
 for each word in headline
   if (lexicon entry for word-feeling association is not unknown)
    add 1 to count[feeling]
    if (word is associated to feeling)
     add 1 to temp[feeling]
    else add 0 to temp[feeling]
 if (temp[feeling]/count[feeling] is greater or equal to 0.5)
  headline is associated to feeling
 else headline is not associated to feeling
```

Fig. 3. Given a headline, use lexicon to find if associated feeling exists

Results of Evaluation. The results of using the testing lexicon and our generated lexicons are displayed in Table 3. The result formulas are like the equations Eq. 1, except we're considering headline-feeling associations instead of word-feeling associations. (Software to recreate these results may be found at http://www.CICLing.org/2013/data/138)

While the results of the testing lexicon may seem too low to properly judge the lexicons, both the f-measures and accuracies of the emotions are within 0.0170 of the lower bounds for the results of the systems that participated in the SemEval task. For all emotions over those systems, the average f-measure was 0.0993, the average accuracy was 0.8791, the highest f-measure was 0.3038, and the highest accuracy was 0.9730 [3].

Table 3. Results from the feeling classification performance of the computed lexicons from each variation (Var.) and method (M), and the human-annotated testing lexicon on the SemEval-2007 Task 14 data set. If a value is **boldface** under any of the Method lexicons, that value is greater than the Testing lexicon (Test.) for either f-measure (f) or accuracy (a).

			Var.1: [0.1, 0.1)						**Var.2: (0.05, 0.15)**					
	Test.		**M1:freq**		**M2:count**		**M3:uniq**		**M1:freq**		**M2:count**		**M3:uniq**	
Feeling	f	a	f	a	f	a	f	a	f	a	f	a	f	a
anger	0.06	0.93	**0.11**	0.86	**0.12**	0.87	**0.11**	0.89	**0.12**	0.89	**0.11**	0.91	**0.07**	**0.96**
disgust	0.00	0.97	**0.21**	0.95	**0.02**	0.93	**0.05**	0.97	0.00	0.96	0.00	0.97	0.00	**0.98**
fear	0.17	0.87	**0.27**	0.77	**0.25**	0.77	**0.25**	0.74	**0.25**	0.81	**0.19**	0.82	**0.21**	**0.87**
joy	0.12	0.85	**0.14**	0.80	0.10	0.81	0.06	**0.87**	0.08	0.83	0.07	**0.87**	0.00	**0.88**
sadness	0.08	0.86	**0.21**	0.80	**0.22**	0.82	**0.17**	0.86	**0.12**	**0.86**	**0.13**	0.85	0.00	**0.88**
surprise	0.05	0.93	**0.07**	0.93	0.00	**0.95**	0.00	**0.96**	0.00	**0.95**	0.00	**0.96**	0.00	**0.96**
positive	0.15	0.81	**0.22**	0.43	**0.22**	0.38	**0.22**	0.37	**0.22**	0.51	**0.22**	0.46	**0.22**	0.38
negative	0.19	0.74	**0.39**	0.60	**0.40**	0.56	**0.39**	0.48	**0.34**	0.66	**0.37**	0.67	**0.38**	0.56
Average	0.10	0.87	**0.20**	0.77	**0.17**	0.76	**0.16**	0.77	**0.14**	0.81	**0.14**	0.81	**0.11**	0.81
Highest	0.19	0.97	**0.39**	0.95	**0.40**	0.95	**0.39**	0.97	**0.34**	0.96	**0.37**	0.97	**0.38**	**0.98**

Discussion. From the results in Table 3, the larger lexicons generated in Variation 1 give higher f-measures than the lexicons in Variation 2, while the smaller more accurate lexicons in Variation 2 give higher accuracies than the lexicons in Variation 1. This latter result likely occurs because most of the human-annotated headlines are not associated to a feeling (when using the 1 or 0 mapping). Because smaller lexicons would only classify a smaller number of words within the headlines, a larger number of these likely-not-associated-to-a-feeling headlines would remain, by default, not associated to any feelings, and thus increase the accuracy. Compared to the testing lexicon, Variations 2's results are still notable because its lexicons are larger than in the testing set (so fewer headlines are left by default with no associations), and yet, it still produces higher accuracies. Variation 1's better performance in f-measures suggest that larger lexicons, with their greater coverage of possible words, increase the precision and recall of feeling classification tasks to find if associations exist, but are less accurate when finding when associations do not exist.

Compared to the testing lexicon, the lexicon computed with Method 1, Variation 1, has the highest f-measures, despite having some of the lowest f-measures in Table 1, which suggests that other tests are needed besides f-measure and accuracy to find the best approach to generate a lexicon. Because Method 1 was based on frequencies, our results add credibility to other frequency-based automatic approaches like Pointwise Mutual Information. Method 3, Variation 2 has the highest accuracies; however, Method 3 does not have as high f-measures, which does not help in identifying if headlines have feeling associations.

For both variations, the computed lexicons performed better for negative emotions like *anger*, *fear* and *sadness*, which farther suggests negative emotions are expressed more in the trigram corpus [2]. The poor performance of *disgust* and *surprise* may result because they had the lowest agreement between the human annotators of the SemEval Affective Text, and, as shown in Table 2, all generated lexicons had a relatively smaller number of word-feeling associations for these feelings, suggesting less accuracy to correctly identify them.

Overall, these results suggest that larger lexicons created using automatic methods can perform feeling classification tasks better than smaller human-annotated ones in terms of f-measure and accuracy. Large lexicons created with less accurate methods (Variation 1) tend to have better f-measure, while smaller lexicons (but still larger than the human-annotated lexicon) with better f-measures (Variation 2) tend to have better accuracy.

6 Conclusion

We proposed a new approach to generate a lexicon by automatic means using data provided by the Google n-grams corpus and NRC lexicon. Our approach consists of using the frequencies of n-grams, the counts of surrounding words or the unique counts of surrounding words at two different variations of tuned parameters to produce lexicons with a relatively large or small number of word-feeling associations. The larger lexicons had more words, but less accuracy than the smaller lexicons. From our evaluation of these computed lexicons against

the testing lexicon, we provide evidence that suggests larger lexicons generated with less accurate methods perform better, and that more measurements, in conjunction with precision, recall and accuracy, are needed to find an approach to generate an effective lexicon.

In addition to the future work mentioned in previous sections, we intend to look into using the n-grams farther by searching for the context around each target word and then searching for an identical context where a word from the training lexicon is used. We will also look into handling target words differently depending on how they are used or their parts of speech.

References

1. Mohammad, S., Turney, P.: Crowdsourcing a word–emotion association lexicon. Computational Intelligence (2012)
2. Kozareva, Z., Navarro, B., Vázquez, S., Montoyo, A.: Ua-zbsa: a headline emotion classification through web information. In: Proceedings of the 4th International Workshop on Semantic Evaluations, SemEval 2007, pp. 334–337. Association for Computational Linguistics, Stroudsburg (2007)
3. Strapparava, C., Mihalcea, R.: Semeval-2007 task 14: affective text. In: Proceedings of the 4th International Workshop on Semantic Evaluations, SemEval 2007, pp. 70–74. Association for Computational Linguistics, Stroudsburg (2007)
4. Lu, Y., Castellanos, M., Dayal, U., Zhai, C.: Automatic construction of a context-aware sentiment lexicon: an optimization approach. In: Proceedings of the 20th International Conference on World Wide Web, WWW 2011, pp. 347–356. ACM, New York (2011)
5. Turney, P.D.: Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In: Flach, P.A., De Raedt, L. (eds.) ECML 2001. LNCS (LNAI), vol. 2167, pp. 491–502. Springer, Heidelberg (2001)
6. Mohammad, S.: From once upon a time to happily ever after: tracking emotions in novels and fairy tales. In: Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, LaTeCH 2011, pp. 105–114. Association for Computational Linguistics, Stroudsburg (2011)
7. Amiri, H., Chua, T.S.: Mining sentiment terminology through time. In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management, CIKM 2012, pp. 2060–2064. ACM, New York (2012)
8. Turney, P.D.: Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL 2002, pp. 417–424. Association for Computational Linguistics, Stroudsburg (2002)
9. Turney, P.D., Littman, M.L.: Measuring praise and criticism: Inference of semantic orientation from association. ACM Trans. Inf. Syst. 21, 315–346 (2003)
10. Yang, C., Lin, K.H.Y., Chen, H.H.: Building emotion lexicon from weblog corpora. In: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions. ACL 2007, pp. 133–136. Association for Computational Linguistics, Stroudsburg (2007)
11. Brants, T., Franz, A.: Web 1t 5-gram, 10 european languages version 1. In: Linguistic Data Consortium, Philadelphia, PA, USA (2009)