EmoTweet-28: A Fine-Grained Emotion Corpus for Sentiment Analysis

Jasy Liew Suet Yan, Howard R. Turtle, Elizabeth D. Liddy
School of Information Studies, Syracuse University
Syracuse, New York, USA
E-mail: jliewsue@syr.edu, turtle@syr.edu, liddy@syr.edu

Abstract

This paper describes EmoTweet-28, a carefully curated corpus of 15,553 tweets annotated with 28 emotion categories for the purpose of training and evaluating machine learning models for emotion classification. EmoTweet-28 is, to date, the largest tweet corpus annotated with fine-grained emotion categories. The corpus contains annotations for four facets of emotion: valence, arousal, emotion category and emotion cues. We first used small-scale content analysis to inductively identify a set of emotion categories that characterize the emotions expressed in microblog text. We then expanded the size of the corpus using crowdsourcing. The corpus encompasses a variety of examples including explicit and implicit expressions of emotions as well as tweets containing multiple emotions. EmoTweet-28 represents an important resource to advance the development and evaluation of more emotion-sensitive systems.

Keywords: emotion corpus, sentiment analysis, microblog text

1. Introduction

Twitter, a popular microblogging site, provides a window into the emotional worlds of significant user populations. Twitter data can be leveraged to study behavior on social media in a non-intrusive manner. Given the large traffic volume (500 million tweets per day), automatic techniques to detect emotions are needed to augment our ability to analyze and understand emotion content. Automatic emotion detection in text at a fine-grained level is potentially useful for applications such as personality detection, public and behavioral health monitoring as well as consumer and market analysis.

An important starting point for building natural language processing (NLP) systems to detect emotions in text is the construction of a ground truth corpus annotated to characterize emotion content. Existing emotion corpora are annotated with relatively coarse-grained categories (e.g., 6 to 8 basic emotion categories) (Mohammad, Zhu, & Martin, 2014; Roberts et al., 2012). Many current corpora are automatically annotated based on emotion hashtags rather than human judgment (Mohammad, 2012; Pak & Paroubek, 2010; Wang et al., 2012).

With the basic emotions (*happiness*, *sadness*, *fear*, *anger*, *disgust* and *surprise*) accepted as the state-of-the-art, existing emotion corpora and other language resources that serve as the basis for building and evaluating mechanisms to detect emotion in tweets are only annotated with those basic categories. As a result, automatic emotion detectors developed using these resources are only able to give us a limited picture of human emotion expressions. Other emotions apart from the basic set, as well as variations within each basic emotion are "virgin territories" that have not yet been explored by researchers in this area. Efforts to increase the utility of automatic emotion detectors have to start with extending language resources to cover other emotion categories.

We describe our efforts to construct a gold standard corpus of 15,553 tweets annotated with 28 emotion categories for the purpose of training and evaluating machine learning models for emotion classification. EmoTweet-28 is the largest carefully curated tweet corpus annotated with fine-grained emotion categories. The contributions of the paper are:

- We present a two-phase methodology describing how we first (Phase 1) used content analysis to inductively identify a set of emotion categories that characterize the emotions expressed in microblog text. We then (Phase 2) expanded the size of the corpus by conducting large-scale content analysis via Amazon Mechanical Turk (AMT).
- We describe the characteristics of the corpus.
- We conduct basic supervised machine learning experiments to evaluate the performance of automatic classification with such fine-grained emotion categories.

2. Related Work

Using Twitter, researchers have explored different strategies to harness large volumes of data for automatic emotion classification. Using a method known as "distant supervision", Pak & Paroubek (2010) applied a method similar to Read (2005) to extract tweets containing happy emoticons to represent positive sentiment, and sad emoticons to represent negative sentiment. This method allows for fast collection of a large self-labeled corpus without the need for manual annotation, but is limited in that it enables the emotion classifier to detect only happiness and sadness.

Mohammad (2012) and Wang et al. (2012) applied an improved method to create a large corpus of self-labeled tweets for emotion classification. Twitter allows the use of hashtags (words that begin with the # sign) as topic indicators. Extracting tweets that contain a predefined list of emotion words appearing in the form of hashtags was used to collect data for these studies. Mohammad (2012) only extracted tweets with emotion hashtags

corresponding to Ekman's six basic emotions (#anger, #disgust, #fear, #joy, #sadness, and #surprise) while Wang et al. (2012) expanded the predefined hashtag list to include emotion words associated with an emotion category, as well as the lexical variants of these emotion words. This approach allows researchers to take advantage of the huge amount of data available on Twitter to train machine learning models. Statistical methods can be used to identify words that frequently co-occur with the emotion hashtags but little is known about the actual linguistic properties that are associated with these emotion categories. Also, this data collection method is biased towards users who choose to express their emotions explicitly using hashtags.

To address some of the criticism associated with the distant supervision method, Purver & Battersby (2012) investigated if the classifiers trained using automatically annotated data (i.e., noisy labels) are recognizing the actual underlying emotion class by comparing models trained with different hashtag and emoticon labels or markers. A corpus of tweets was first collected using a predefined list of emotion markers, which only included emoticons and emotion word hashtags that were considered to be conventional markers for six emotion classes (i.e., *happy*, *sad*, *anger*, *fear*, *surprise* and *disgust*). The classifiers demonstrated reasonable performance when trained and tested on tweets containing the same label convention or emotion marker. Classifier performance was less reliable across label conventions (i.e., training on one emotion marker and testing on the others) and against a set of manually annotated examples. The distant supervision method was suitable for only some emotions like *happiness*, *sadness* and *anger* but did poorly in distinguishing other emotions.

Manual annotation has also been used in the development of emotion tweet corpora. Roberts et al. (2012) annotated a tweet corpus sampled by topics expected to evoke emotions with seven emotion categories (*anger*, *disgust*, *fear*, *joy*, *love*, *sadness*, and *surprise*). With the exception of *love*, the other six emotion categories are adopted from Ekman's six basic emotions. While the data may not be representative of Twitter as a whole, manual annotation allows for tweets that are not explicitly tagged with an emotion word (#emotion) to be included in the training data. This way, the machine learning model can also learn from tweets containing explicit expressions that do not include one of the tagged emotion word and implicit expressions of emotion.

The studies reviewed so far employed only a small set of emotion categories. These emotion categories are too limited to capture the richness of emotions expressed in tweets. To address this limitation, we developed a set of emotion categories that accurately describes the emotions that are expressed in tweets.

3. Methodology

3.1 Data Collection

Four different sampling strategies were used to retrieve tweets to be included in the corpus: random sampling (RANDOM), sampling by topic using selected keywords or topical hashtags (TOPIC), and two variations of sampling by user type using @username (SEN-USER and AVG-USER). Tweets were either retrieved from the Twitter API or acquired from publicly available data sets. Tweets were pre-processed to remove spam, duplicates, repeated retweets, and non-English tweets. A total of 15,553 tweets were included in the corpus, where 5,553 tweets were annotated in Phase 1 and 10,000 tweets were annotated in Phase 2. The distribution of tweets for each sample is shown in Table 1.

Sample	**Sample Size**		
	P1	**P2**	**Total**
RANDOM	1450	2500	3950
TOPIC	1310	2500	3810
SEN-USER	1493	2500	3993
AVG-USER	1300	2500	3800
Total	**5553**	**10000**	**15553**

Table 1: Distribution of tweets for 4 samples

3.1.1. Random Sampling [RANDOM]

The first sampling strategy was intended to collect a random sample of tweets that is representative of the overall population on Twitter. The sample produced using this strategy might not be as rich with emotional content as the other samples. Since the Twitter API required query terms to retrieve tweets, 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 for the random sample. An initial sample of 48,577 tweets was collected. Then, a random number generator was used to select tweets to be included in the corpus. The tweets were created between May – July 2014.

3.1.2. Sampling by Topic [TOPIC]

The second sampling strategy was based on topics or events. Tweets were sampled based on hashtags of events expected to contain emotional content. A wide range of topics were included to reduce the effect of emotional biases associated with certain topics (e.g., disaster-related topics are more likely to contain more negative emotions). The tweets for this sample were sampled from three sources: 1) the SemEval 2014 tweet data set (Nakov et al., 2013; Rosenthal et al., 2014), 2) the 2012 US presidential elections data set (Mohammad et al., 2014), and 3) tweets retrieved using the Twitter API from February – December 2014 using query terms shown in Table 2.

3.1.3. Sampling by User [USER]

The final two sampling strategies were based on usernames. These two sampling strategies were aimed at striking a balance between including users who were representative of "average" Twitter users and active users who generated a relatively large number of tweets for analysis. One sample was collected from "average" Twitter users [AVG-USER] and another sample was collected from US political leaders (active Twitter users)

[SEN-USER]. While sampling tweets from selected individuals limits the generality of findings, it allows exploration of the emotion variation and distribution in individual streams of tweets and examination of any differences when compared to the TOPIC and RANDOM samples.

Data Source: Topic Description	Total	Sample Size	
		P1	P2
SemEval 2014: Topics related to famous characters (e.g., Gadafi, Steve Jobs), products (e.g. Kindle, Android phone), and events (e.g., Japan earthquake, NHL playoffs)	9520	910	400
2012 US presidential elections: #4moreyears, #Barack #election2012, #ObamaBiden2012, #mitt2012, #dems2012, #gop2012, etc.	168975	200	1100
Twitter API: #Sochi2014, #Oscar2014, #PrayForMH370, #MH17, #ValentinesDay, #anniversary, #graduation, #americanairlines, etc.	6621	200	1000

Table 2: Description of topics included in TOPIC

[SEN-USER]: We first collected the @usernames of 89 US Senators, who were active users with a large number of followers from www.tweetcongress.org. The tweet streams were then collected from the Twitter API using the @usernames as the query terms. The number of tweets retrieved for each @username ranged between 43 and 386. We drew a sample from a total of 16,393 tweets created between March 2008 and April 2013.

[AVG-USER]: Another random sample of 10,000 tweets was collected using the same technique described in RANDOM. From this sample, we randomly selected 82 @usernames belonging to individuals and not organizations or news agencies. We then collected tweet streams using these 82 @usernames as the query terms from the Twitter API. The number of tweets for each @username ranged between 2 and 248. Similar to SEN-USER, we drew a sample of tweets to be annotated from 31,556 tweets created between July – August 2014. We included only users with at least 100 retrieved tweets in the sample.

3.2 Phase 1: Small-scale Content Analysis

A key aspect of our annotation methodology is the use of open coding in which the categories to be used for annotation are derived from the data itself. Rather than use a predefined set of emotion categories, the categories used here were developed collectively by the annotators based on what they found in the tweets. Annotators were instructed to annotate the data based on the four facets of emotions described in Table 3.

Annotators were first asked to annotate the emotional valence of the tweet. Tweets were annotated as positive, negative, neutral or none (i.e., no emotion). We included a class for neutral emotions to account for emotions that were neither positively nor negatively valenced (e.g., surprise). For tweets that were labeled as positive, negative or neutral, annotators were then asked to create their own emotion labels to describe the emotion(s) expressed in the tweets. Annotators were asked to assign only one emotion label to each tweet (i.e., the best emotion tag to describe the overall emotion expressed by the tweeter) (Examples 1, 2 and 3). However, in cases where a tweet contains multiple emotions, annotators were asked to first identify the primary emotion expressed in the tweet and then include the other emotions observed (Example 4). As illustrated by the examples, the annotated corpus captures not only explicit expressions of emotion using emotion words but also implicit expressions of emotion.

Facet	Description	Codes
Valence	Expressing pleasure or displeasure towards events, objects or situations	Positive: Expressing pleasure Negative: Expressing displeasure Neutral: Emotion expressed is neither positive nor negative No Emotion
Arousal	Level of arousal/activation to the stimuli	1: Calm (Very low intensity) 2: Low intensity 3: Moderate intensity 4: High intensity 5: Very high intensity
Emotion Tag	Emotion category that best describes the emotion expressed in a tweet	Open coding
Emotion Cues	Words/phrases that influence annotators to annotate the tweet with a particular emotion tag	Open coding

Table 3: Classification schemes for 4 emotion facets

Example 1: IM SO HAPPY SOFIA IS OKAY THIS IS A MIRACLE
Valence: Positive; **Arousal:** 2
Emotion Tag [Emotion Cues]: Happiness [SO HAPPY, MIRACLE]

Example 2: OMFG MY DREAMS HAVE COME TRUE SOSUKE, AI AND MOMOTAROU ARE GETTING CHARACTER SONGS AS WELL AS STYLE FIVE @AnimatedSal
Valence: Positive; **Arousal:** 4
Emotion Tag [Emotion Cues]: Happiness [OMFG MY DREAMS HAVE COME TRUE]

Example 3: Wow that freeze! #PlayPokemon
Valence: Neutral; **Arousal:** 4
Emotion Tag [Emotion Cues]: Surprise [Wow]

Example 4: Saw Argo yesterday, a movie about the 1979 Iranian Revolution. Chilling, sobering, and inspirational at the same time
Valence: Positive, Negative; **Arousal:** 4
Emotion Tag [Emotion Cues]: Inspiration [inspirational], Fear [Chilling, sobering]

To annotate all 5,553 tweets, 17 students were recruited as annotators. Students undertook the task as part of a class project (a Natural Language Processing course) or to gain research experience in content analysis. Students were divided into groups and each group is assigned to annotate one of the four samples. All annotators went through the same training phase to encourage consistent annotation. The primary researcher performed annotations in every sample to ensure that the tweets were consistently annotated. Each tweet was annotated by at least three annotators. The annotations were done in an incremental fashion. In the first round, annotators were asked to perform open coding for emotion tag on 300 tweets for the sample they were assigned to.

To refine the set of emotion tags that emerged from data, annotators were then asked to perform a card sorting activity to group semantically similar emotion tags into the same category. Annotators were asked to collectively pick the most descriptive emotion tag to represent each category. Once the emotion categories were identified the original emotion tag labels in the first round of annotation for each tweet were replaced by the agreed-upon category labels. Annotators incrementally annotated more tweets in subsequent rounds until a point of saturation was reached, where new emotion categories stopped emerging from data. The final 28 emotion categories are shown in Table 7. The primary researcher met with the annotators after every round of annotation to discuss the disagreements, and 100% agreement for valence and emotion tag was achieved after discussions.

3.3 Phase 2: Large-scale Content Analysis

Using the annotation scheme developed in Phase 1, a larger set of manual annotations was obtained using Amazon Mechanical Turk (AMT) in Phase 2. To streamline the annotation process across a large pool of annotators, we developed a Web annotation application shown in **Figure 1**, which was tailored to our annotation scheme. The facets of emotion to be annotated were presented as a series of questions. 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 were allowed to suggest a new emotion tag for any tweets where none of the listed emotion category was applicable.

Recruitment of workers was done through Human Intelligence Tasks (HITs) on the online AMT platform. Of the 30 tweets in one HIT, 25 were unlabeled tweets and 5 were gold standard tweets with full agreement from Phase 1. Each tweet was annotated by at least three 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. Only about one third of the tweets had full agreement for emotion tag among all annotators (32%). To establish ground truth for machine learning experiments the primary researcher manually reviewed all annotations and resolved disagreements.

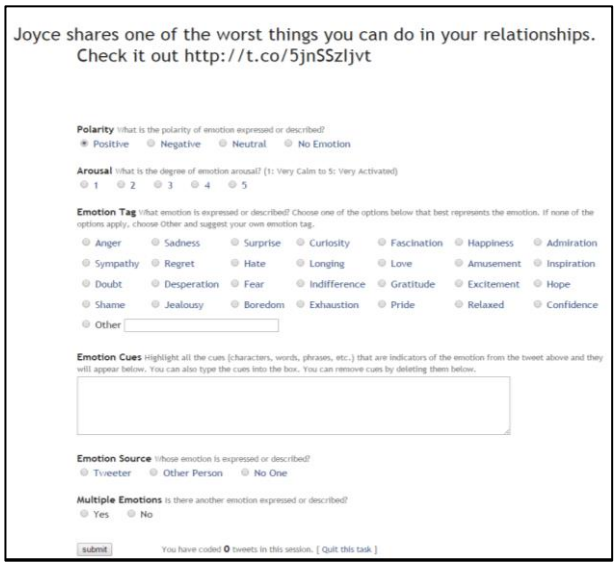

Figure 1: Web annotation application for data collection in Phase 2

4. Inter-annotator Agreement

Table 4 presents the inter-annotator agreement statistics for presence or absence of emotion, valence, arousal, 28 emotion categories and emotion cues for all tweets with three annotations. Krippendorff's alpha (α) is used as the primary measure of agreement for valence, arousal and emotion category as α can be applied for any number of annotators as well as for both nominal and ordinal variables. Percent agreement (%) and Fleiss' Kappa (κ) are also presented alongside α as a means to compare our results with common standards or guidelines.

Facet	Agreement	P1	P2	P1+P2
Emo/ Non-Emo	**%**	81	66	76
	α	0.62	0.29	0.51
	κ	0.62	0.29	0.51
Valence	**%**	77	60	71
	α	0.61	0.34	0.52
	κ	0.61	0.34	0.52
Arousal	**α**	0.59	0.32	0.5
EmoCat-28	**%**	66	51	61
	α	0.50	0.28	0.43
	κ	0.50	0.28	0.43
EmoCues	**MASI**	0.55	0.48	0.52

Table 4: Inter-annotator agreement statistics for emotion/non-emotion, valence, arousal, emotion category and emotion cue

Emotion cue captures the segment of text marked by annotators as the indicator of an emotion category. Unlike valence, arousal and emotion categories, emotion cue does not have a pre-defined set of categories and the boundary of the marked up text is not fixed. The size of an emotion cue varies from a single word to long strings of words within a tweet. We adopt the measure of agreement on set-valued items (MASI) to determine the agreement

between sets of text spans among multiple annotators for each tweet (Aman & Szpakowicz, 2007). MASI has been applied previously to quantify the reliability in co-reference annotation (Passonneau, 2004) and automatic summarization (Passonneau, 2006).

Mean α for 28 emotion categories across all rounds in P1 (annotated by expert annotators) is 0.50. With limited training, α scores in P2 decrease almost by half for all facets of emotion as shown in Table 4. It is important to note that agreement based on 28 emotion categories is not a great deal lower that that observed for other more coarse-grained facets of emotion. Annotators across P1 and P2 achieve overall $\alpha = 0.43$ when asked to identify 28 emotion categories, which is not a drastic drop compared to $\alpha = 0.52$ obtained from the four-class valence annotation. MASI scores for the emotion cues are more stable across P1 (MASI = 0.55) and P2 (MASI = 0.48), thus showing that there is less discord among expert and novice annotators when asked to identify written linguistic cues associated with emotion.

Since α is affected by dissimilar scales and the number of categories, care must be taken when making comparisons across different facets of emotion. Typically, a larger number of categories would lead to more disagreements, and thus lower α (Sim & Wright, 2005). Our P1 results are consistent with this general observation except for the anomaly in that valence annotation (4 classes: "positive", "negative", "neutral" and "no emotion") in P2 obtains slightly higher α compared with the binary emotion versus non-emotion annotation (2 classes: "has emotion" and "no emotion"). This led us to conclude there are high enough agreements among annotators in making the distinction between positive, negative and neutral instances to offset some of the disagreements in the binary emotion versus non-emotion annotation.

We acknowledge that overall inter-annotator agreement in detecting the 28 emotion categories is at best fair to good (κ between 0.40 – 0.75) according to the guidelines described in Fleiss, Levin, & Paik (2013). Emotion annotation is a subjective and difficult task. The overall inter-annotator agreement scores could be increased by removing some of the emotion categories with poor agreement (e.g., *relaxed*, *doubt* and *confidence*) or retraining annotators until a κ of above 0.75 is achieved for all facets of emotion. However, the use of inter-annotator agreement here is intended to show a realistic assessment of human performance in the exploratory annotation of the emotion categories that emerged from the open coding task.

All tweets in the corpus are assigned gold labels, which act as ground truth. For P1, all disagreements were first resolved through discussion with expert annotators. Essentially, expert annotators achieved 100% agreement in P1. For P2, we assigned the gold labels after manually reviewing all annotations provided by AMT workers. The manual review procedure was necessary to reduce as much as possible the noise from a large group of novice annotators.

5. Corpus Characteristics

EmoTweet-28 contains 15,553 tweets from P1 and P2. Overall, the corpus is composed of 247,872 words, of which 42,620 are unique terms. Message length is short with 16 words on average per tweet. The shortest tweet contains only one word while the longest tweet contains 40 words.

5.1 Emotion Distributions

This section describes the distribution of gold labels among the facets of emotion. As shown in Table 5, the overall distribution between tweets containing emotion and those that do not is roughly balanced; slightly over half of the tweets (51%) contain emotion. The ratios between emotion and non-emotion tweets respectively for RANDOM, TOPIC, SEN-USER and AVG-USER are similar. The biggest contribution of emotion tweets comes from TOPIC, and the lowest from SENUSER. The number of emotion tweets exceeds the number of non-emotion tweets in TOPIC and AVG-USER but the reverse is observed for RANDOM and SEN-USER.

Table 6 summarizes results for emotion valence. The overall corpus contains more than twice as many positive tweets than negative. This skew is especially apparent for SEN-USER with three quarters of the tweets annotated as positive and barely any as neutral. RANDOM, TOPIC and AVG-USER samples are similar in the proportion of positive, negative, and neutral tweets and are likely to be more representative samples of the true distribution on Twitter. About 7% of the corpus consists of tweets assigned with multiple valence labels (e.g., presence of positive and negative emotions in the same tweet).

Each tweet containing emotion is assigned a final arousal score, which is computed based on the mean arousal ratings provided by all the annotators. The data follows a roughly normal distribution with a slight skew to the right. On a 1 to 5 scale, mean arousal is 3.24.

Table 7 summarizes the frequency distribution of emotion categories. Tweets that are assigned with multiple emotion categories are counted more than one time. As expected, the frequency of emotion classes becomes even more unbalanced and sparse with a greater number of classes compared to valence. Of the 28 emotion categories, *happiness* is the most frequently occurring emotion (12% of the full corpus) whereas *jealousy* is the least frequent (0.2%). All four samples share one similarity: *happiness* occurs the most frequently in each sample. Other than that, the proportion of emotion categories differs across the four samples. For example, political leaders (SEN-USER) express more gratitude and much less anger on Twitter than a typical user (AVG-USER) indicating that leaders take a more controlled and strategic approach when expressing their emotions on Twitter. RANDOM, TOPIC and AVG-USER contribute at least a few positive instances of each emotion category. Three emotion categories are notably absent from SEN-USER: *boredom*, *indifference* and *jealousy*.

Class	P1	P2	P1+P2	RANDOM	TOPIC	SEN-USER	AVG-USER
Emotion	2916 (53%)	4953 (50%)	7869 (51%)	1775 (45%)	2281 (60%)	1615 (40%)	2198 (58%)
Non-Emotion	2637 (47%)	5047 (50%)	7684 (49%)	2175 (55%)	1529 (40%)	2378 (60%)	1602 (42%)
Total	**5553**	**10000**	**15553**	**3950**	**3810**	**3993**	**3800**

Table 5: Distribution of emotional and non-emotional tweets

Class	P1	P2	P1+P2	RANDOM	TOPIC	SEN-USER	AVG-USER
Positive	1840 (63%)	2846 (57%)	4686 (60%)	1022 (58%)	1306 (57%)	1259 (78%)	1099 (50%)
Negative	744 (26%)	1493 (30%)	2237 (28%)	538 (30%)	689 (30%)	276 (17%)	734 (33%)
Neutral	155 (5%)	222 (4%)	377 (5%)	87 (5%)	107 (5%)	24 (1%)	159 (7%)
Multiple Valence	177 (6%)	392 (8%)	569 (7%)	128 (7%)	179 (8%)	56 (3%)	206 (9%)
Total	**2916**	**4953**	**7869**	**1775**	**2281**	**1615**	**2198**

Table 6: Distribution of tweets based on emotion valence

Category	P1		P2		P1 + P2		RANDOM	TOPIC	SEN-USER	AVG-USER
	n	%	n	%	n	%	n	n	n	n
Admiration	158	2.8	245	2.5	403	2.6	112	70	113	108
Amusement	237	4.3	423	4.2	660	4.2	180	161	10	309
Anger	444	8.0	757	7.6	1201	7.7	267	399	137	398
Boredom	12	0.2	36	0.4	48	0.3	9	11	0	28
Confidence	19	0.3	91	0.9	110	0.7	32	34	16	28
Curiosity	30	0.5	63	0.6	93	0.6	25	28	4	36
Desperation	8	0.1	50	0.5	58	0.4	12	19	5	22
Doubt	50	0.9	108	1.1	158	1.0	32	49	12	65
Excitement	265	4.8	421	4.2	686	4.4	85	324	163	114
Exhaustion	10	0.2	39	0.4	49	0.3	10	15	4	20
Fascination	54	1.0	150	1.5	204	1.3	64	45	37	58
Fear	77	1.4	162	1.6	239	1.5	48	73	57	61
Gratitude	221	4.0	300	3.0	521	3.3	114	66	263	78
Happiness	778	14.0	1009	10.1	1787	11.5	302	483	604	398
Hate	63	1.1	129	1.3	192	1.2	54	44	4	90
Hope	187	3.4	335	3.4	522	3.4	102	208	108	104
Indifference	28	0.5	40	0.4	68	0.4	18	14	0	36
Inspiration	21	0.4	54	0.5	75	0.5	21	17	29	8
Jealousy	5	0.1	29	0.3	34	0.2	25	3	0	6
Longing	41	0.7	80	0.8	121	0.8	42	30	6	43
Love	234	4.2	447	4.5	681	4.4	248	217	31	184
Pride	85	1.5	128	1.3	213	1.4	42	29	123	19
Regret	49	0.9	104	1.0	153	1.0	43	42	9	59
Relaxed	26	0.5	51	0.5	77	0.5	14	29	11	23
Sadness	158	2.8	363	3.6	521	3.3	138	176	61	146
Shame	26	0.5	64	0.6	90	0.6	16	24	7	43
Surprise	93	1.7	173	1.7	266	1.7	57	99	26	84
Sympathy	35	0.6	66	0.7	101	0.6	13	36	44	8

Table 7: Frequency distribution of 28 emotion categories in the corpus

5.2 Multiple Emotions in a Tweet

Although tweets are short and contain a maximum of 140 characters, we also captured tweets tagged with multiple emotion categories during the annotation process. Users can be very expressive in conveying their emotions on Twitter even in such a short span of text. Such tweets have usually been excluded from existing gold standard corpora (Hasan, Rundensteiner, & Agu, 2014; Mohammad et al., 2014) to reduce complexity. In fine-grained emotion analysis, multiple emotions occur naturally so a corpus should represent the occurrences of such cases and not ignore them because it is easier. If the portion of tweets containing multiple emotion categories is high, including them in the corpus would help increase the number of positive examples for each emotion category.

Tweets that contain multiple emotion can be characterized in two ways: 1) expression of multiple emotions with the same valence being labeled as *multiple emotions* (Example 5), and 2) expressing multiple emotions with distinct valence being labeled as *multiple valence* (Example 6). The tweeter in Example 5 expressed three positive emotions in a single tweet: *gratitude* (*thank you so much*), *love* (*As a fan of the series*), and *excitement* (*i'm really looking forward to*). In Example 6, the tweeter expressed both a positive emotion, *happiness* (*Yay freedom!*) and a negative emotion, *anger* (*Fffffffff*) in the same tweet.

Example 5: @yenpress thank you so much for licensing kagerou project!!! As a fan of the series i'm really really looking forward to the release!!!!
[Multiple Emotions Same Valence: Gratitude, Love, Excitement]

Example 6: Yay freedom! *looks at traffic map* Ffffffffff-
[Multiple Emotions Different Valence: Happiness, Anger]

Category Count/Tweet	P1	P2	P1+P2
Single	5102 (92%)	9135 (91%)	14237 (92%)
Multiple	451 (8%)	865 (9%)	1316 (8%)
- Multiple: Same Valence	*274 (5%)*	*467 (5%)*	*741 (5%)*
- Multiple: Different Valence	*177 (3%)*	*398 (4%)*	*575 (3%)*
Total	**5553**	**10000**	**15553**

Table 8: Distribution of tweets containing single and multiple emotion categories

As shown in Table 8, the corpus contains a large number of tweets tagged with a single emotion category (92%) and only 8% of tweets tagged with more than one emotion category. Mohammad et al. (2014) reported 2% of their 2012 US presidential elections corpus comprises of tweets with two or more contrasting emotions. Our findings are consistent with previous observation although the proportion of tweets with multiple emotion categories is higher in our corpus. The emotion categories in our annotation scheme are more fine-grained which naturally lead to more tweets being tagged with multiple emotion categories.

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). Tweets annotated with only a single emotion produce only 6553 positive examples. The inclusion of tweets annotated with multiple emotions increases the number of positive examples to 9331. This is especially beneficial for categories that suffer from sparseness of positive examples such as *jealousy*, *boredom* and *exhaustion*. Overall, including tweets containing multiple emotions gives each emotion category a boost in frequency, notably for *happiness* and *love*.

6. Emotion Classification

We frame the classification problems as a multi-label classification task, where each instance could be assigned to more than one emotion category label. Using Weka (Hall et al., 2009), separate binary classifiers were built for each emotion category to detect if an emotion category were present or absent in a tweet. The precision, recall and F1 of an emotion classifier based on Support Vector Machines (SVM) with Sequential Minimal Optimization (SMO) and evaluated using 10-fold cross validation are shown in Table 9. Only stemmed and lowercased unigrams with minimum word frequency of 3 were used as features. We also normalized the hyperlinks (URLs)[1] in the tweets and include a feature to indicate the presence or absence of URL in a tweet. We ran a large number of experiments but present only a representative example in this paper.

The F1 scores ranged from the highest of 0.92 for *gratitude* to the lowest of 0.09 for *indifference*. The classifier performed extremely well for certain emotion categories with clear lexical patterns while poorer performance is observed for the sparse and obscure categories. These binary classifiers use very basic combination of features and machine learning algorithm and are not yet optimized to yield the best performance per category.

7. Conclusion and Future Work

We present EmoTweet-28, a tweet corpus developed using 4 different sampling strategies and annotated with 28 emotion categories. The corpus contains tweets annotated with multiple emotions and captures the language used to express an emotion explicitly and implicitly. The corpus annotated with fine-grained emotion categories represents an important resource to advance the development and evaluation of more

[1] URLs in the tweets are normalized to http://URL.

emotion-sensitive systems. We show classifiers can perform extremely well on certain fine-grained emotion categories by training and testing a basic unigram SMO classifier using the corpus. In the future, we will perform linguistic analysis to extract salient linguistic patterns associated with each emotion category and leverage them to improve classification performance.

Category	Precision	Recall	F1
Admiration	0.370	0.201	0.260
Amusement	0.869	0.645	0.741
Anger	0.478	0.321	0.384
Boredom	0.818	0.375	0.514
Confidence	0.303	0.091	0.140
Curiosity	0.638	0.548	0.590
Desperation	0.500	0.069	0.121
Doubt	0.269	0.089	0.133
Excitement	0.655	0.474	0.550
Exhaustion	0.611	0.224	0.328
Fascination	0.553	0.309	0.396
Fear	0.491	0.230	0.313
Gratitude	0.928	0.914	0.921
Happiness	0.622	0.506	0.558
Hate	0.788	0.542	0.642
Hope	0.781	0.580	0.666
Indifference	0.235	0.059	0.094
Inspiration	0.816	0.413	0.549
Jealousy	0.765	0.382	0.510
Longing	0.529	0.306	0.387
Love	0.659	0.519	0.581
Pride	0.862	0.676	0.758
Regret	0.514	0.242	0.329
Relaxed	0.737	0.182	0.292
Sadness	0.650	0.461	0.539
Shame	0.622	0.311	0.415
Surprise	0.556	0.278	0.371
Sympathy	0.705	0.426	0.531
Macro-avg	**0.619**	**0.370**	**0.450**
Micro-avg	**0.656**	**0.455**	**0.537**

Table 9: Precision, recall and F1 for 28 emotion categories

8. Acknowledgements

We thank the annotators who volunteered in performing the annotation task. We are immensely grateful to Christine Larsen who partially funded the data collection under the Liddy Fellowship.

9. References

Aman, S., & Szpakowicz, S. (2007). Identifying expressions of emotion in text. In *Text, Speech and Dialogue*, pp. 196–205.

Fleiss, J. L., Levin, B., & Paik, M. C. (2013). The measurement of interrater agreement. In *Statistical methods for rates and proportions*. John Wiley & Sons.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter*, *11*(1), 10–18.

Hasan, M., Rundensteiner, E., & Agu, E. (2014). EMOTEX: Detecting emotions in Twitter messages. Presented at the 2014 ASE BIGDATA/SOCIALCOM/ CYBERSECURITY Conference, Stanford University.

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology*. Newbury Park, CA: Sage.

Mohammad, S. M. (2012). #Emotional tweets. In *Proceedings of the 1st Joint Conference on Lexical and Computational Semantics*, pp. 246–255. Montreal, QC.

Mohammad, S. M., Zhu, X., & Martin, J. (2014). Semantic role labeling of emotions in tweets. In *Proceedings of the ACL 2014 Workshop on Computational Approaches to Subjectivity, Sentiment, and Social Media (WASSA)*, pp. 32–41. Baltimore, MD.

Nakov, P., Kozareva, Z., Ritter, A., Rosenthal, S., Stoyanov, V., & Wilson, T. (2013). SemEval-2013 task 2: Sentiment analysis in Twitter. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*, pp. 312–320.

Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In *7th International Conference on Language Resources and Evaluation (LREC)*, pp. 1320–1326.

Passonneau, R. (2004). *Computing reliability for coreference annotation*.

Passonneau, R. (2006). Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, pp. 831–836.

Purver, M., & Battersby, S. (2012). Experimenting with distant supervision for emotion classification. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 482–491. Stroudsburg, PA, USA.

Roberts, K., Roach, M. A., Johnson, J., Guthrie, J., & Harabagiu, S. M. (2012). EmpaTweet: Annotating and detecting emotions on Twitter. In *8th International Conference on Language Resources and Evaluation (LREC)*, pp. 3806–3813.

Rosenthal, S., Nakov, P., Ritter, A., & Stoyanov, V. (2014). Semeval-2014 task 9: Sentiment analysis in Twitter. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pp. 73–80. Dublin, Ireland.

Sim, J., & Wright, C. C. (2005). The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy*, *85*(3), 257–268.

Wang, W., Chen, L., Thirunarayan, K., & Sheth, A. P. (2012). Harnessing Twitter "big data" for automatic emotion identification. In *2012 International Conference on Privacy, Security, Risk and Trust (PASSAT) and 2012 International Conference on Social Computing (SocialCom)*, pp. 587–592.