Exposing a Set of Fine-Grained Emotion Categories from Tweets

Jasy Liew Suet Yan, Howard R. Turtle
School of Information Studies, Syracuse University
Syracuse, New York, USA
jliewsue@syr.edu, turtle@syr.edu

Abstract

An important starting point in analyzing emotions on Twitter is the identification of a set of suitable emotion classes representative of the range of emotions expressed on Twitter. This paper first presents a set of 48 emotion categories discovered inductively from 5,553 annotated tweets through a small-scale content analysis by trained or expert annotators. We then refine the emotion categories to a set of 28 and test how representative they are on a larger set of 10,000 tweets through crowdsourcing. We describe the two-phase methodology used to expose and refine the set of fine-grained emotion categories from tweets, compare the inter-annotator agreement between annotations generated by expert and novice annotators (crowdsourcing) and show that it is feasible to perform fine-grained emotion classification using gold standard data generated from these two phases. Our main goal is to offer a more representative and finer-grained framework of emotions expressed in microblog text, thus allowing study of emotions that are currently underexplored in sentiment analysis.

1 Introduction

The ways that individuals express themselves in tweets provide windows into their emotional worlds. Twitter, a popular microblogging site with 500 million tweets being sent a day, is particularly rich with emotion expressions. These emotion expressions can be harnessed for sentiment analysis and to build more emotion-sensitive systems. The availability of tweets has paved the way for studies of how emotions expressed on microblogs affect stock market trends [Bollen *et al*., 2011a], relate to fluctuations in social and economic indicators [Bollen *et al*., 2011b], serve as a measure for the population's level of happiness [Dodds and Danforth, 2010], provide situational awareness for both the authorities and the public in the event of disasters [Vo and Collier, 2013], and reflect clinical depression [Park *et al*., 2012].

An important starting point in analyzing emotions on Twitter is the identification of a set of suitable emotion classes. This set of emotion classes should be representative of the emotions expressed in tweets. No consensus has emerged as to how many classes are needed to represent the emotions expressed in text [Farzindar and Inkpen, 2015]. Previous studies have focused on adapting conventional emotion theories from psychology to represent emotions expressed on Twitter and has not attempted to discover the actual range of emotions expressed or how these emotions are actually characterized in tweets. The most commonly used emotion categories are adopted from the basic emotion framework, Ekman's six basic emotions (happiness, sadness, fear, anger, disgust, and surprise) [Ekman, 1971] or Plutchik's eight basic emotions comprising Ekman's six basic emotion, plus the addition of trust and anticipation [Plutchik, 1962].

Instead of borrowing a set of emotion categories from existing emotion theories in psychology, this paper aims to expose a set of categories that are representative of the emotions expressed on Twitter by analyzing the range of emotions humans can reliably detect in microblog text. Our main goal is to offer a more representative and finer-grained framework of emotions expressed in microblog text, thus allowing study of emotions that are currently underexplored in sentiment analysis. In this paper, we address the general research question of what emotions can humans detect in microblog text. We first uncover the set of emotion categories inductively from data and then further refine that set into a manageable set that both humans and machine learning systems are able to reliably detect.

2 Theoretical Background

Generally, we define emotion in text as "a subset of particularly visible and identifiable feelings" [Besnier, 1990; Kagan, 1978] that are expressed in written form through descriptions of expressive reactions (furrowed brow, smile), physiological reactions (increase in heart rate, teeth grinding), cognitions (thoughts of abandonment), behaviors (escape, attack, avoidance) as well as other socially prescribed set of responses [Averill, 1980; Cornelius, 1996]. The classification of emotion in text is largely based on two common models of emotion: 1) the dimensional model, and 2) the categorical model [Calvo and Mac Kim, 2012; Zachar and Ellis, 2012].

Proceedings of the 4th Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2016), IJCAI 2016, pages 8-14, New York City, USA, July 10, 2016.

The dimensional model organizes emotions into more general dimensions representing the underlying fundamental structure. Emotions can be identified through the composition of two or more independent dimensions [Zachar and Ellis, 2012]. Attempts to identify the dimensions have been conducted through multidimensional scaling of human similarity judgments of emotion expressions based on facial expressions [Abelson and Sermat, 1962], vocal expressions [Green and Cliff, 1975] and emotion terms [Russell, 1978]. The two common dimensions that emerged from these studies are pleasure-displeasure (valence) and degree of arousal (intensity). Similar findings are found in semantic differential studies on emotion terms with the addition of another dimension, dominance-submissiveness [Russell and Mehrabian, 1977]. Valence (also referred to as polarity) classifies emotion as either being positive, negative or neutral [Alm *et al*., 2005; Strapparava and Mihalcea, 2007]. Intensity is somewhat similar to the degree of arousal although it is generally used to measure the strength of the emotion (i.e., very weak to very strong) [Aman and Szpakowicz, 2007]. It can be operationalized as a nominal variable with labels representing varying intensities or measured on a numeric scale.

The categorical model organizes emotions into categories that are formed around prototypes. Each emotion category has a set of distinguishable properties and is assigned a label that best describes the category (e.g., happy, sad and angry). The basic emotion framework follows the categorical model, where emotion is organized and represented using a category system. Each category represents a prototypical emotion. Using a hierarchical classification approach, [Shaver *et al*., 2001] expanded the basic emotions into 25 finer categories through similarity sorting of 135 emotion words. These finer categories are more representative of the emotions that can be expressed using English words.

The dimensional model offers a more coarse-grained representation of emotion while the categorical model can be used to represent emotion at a finer-grained level. In addition, the categorical model uses emotion labels that are more intuitive, thus making recognition of the emotion easier for humans. Therefore, we adopted the categorical model in line with our goal to develop a fine-grained emotion taxonomy for microblog text.

2 Methodology

We used content analysis to identify a stable set of emotion categories that is representative of the range of emotions expressed in tweets. The small-scale content analysis was first conducted (Phase 1) by training a group of annotators to annotate a sample of 5,553 tweets. Three tasks were completed to uncover this set of emotion categories: 1) inductive coding, 2) card sorting, and 3) emotion word rating. In Phase 2, we tested the representativeness of the emotion categories derived from Phase 1 using large-scale content analysis. Annotations were collected through crowdsourcing using Amazon Mechanical Turk (AMT).

2.1 Data Collection

Data consisted of tweets (i.e., microblog posts) retrieved from Twitter. Four different sampling strategies were used to retrieve the tweets to be included in the corpus: random sampling (RANDOM), sampling by topic (TOPIC), and two variations of sampling by user type (SEN-USER and AVG-USER). For the RANDOM sample, nine stopwords (the, be, to, of, and, a, in, that, have) reported to be words most frequently used on Twitter were used to retrieve tweets. Topic sampling was done by retrieving tweets that contain selected topical hashtags or keywords. Sampling by user type retrieved tweets using selected user names (@usernames). One user sample contained tweets retrieved from US Senators (SEN-USER). Tweets from the second user sample were retrieved using randomly selected user names (AVG-USER). Tweets were either retrieved using the Twitter API or acquired from two publicly available data sets: 1) the SemEval 2014 tweet data set [Nakov *et al*., 2013; Rosenthal *et al*., 2014], and 2) the 2012 US presidential elections data set [Mohammad *et al*., 2014]. The data set containing 15,553 tweets received roughly equal contribution from each of the four sampling strategies.

2.2 Phase 1: Small-scale Content Analysis

2.2.1 Task 1: Inductive Coding

We adapted grounded theory [Glaser and Strauss, 1967] to expose a set of fine-grained emotion categories from tweets. This method used inductive coding to derive the classification scheme through observation of content [Potter and Levine-Donnerstein, 1999]. Annotators engaged in three coding activities central to this method: open coding, axial coding, and selective coding [Corbin and Strauss, 2008]. In open coding, annotators read the content of each tweet to capture all possible meanings, and took a first pass at assigning concepts to describe the interpretation of the data. No restriction was posed on analysis in this phase, and minimal instructions were provided to avoid predisposing annotators. Axial coding then involved the process of drawing the relationships between concepts and categories. Based on their knowledge of emotion, annotators started with a set of self-defined emotion tags. They then met in groups with the primary researcher to start drawing relationships between different emotion tags suggested by individuals in the group. Emotion tags were examined, accepted, modified, and discarded. Discrete emotion categories started to form in this phase, and were systematically applied to more data. Annotators switched back and forth between axial coding and open coding until a stable set of categories was identified. Finally, selective coding represented an integration phase where the identified discrete categories were further developed, defined and refined under a unifying theme of emotion. Annotators then continued to validate the classification scheme by applying and refining it on more data until no new category emerged.

Graduate students who were interested in undertaking the task as part of a class project (e.g., Natural Language Processing course) or to gain research experience in content

analysis (e.g., independent study) were recruited as annotators. Annotators were not expected to possess special skills except for the required abilities to read and interpret English text. A total of eighteen annotators worked on the annotation task over a period of ten months. To derive an emotion framework based on collective knowledge, each tweet was annotated by at least three annotators. Thus, annotators were divided into groups of at least three. Each group was assigned to work on one of the four samples.

All the annotators went through the same training procedures to reduce as much as possible the variation among different individuals. Each annotator first attended a one hour training session to discuss the concept of emotion with the researcher and to receive instructions on how to perform annotations of the tweets. Annotators were not given any emotion categories and were asked to suggest the best-fitting emotion tags or labels to describe the emotion expressed in each tweet (Example 1). For tweets containing multiple emotions, annotators were asked to first identify the primary emotion expressed in the tweet, and then also include the other emotions observed (Example 2).

Example 1: Alaska is so proud of our Spartans! The 4-25 executed every mission in Afghanistan with honor & now, they're home http://t.co/r8pLpnud **[Pride]**

Example 2: Saw Argo yesterday, a movie about the 1979 Iranian Revolution. Chilling, sobering, and inspirational at the same time. **[Inspiration, Fear]**

Annotation was done in an iterative fashion. In the first iteration, also referred to as the training round, all annotators annotated the same sample of 300 tweets from SEN-USER. Upon completing the training round, annotators were assigned to annotate at least 1,000 tweets from one of the four samples (RANDOM, TOPIC, AVG-USER or SEN-USER) in subsequent iterations. Every week, annotators worked independently on annotating a subset of 150 – 200 tweets but met with the researcher in groups to discuss disagreements, and 100% agreement for emotion tag was achieved after discussion. In these weekly meetings, the researcher also facilitated the discussions among annotators working on the same sample to merge, remove, and refine suggested emotion tags. Output of Task 1 included 4,010 annotated tweets in the gold standard corpus and 246 emotion tags.

2.2.2 Task 2: Card Sorting

Some of the 246 emotion tags were simply morphological variations and many were semantically similar. Task 2 served as an intermediate step to refine the emotion tags emerging from data into a more manageable set of higher level emotion categories. Annotators were asked to perform a card sorting exercise in different teams to group emotion tags that are variants of the same root word or semantically similar into the same category. Annotators were divided into 5 teams, and each team received a pack of 1' x 5' cards containing only the emotion tags used by the all members in their respective teams.

Each team consisted of 2 - 3 members who worked on the same sample. Teams were instructed to follow the four-step procedures described below:

a) Group all the emotion tags into categories. Members were allowed to create a "Not Emotion" category if needed.
b) Create a name for the emotion category. Collectively pick the most descriptive emotion tag or suggest a new name to represent each category.
c) Group all the emotion categories based on valence: positive, negative and neutral.
d) Match emotion categories generated from other team's card sorting activity to the emotion categories proposed by your team.

Team	Sample	Number of Emotion Categories			
		Positive	Negative	Neutral	Total
G1	SEN-USER	8	13	2	23
G2	TOPIC	16	14	5	35
G3	TOPIC	16	18	8	42
G4	AVG-USER	14	18	15	47
G5	RANDOM	14	16	9	39

Table 1: Number of categories proposed by each card sorting team

Members in the same team were allowed to discuss their decisions with each other during the card sorting exercise with minimal intervention from the researcher. The session concluded when all members completed the four-step procedure and reached a consensus on final groupings of the emotion tags. No limit was placed on the number of categories or the number of emotion tags within each category so the number of categories proposed varied across the five teams as shown in Table 1. Some teams decided to put the emotion tags into fewer higher-level categories, while others who chose to capture more subtle emotions generated more emotion categories. Finally, the researcher merged, divided, and verified the final emotion categories to be included in the classification scheme.

Once the final 48 emotion categories shown in Table 2 were identified (see Emotion-Category-48 column), the original emotion tag labels generated from the open coding exercise were systematically replaced by the appropriate emotion category labels. Annotators then incrementally annotated more tweets (150 - 200 tweets per round) to ensure that a point of saturation was reached. No new emotion category emerged from data in this coding phase. Another 1,543 annotated tweets with gold labels were added to the corpus.

2.2.3 Task 3: Emotion Word Rating

We found it methodologically challenging and time consuming to provide rigorous training to a large number of annotators in order to grow the size of the corpus with 48 emotion categories. A word rating study was conducted as a systematic method to merge and distill the number of categories into a more manageable set. The motivation behind the word rating study came from prior studies showing that emotion words with greater similarity tend to be in close proximity to one another on a two-dimensional

pleasure and degree of arousal space [Russell, 1980]. In order to plot our emotion categories in this two-dimensional space, we collected the pleasure and arousal ratings for each emotion category. A set of 50 emotion words were selected for the emotion rating task. We included the 48 emotion category names and added 2 emotion words that were deemed to be more appropriate category names than the ones determined by the annotators in Task 2. These two emotion words were *"longing"* for the category *"yearning"* and *"torn"* for *"ambivalence"*.

To obtain a complete set of pleasure and arousal ratings for our set of 50 emotion words, we conducted an emotion word rating study on AMT. We adapted the instrument that was used in [Bradley & Lang, 1999] to collect the ratings. We implemented the study using exactly the same 9-point scale for the pleasure and arousal ratings. The validity of the scales are described in [Bradley & Lang, 1994]. The same set of instructions was reused but modified to fit the crowdsourcing context.

Human raters were recruited from the pool of workers available on AMT. The rating instrument was offered to the workers via a Human Intelligence Task (HIT), and workers received payment of US$ 0.20 upon completion and approval of the HIT. HITs were restricted to workers in the US to increase the likelihood that ratings came from native English speakers. Each respondent first read the instructions on how to use the pleasure and arousal scales. Respondents were then instructed to make a pleasure rating and an arousal rating for each of the 50 emotion words.

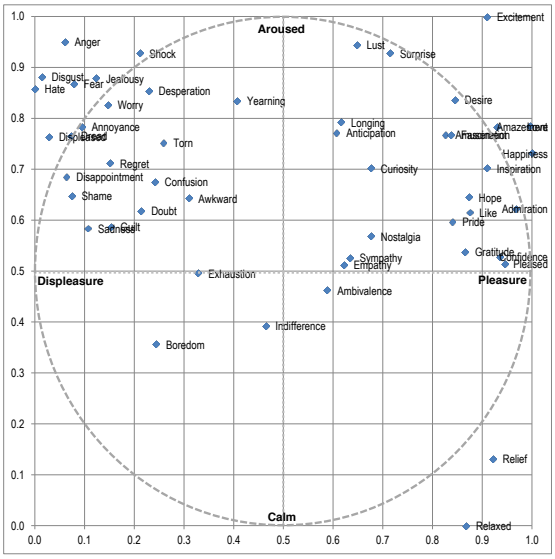

Figure 1: Two-dimensional pleasure and arousal plot for 50 emotion words based on AMT ratings (x-axis represents pleasure, y-axis represents arousal)

After removing incomplete and rejected responses, mean rating and standard deviation were computed from 76 usable responses.

Figure 1 shows the plot for all 50 emotion words based on AMT ratings normalized using feature scaling. Emotion categories that are semantically-related and relatively close in proximity to one another on the plot are merged. The merge process involved some subjective decision and reduced the number of emotion categories from 48 to the final set of 28 shown in Table 2 (see Emotion-Category-28 column). Category name "ambivalence" was substituted by its more descriptive member term, "torn" and "yearning" was substituted by "longing". Also, two emotion categories from the original 48, "desire" and "lust" were dropped altogether from the final set of 28 because it is not clear that they should be considered separate emotional states [Ortony and Turner, 1990]. Based on their conceptualization in our annotation scheme, they were considered to be more general feelings of wanting rather than distinct emotional states. Finally, the 48 emotion category labels in the corpus were systematically replaced by the corresponding 28 emotion category labels.

The set of 28 categories is derived from the corpus and is a "good" representation of the set of emotions expressed therein. It is substantially more refined than the traditional 5 to 8 category set yet is small enough that human annotators are comfortable with the distinctions.

2.3 Phase 2: Large-scale Content Analysis

Manual annotations for an additional 10,000 tweets were obtained using AMT in Phase 2. For emotion tag, workers were given a set of 28 emotion categories to choose from plus an "other" option with a text box so they could suggest a new emotion tag where none of the listed emotion category was applicable. The order in which the emotion categories were presented to the workers was randomized across the four samples in order to control for order effect. If a tweet was flagged as containing multiple emotions, annotators were asked to provide all relevant emotion tags.

Recruitment of workers was done through Human Intelligence Tasks (HITs) on the online AMT platform. AMT workers must fulfill at least the basic requirement of being able to read and understand English text. We set the HIT approval rate for all requesters' HITs to greater than or equal to 95% and the number of HITs approved to greater than or equal to 1000 to increase the probability of recruiting first-rate workers.

In the design of the HIT, workers were provided clear and simple instructions describing the task, the annotation site link, as well as a batch id required to retrieve a subset of 30 tweets to work on. Of the 30 tweets in each HIT, 25 were new tweets and 5 were gold standard tweets intended to be used for quality control. Each HIT was assigned to three different annotators. Each HIT bundled a different subset of 30 tweets so a worker could attempt more than one HIT. Workers were paid US$ 0.50 for every completed and approved HIT containing 30 tweets.

Emotion-Category-28	Emotion-Category-48	Emotion-Category-28	Emotion-Category-48
Admiration	Admiration	Hate	Hate, Disgust
Amusement	Amusement	Hope	Hope
Anger	Anger, Annoyance, Displeased, Disappointment	Indifference	Indifference
Boredom	Boredom	Inspiration	Inspiration
Confidence	Confidence	Jealousy	Jealousy
Curiosity	Curiosity	Longing	*Longing, Nostalgia
Desperation	Desperation	Love	Love, Like
Doubt	Doubt, Confusion, *Torn	Pride	Pride
Excitement	Excitement, Anticipation	Regret	Regret, Guilt
Exhaustion	Exhaustion	Relaxed	Relaxed, Relief
Fascination	Fascination, Amazement	Sadness	Sadness
Fear	Fear, Dread, Worry	Shame	Shame, Awkward
Gratitude	Gratitude	Surprise	Surprise, Shock
Happiness	Happiness, Pleased	Sympathy	Sympathy, Empathy

Table 2: Mapping between the final set of 28 emotion categories to the original set of 48 (category names preceded by * were modified)

Phase	Phase 1				Phase 2				Phase 1 + Phase 2			
Category	**n**	**P**	**R**	**F1**	**n**	**P**	**R**	**F2**	**n**	**P**	**R**	**F1**
Admiration	158	0.417	0.190	0.261	245	0.328	0.155	0.211	403	0.370	0.201	0.260
Amusement	237	0.744	0.515	0.608	423	0.888	0.617	0.728	660	0.869	0.645	0.741
Anger	444	0.288	0.203	0.238	757	0.495	0.346	0.407	1201	0.478	0.321	0.384
Boredom	12	0.400	0.167	0.235	36	0.714	0.417	0.526	48	0.818	0.375	0.514
Confidence	19	0.000	0.000	0.000	91	0.286	0.088	0.134	110	0.303	0.091	0.140
Curiosity	30	0.586	0.567	0.576	63	0.591	0.413	0.486	93	0.638	0.548	0.590
Desperation	8	1.000	0.125	0.222	50	0.417	0.100	0.161	58	0.500	0.069	0.121
Doubt	50	0.125	0.020	0.034	108	0.256	0.102	0.146	158	0.269	0.089	0.133
Excitement	265	0.457	0.377	0.413	421	0.675	0.463	0.549	686	0.655	0.474	0.550
Exhaustion	10	0.000	0.000	0.000	39	0.706	0.308	0.429	49	0.611	0.224	0.328
Fascination	54	0.417	0.185	0.256	150	0.587	0.360	0.446	204	0.553	0.309	0.396
Fear	77	0.240	0.078	0.118	162	0.556	0.216	0.311	239	0.491	0.230	0.313
Gratitude	221	0.943	0.905	0.924	300	0.913	0.877	0.895	521	0.928	0.914	0.921
Happiness	778	0.589	0.500	0.541	1009	0.596	0.477	0.530	1787	0.622	0.506	0.558
Hate	63	0.778	0.444	0.566	129	0.812	0.535	0.645	192	0.788	0.542	0.642
Hope	187	0.660	0.508	0.574	335	0.781	0.564	0.655	522	0.781	0.580	0.666
Indifference	28	0.500	0.071	0.125	40	0.308	0.100	0.151	68	0.235	0.059	0.094
Inspiration	21	0.923	0.571	0.706	54	0.731	0.352	0.475	75	0.816	0.413	0.549
Jealousy	5	1.000	0.400	0.571	29	0.846	0.379	0.524	34	0.765	0.382	0.510
Longing	41	0.545	0.146	0.231	80	0.487	0.238	0.319	121	0.529	0.306	0.387
Love	234	0.608	0.444	0.514	447	0.645	0.538	0.587	681	0.659	0.519	0.581
Pride	85	0.817	0.682	0.744	128	0.907	0.688	0.782	213	0.862	0.676	0.758
Regret	49	0.500	0.102	0.169	104	0.571	0.308	0.400	153	0.514	0.242	0.329
Relaxed	26	0.200	0.038	0.065	51	0.550	0.216	0.310	77	0.737	0.182	0.292
Sadness	158	0.609	0.335	0.433	363	0.612	0.444	0.514	521	0.650	0.461	0.539
Shame	26	0.600	0.231	0.333	64	0.545	0.281	0.371	90	0.622	0.311	0.415
Surprise	93	0.342	0.140	0.198	173	0.627	0.301	0.406	266	0.556	0.278	0.371
Sympathy	35	0.813	0.371	0.510	66	0.625	0.379	0.472	101	0.705	0.426	0.531
None	2637				5047				7684			
Macro-avg		**0.539**	**0.297**	**0.363**		**0.609**	**0.366**	**0.449**		**0.619**	**0.370**	**0.450**
Micro-avg		**0.580**	**0.400**	**0.474**		**0.647**	**0.440**	**0.524**		**0.656**	**0.455**	**0.537**

Table 3: Precision, recall and F1 of SMO classifiers across P1, P2 and P1+P2

3 Inter-annotator Agreement

Table 4 presents the inter-annotator agreement for 28 emotion categories based on tweets with three annotations. Overall α across Phase 1 and Phase 2 is 0.43. Mean α for 28 emotion categories in Phase 1 is 0.50. Emotion annotation especially at a fine-grained level is a subjective and difficult task. It is possible to generate reliable data when annotators are given sufficient training. With limited training, α in Phase 2 decreases almost by half to 0.28.

Sample	EmoCat-28		
	%	κ	α
Phase 1	66	0.50	0.50
Phase 2	51	0.28	0.28
Phase 1+Phase 2	61	0.43	0.43

Table 4: Inter-annotator agreement (percent agreement, Fleiss' κ and Krippendorff's α)

For Phase 1, all disagreements were first resolved through discussion with expert annotators. Essentially, expert annotators achieved 100% agreement in Phase 1. In Phase 2, about one third of the tweets had full agreement for emotion tag among all annotators (32%). To avoid throwing away any data, the researcher manually reviewed all annotations and resolved the disagreements. Such effort was deemed necessary to reduce as much noise as possible in the corpus, and to ensure that the classification schemes were applied consistently across the two phases of data collection. Similar to the Phase 1, each tweet in Phase 2 was assigned final labels for emotion category.

4 Emotion Distribution

Slightly over half (51%) of the tweets contain emotion. Table 3 shows imbalance in the frequencies of the emotion categories. Of the 28 emotion categories, the full corpus (Phase 1 and Phase 2) contains the highest instances of *happiness* (12%) and the lowest instances of jealousy (0.2%). Only 9 categories have less than 100 instances. The frequency distribution of the emotion categories in Phase 1, Phase 2 and Phase 1 + Phase 2 are roughly similar.

The corpus contains a significant portion of tweets tagged with a single emotion category (92%) and only 8% of tweets tagged with more than one emotion category. Although tweets containing multiple emotions represent only 8% of the corpus, including such tweets in the corpus leads to over 40% overall increase in the number of positive examples (i.e., instances of an emotion category).

5 Comparing Machine Learning Results from Phase 1 and Phase 2

Since a tweet might be assigned multiple emotion categories, we frame the problem as a multi-label classification task. A separate binary classifier was built for each emotion category to detect if an emotion category were present or absent in a tweet (emotion X or not emotion X).

We conducted a wide range of classification experiments to better understand the impact of classifier and feature set selection on classification accuracy [Liew, 2016]. We present here results for a single representative selection: Sequential Minimal Optimization (SMO), an SVM variant [Platt, 1998] trained with features that include unigrams occurring three or more times in the corpus that are stemmed and lowercased. Classifiers were evaluated using ten-fold cross validation.

The precision, recall and F1 for SMO across Phase 1, Phase 2 and Phase 1 + Phase 2 are shown in Table 3. A general upward trend in precision (P), recall (R) and F1 are observed across the three data sets. There are two key takeaways from our preliminary experiments. First, using the combined data from P1 and P2 generally yields higher performance than using P1 or P2 data alone. For a majority of the emotion categories, the classifiers used for emotion classification achieved similar performance using gold standard data generated Phase 1 and Phase 2 respectively. Second, classifiers provided with more training examples usually produce higher overall performance as evidenced by higher F1 when larger data sets are used. The results for individual emotion categories shows that more data does not always leads to higher performance. The classifiers may behave differently depending on the linguistic characteristics of the category. More experiments will be conducted in future work to identify the salient linguistic features for each emotion category.

6 Conclusion

We describe a two-phase methodology to uncover a set of 28 emotion categories representative of the emotions expressed in tweets. There are two main contributions: 1) the introduction an emotion taxonomy catered for emotion expressed in text and 2) the development of a gold standard corpus that can be used to train and evaluate more fine-grained emotion classifiers.

The set of 28 emotions is derived using an integrative view of emotion and grounded on linguistic expressions of emotion in text. In Phase 1, inductive coding was first used to expose a set of emotion categories from 5,553 tweets. The categories were then further merged and refined using card sorting and emotion word rating. In Phase 2, we then tested the representativeness of the emotion categories on a larger data set of 10,000 tweets using crowdsourcing. No new emotion categories emerged from Phase 2, indicating that the 28 emotion categories are sufficient to capture the richness of emotional experiences expressed in tweets. However, the classifiers perform poorly on some categories such as *confidence*, *desperation*, *doubt* and *indifference*. We intend to perform a closer examination of the low performing categories to determine if they should be removed.

Acknowledgments

We thank the annotators who volunteered in performing the annotation task. We are grateful to Dr. Elizabeth D. Liddy for her insights in the study.

References

[Abelson and Sermat, 1962] Robert Abelson, and Vello Sermat. Multidimensional Scaling of Facial Expressions. *Journal of Experimental Psychology*, 63(6):546–554, 1962.

[Alm et al., 2005] Cecilia Alm, Dan Roth, and Richard Sproat. Emotions from Text: Machine Learning for Text-Based Emotion Prediction. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 579–586, Stroudsburg, PA, USA, 2005.

[Aman and Szpakowicz, 2007] Saima Aman, and Stan Szpakowicz. Identifying Expressions of Emotion in Text. In *Text, Speech and Dialogue*, pages 196–205, 2007.

[Averill, 1980] James R. Averill. A Constructivist View of Emotion. *Emotion: Theory, Research, and Experience*, 1:305–339, Academic Press, New York, 1980.

[Besnier, 1990] Niko Besnier. Language and Affect. *Annual Review of Anthropology*, 19:419–451, 1990.

[Bollen *et al*., 2011a] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter Mood Predicts the Stock Market. *Journal of Computational Science*, 2(1):1–8, 2011.

[Bollen *et al*., 2011b] Johan Bollen, Alberto Pepe, and Huina Mao. Modeling Public Mood and Emotion: Twitter Sentiment and Socio-Economic Phenomena. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, pages 450–53, 2011.

[Bradley *et al*., 1994] Margaret M. Bradley, and Peter J. Lang. Measuring Emotion: The Self-Assessment Manikin and the Semantic Differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1): 49–59, 1994.

[Bradley and Lang, 1999] Margaret M. Bradley, and Peter J. Lang. Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings. University of Florida: Technical Report C-1, The Center for Research in Psychophysiology, 1999.

[Calvo and Mac Kim, 2012] Rafael A. Calvo, and Sunghwan Mac Kim. Emotions in Text: Dimensional and Categorical Models. *Computational Intelligence*, 29(3):527–43, 2012.

[Corbin and Strauss, 2008] Juliet Corbin, and Anselm Strauss. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. Sage, 2008.

[Cornelius, 1996] Randolph R. Cornelius. *The Science of Emotion: Research and Tradition in the Psychology of Emotions*. Upper Saddle River, Prentice Hall, New Jersey, 1996.

[Dodds and Danforth, 2010] Peter S. Dodds, and Christopher M. Danforth. Measuring the Happiness of Large-Scale Written Expression: Songs, Blogs, and Presidents. *Journal of Happiness Studies*, 11(4):441–56, 2010.

[Ekman, 1971] Paul Ekman. Universals and Cultural Differences in Facial Expressions of Emotion. *Nebraska Symposium on Motivation*, 19:207–83, 1971.

[Farzindar and Inkpen, 2015] Atefeh Farzindar, and Diana Inkpen. Natural Language Processing for Social Media. *Synthesis Lectures on Human Language Technologies*, 8(2):1–166, 2015.

[Glaser and Strauss, 1967] Barney G. Glaser, and Anselm L. Strauss. *The Discovery of Grounded Theory: Strategies for Qualitative Research*, Aldine Publishing, Chicago 1967.

[Green and Cliff, 1975] Rex S. Green, and Norman Cliff. Multidimensional Comparisons of Structures of Vocally and Facially Expressed Emotion. *Perception & Psychophysics*, 17(5):429–438, 1975.

[Kagan, 1978] Jerome Kagan. On Emotion and Its Development: A Working Paper. In *The Development of Affect*, pages 11–41, Genesis of Behavior 1, 1978.

[Liew, 2016] Jasy Liew Suet Yan. *Fine-Grained Emotion Detection in Microblog Text*. Syracuse, NY, USA: Syracuse University, 2016.

[Mohammad *et al*., 2014] Saif Mohammad, Xiaodan Zhu, and Joel Martin. Semantic Role Labeling of Emotions in Tweets." In *Proceedings of the ACL 2014 Workshop on Computational Approaches to Subjectivity, Sentiment, and Social Media*, pages 32–41, Baltimore, MD, USA, 2014.

[Nakov *et al*., 2013] Preslav Nakov, Zornitsa Kozareva, Alan Ritter, Sara Rosenthal, Veselin Stoyanov, and Theresa Wilson. SemEval-2013 Task 2: Sentiment Analysis in Twitter." In *Proceedings of the 7th International Workshop on Semantic Evaluation* , 2:312–320, 2013.

[Ortony and Turner, 1990] Andrew Ortony, and Terence J. Turner. What's Basic about Basic Emotions? *Psychological Review*, 97(3):315–331, 1990.

[Park *et al*., 2012] Minsu Park, Chiyoung Cha, and Meeyoung Cha. Depressive Moods of Users Portrayed in Twitter. In *Proceedings of the ACM SIGKDD Workshop on Healthcare Informatics*, pages 1–8, 2012.

[Platt, 1998] John C. Platt. Fast Training of Support Vector Machines Using Sequential Minimal Optimization. In *Advances in Kernel Methods*, pages 41–65, MIT Press, 1998.

[Plutchik, 1962] Robert Plutchik. *The Emotions: Facts, Theories, and a New Model*. Studies in Psychology, Random House, New York, 1962.

[Potter and Levine-Donnerstein, 1999] W. James Potter, and Deborah Levine-Donnerstein. Rethinking Validity and Reliability in Content Analysis. *Journal of Applied Communication Research*, 27(3):258–284, 1999.

[Rosenthal *et al*., 2014] Sara Rosenthal, Preslav Nakov, Alan Ritter, and Veselin Stoyanov. Semeval-2014 Task 9: Sentiment Analysis in Twitter." In *Proceedings of the 8th International Workshop on Semantic Evaluation*, pages 73–80. Dublin, Ireland, 2014.

[Russell, 1978] James A. Russell. Evidence of Convergent Validity on the Dimensions of Affect. *Journal of Personality and Social Psychology*, 36(10):1152–1168, 1978.

[Russell, 1980] James A. Russell. A Circumplex Model of Affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, 1980.

[Russell and Mehrabian, 1977] James A. Russell, and A. Mehrabian. Evidence for a Three-Factor Theory of Emotions. *Journal of Research in Personality*, 11(3):273–294, 1977.

[Shaver et al., 2001] Phillip Shaver, Judith Schwartz, Donald Kirson, and Cary O'Connor. Emotion Knowledge: Further Exploration of a Prototype Approach. In *Emotions in Social Psychology*, pages 26–56. Psychology Press, 2001.

[Strapparava and Mihalcea, 2007] Carlo Strapparava, and Rada Mihalcea. Semeval-2007 Task 14: Affective Text. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 70–74. Prague, 2007.

[Vo and Collier, 2013] Bao-Khanh H. Vo, and Nigel Collier. Twitter Emotion Analysis in Earthquake Situations. *International Journal of Computational Linguistics and Applications*, 4(1):159–173, 2013.

[Zachar and Ellis, 2012] Peter Zachar, and Ralph D. Ellis. *Categorical versus Dimensional Models of Affect: A Seminar on the Theories of Panksepp and Russell*. Vol. 7. John Benjamins Publishing Company, 2012.