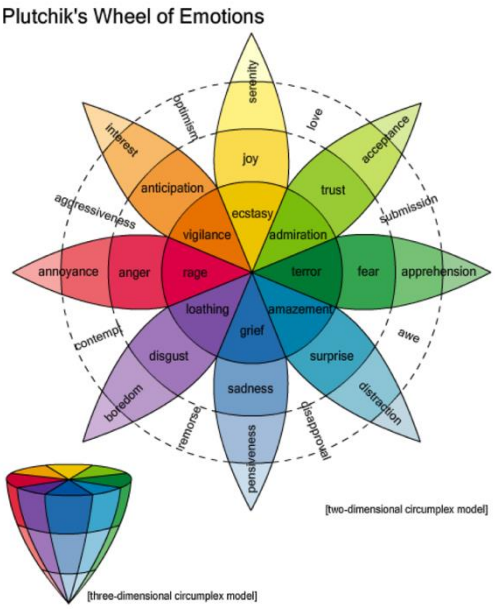
Vadym Dudarenko 444820
Ismayil Ismayilov 444459
Ivan Grakhovski 444422

"Crowdsourcing a Word-Emotion Association Lexicon" by Mohammad, Saif M., Kiritchenko, Svetlana, and Zhu, Xiaodan

The paper describes the creation of a large-scale word-emotion association lexicon using crowdsourcing.

The sentiment analysis, which involves determining the opinions and emotions of speakers towards a target entity (company, product, person, etc.) through computational assistance. The area has many applications, such as in customer relations management, and has seen significant progress in recent years with regards to determining positive or negative polarity in words, phrases, and documents. Despite advancements, there is still much research to be done in automatically analyzing emotions in text.

The authors decide to annotate words with Plutchik's eight basic emotions instead of hundreds of emotions. This choice is based on its well-established foundation in psychological, physiological, and empirical research and its balanced representation of positive and negative emotions. It also serves as a superset of other emotion categorizations.



Challenges

They used Amazon's Mechanical Turk service as a platform to obtain large-scale emotion annotations. *One of the drawbacks of this platform:* The main challenges include quality control and attracting enough interested Turkers (The people who provide responses in this platform are called Turkers). Quality control can be affected by cheaters and annotators with limited knowledge and understanding of the task. However, clear and simple instructions can improve the accuracy of annotations and agreement among annotators. The

number of Turkers willing to do the task is also dependent on the task's interest level and compensation.

Another challenge of emotion annotation lies in the complexity of emotions associated with words, which can vary based on context and word sense. Native and fluent speakers of a language are capable of identifying these emotions, but crowdsource annotation brings additional challenges such as ensuring clear instructions, avoiding long definitions, and discarding malicious annotations. To address these issues, annotators need to be familiar with the words they are annotating and compensation should be proportionate to time spent on the task. The goal is to maintain a balance between a fine-grained sense-inventory and a clear understanding of word emotions for accurate annotations.

Dealing with challenges:

To address the challenges of emotion annotation, the authors have developed a pre-question that presents the annotators with four different words and asks them to choose the one closest in meaning to the target word. The purpose of this pre-question is to guide the annotators to the intended sense of the target word and avoid misunderstandings. Three of the four options are irrelevant distractors, while the remaining option is a synonym for one of the senses of the target word. By presenting the annotators with a word choice problem instead of a definition, they can avoid long definitions and time-consuming reading while also conveying the word sense for which annotations are to be provided.

Additionally, this pre-question serves as a quality control mechanism. If an annotator is not familiar with the target word or is randomly clicking options, there is a 75% chance that they will get the answer to this question wrong, and all of their responses for the target term can be discarded. This helps ensure the accuracy of the annotations and reduces the potential for malicious or erroneous annotations.

The word choice problems were generated using the Macquarie Thesaurus, which divides vocabulary into approximately a thousand categories with a head word that represents the category meaning. The word choice question for a target term is generated by selecting four alternatives including the correct head word and three distractors randomly selected from other categories. These alternatives are presented to the annotator in a random order.

Processing:

They 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.

Once the assignments were collected, they used automatic scripts to validate the annotations:

1) A subset of the discarded assignments were officially rejected because instructions were not followed.
2) Some assignments included at least one unanswered question. These assignments were discarded and rejected.
3) 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.

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. After this post-processing, 8,883 of the initial 10,170 terms remained, each with three or more valid assignments.

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**

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. 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. 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.

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**

The authors consolidate the polarity annotations in the same manner as for emotion annotations:

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**

Conclusion:

In this paper, the authors demonstrate how crowdsourcing can be used to create a large term-emotion association lexicon, EmoLex, efficiently and economically. Emotion detection and generation have various practical applications in various fields, but there are limited resources available, especially for non-English languages. The authors used Amazon Mechanical Turk to generate the lexicon, which includes entries for over 10,000 word-sense pairs and the association with 8 basic emotions. The authors addressed various challenges in

crowdsourcing the lexicon creation and used automatically generated word choice questions to detect and reject erroneous annotations and reject annotations from unqualified or malicious annotators. The quality of the lexicon was compared to existing data and showed high quality results. The authors also identified simultaneous emotions evoked by the same term, frequent emotion associations, and terms that directly refer to emotions. All the 10,170 terms in the lexicon are annotated with their positive, negative, or neutral semantic orientation.

Opinion:

One of the main contributions of this paper is the use of crowdsourcing to create a lexicon of emotional associations. The authors note that previous lexicons of emotional associations have been created by experts in the field, which can be time-consuming and expensive. By using crowdsourcing, the authors were able to gather data from a large number of participants in a relatively short amount of time. Additionally, the authors note that the use of crowdsourcing allows for the lexicon to capture emotional associations that may be specific to certain cultures or languages.

The lexicon created in the study has been widely used in various studies and research in the field of natural language processing and computational social science. The authors note that the lexicon is able to accurately capture emotional associations for words.