Distant Supervision for Emotion Classification with Discrete Binary Values

Jared Suttles[1] and Nancy Ide[1]

Department of Computer Science, Vassar College
Poughkeepsie, New York 12604 USA
{jasuttles, ide}@cs.vassar.edu

Abstract. In this paper, we present an experiment to identify emotions in tweets. Unlike previous studies, which typically use the six basic emotion classes defined by Ekman, we classify emotions according to a set of eight basic *bipolar* emotions defined by Plutchik (Plutchik's "wheel of emotions"). This allows us to treat the inherently multi-class problem of emotion classification as a binary problem for four opposing emotion pairs. Our approach applies *distant supervision*, which has been shown to be an effective way to overcome the need for a large set of manually labeled data to produce accurate classifiers. We build on previous work by treating not only emoticons and hashtags but also emoji, which are increasingly used in social media, as an alternative for explicit, manual labels. Since these labels may be noisy, we first perform an experiment to investigate the correspondence among particular labels of different types assumed to be indicative of the same emotion. We then test and compare the accuracy of independent binary classifiers for each of Plutchik's four binary emotion pairs trained with different combinations of label types. Our best performing classifiers produce results between 75-91%, depending on the emotion pair; these classifiers can be combined to emulate a single multi-label classifier for Plutchik's eight emotions that achieves accuracies superior to those reported in previous multi-way classification studies.

1 Introduction

The development of web- and mobile-based media devoted to persistent social interaction among users ("social networks") has provided a massive, continuous stream of data reflecting the public's opinions about and reactions to phenomena from political and world events to movies and consumer products. Over the past ten years, there has been no shortage of studies attempting to mine this data to inform decisions about product design, brand identity, corporate strategy, government policies, etc., as well as improve social-psychological correlational studies and predictive models of human behavior. Recently, many analyses have focused on the microblogging service Twitter, which provides a continuous stream of user-generated content in the form of short texts under 140 characters in length. Much of this work involves *sentiment* analysis, in which user attitudes toward a particular topic or product are classified as positive, negative, or

neutral (e.g., [13,21]). Other studies have tackled the broader problem of detecting *emotions* in tweets, often for the purpose of modeling collective emotional trends [4, 5, 10, 24].

In this paper, we present an experiment to identify emotions in tweets. Unlike previous studies, which typically use the six basic emotion classes defined by Ekman [11, 12], we classify emotions according to a set of eight basic *bipolar* emotions defined by Plutchik (Plutchik's "wheel of emotions" [22]). This allows us to treat the inherently multi-class problem of emotion classification as a binary problem for four opposing emotion pairs. Our approach applies *distant supervision* (see e.g. [17]), which has been shown to be an effective way to overcome the need for a large set of manually labeled data to produce accurate classifiers (e.g., [13, 24]). We build on previous work by treating not only emoticons and hashtags but also emoji, which are increasingly used in social media, as an alternative for explicit, manual labels. Since these labels may be noisy, we first perform an experiment to investigate the correspondence among particular labels of different types assumed to be indicative of the same emotion. We then test and compare the accuracy of independent binary classifiers for each of Plutchik's four binary emotion pairs trained with different combinations of label types. Our best performing classifiers produce results between 75-91%, depending on the emotion pair; these classifiers can be combined to emulate a single multi-label classifier for Plutchik's eight emotions that achieves accuracies superior to those reported in previous multi-way classification studies.

2 Previous work

Several studies have focused on the task of identifying *emotions* in different text types, including stories [2, 18, 25], spoken data [7, 8, 15], blogs [16, 20], and microblogs (tweets) [19,24,27]. Earlier studies relied on datasets that were manually annotated for emotion and were typically *keyword-based*, identifying the presence of an emotion based on the appearance of pre-determined lexical markers. It is well-recognized that this approach has drawbacks: determining the contents of the emotional lexicon is subjective, and there is no guarantee that the lexicon is comprehensive; furthermore, the selected words may be ambiguous. These problems are compounded when performing sentence-level analyses where very little context is available, which is clearly a factor in studies involving context-poor Twitter messages.

To address this and the problem of generating large annotated datasets for training, several studies have attempted to exploit the widespread use of emoticons and other indicators of emotional content in tweets by treating them as noisy labels in order to automatically obtain very large training sets (see e.g., [13, 19, 24]). This strategy of *distant supervision* [17] has been used to achieve accuracy scores as high as 80-83% for distinguishing positive and negative sentiment [13]. Studies using distant supervision commonly rely on a set of Western-style emoticons (e.g., "`:-)`", "`:(`", etc.) and Eastern-style emoticons (e.g., "`(^_^)`", "`(>_<)`", etc.) as emotional labels [21,26,29]. The means by which

these labels are associated with specific emotions varies from study to study–the most common strategy is to manually classify emoticons such as those available from on-line emoticon lists (e.g., Wikipedia *List of Emoticons*[1], Yahoo messenger classification[2]) as indicative of a specific emotion. The most commonly-used scheme for emotion classification is Ekman's [11,12], which identifies six primary emotions based on facial expressions.

Recently, there has been work exploring the use of Twitter *hashtags* to collect datasets indicative of emotional states for distant supervision. Hashtags, consisting of a tag or word prepended with "#" are typically used to indicate the tweet's topic in order to facilitate search and increase visibility. However, the practice of using hashtags has extended to other kinds of labeling, in particular, noting attitudes such as `#sarcasm` and `#irony` as well as emotional states (`#angry, #happy, etc.`). Previous studies collected tweets with specific hashtags to create datasets of sarcastic tweets [14]; recently, this approach has been applied to hashtags signaling the presence of particular emotions [19, 24, 30]. Again, the means by which hashtags are associated with particular emotions varies, but most studies use the names of Ekman's six basic emotions as relevant hashtags [6, 19], sometimes together with a few closely related terms [24]. However, the number of messages containing this small set of words as hashtags is typically very small, as noted in [24]. To increase the number of relevant terms, others have relied on pre-compiled lists of emotion words from psychological literature [30]. Our strategy, described in Section 3.2, differs from previous studies by using hashtags extracted from a large database of current tweets that have been manually labeled for emotional content.

3 Methodology

3.1 Emotional Binaries

Our work relies on a set of eight basic *bipolar* emotions as defined in Plutchik's psychoevolutionary theory of emotion [22] rather than the six basic emotion classes defined by Ekman [11] or previously-used minor variants [2,27]. Ekman's basic emotions include ANGER, DISGUST, FEAR, HAPPINESS, SADNESS, and SURPRISE; Plutchik's theory defines eight primary emotions, consisting of a superset of Ekman's and with two additions: TRUST and ANTICIPATION. These eight emotions are organized into four bipolar sets: JOY vs. SADNESS, ANGER vs. FEAR, TRUST vs. DISGUST, and SURPRISE vs. ANTICIPATION. Plutchik's "wheel of emotions" (see Figure 1) represents the relations among emotions as a color wheel; like colors, emotions can vary in intensity (proximity to the center indicates intensity) and mix to create additional emotions (*primary dyads*, appearing in the white spaces between primary emotions). Most relevant to our work is Plutchik's

[1] http://en.wikipedia.org/wiki/List_of_emoticons
[2] `http://messenger.yahoo.com/features/emoticons/`

definition of *emotional opposites*, represented in the spatial oppositions in the wheel, which are considered to be mutually exclusive.[3]

We adopted Plutchik's model over Ekman's for several reasons. First, it includes LOVE, an emotion very frequently expressed on Twitter. In Plutchik's scheme, LOVE is defined as a *primary dyad*, i.e., a combination of the two primary emotions JOY and TRUST; Ekman's set of six emotions, grounded in physiological rather than psychological research, omits LOVE and is in general more focused on negative emotions. The main advantage of using Plutchik's theory for our work is that it allows us to exploit his notion of *emotional polar opposites* to treat emotion detection as a binary rather than multi-way classification problem. Whereas previous studies used multiple category classification for emotion detection (see, e.g., [7,8,28]) or simulated binary classification by distinguishing one emotion class from all others (e.g. ANGER v. NOT ANGER, [19,24]), the classifiers used in this study make binary decisions concerning which of each pair of opposing emotions is most probable, thereby likening the problem to that of distinguishing two opposing classes (e.g., positive vs. negative sentiment) rather than presence or absence of a class among several others. This simplification enables development of four independent binary classifiers, one for each binary emotion pair, that can be combined to emulate a single multi-label classifier for Plutchik's eight primary emotions.

3.2 Emotion Lexicon

Our lexicon comprises a combination of emotional labels including hashtags, traditional emoticons, and emoji. It is assumed that the use of any of these symbols reflects the emotion of the author of the tweet, even when the emotional state of another individual is the topic. Support for this assumption is provided by studies on internet-based social interactions and the representation of emotions (e.g., [9]), which show that emoticons are used to increase the intensity of emotions already conveyed by the lexical content. It has also been suggested that "emotional punctuations" (e.g., noting laughter) in spoken transcriptions are similar to written emoticons, with both acting as punctuation for the surrounding language [23].

Our lexicon of 69 emoticons was derived from Wikipedia[4]. The emotion class assignments were based on those used in previous studies [1, 13, 21, 24]. Our lexicon also includes *emoji*[5], which originally developed in Japan but have come into widespread use since their inclusion in Unicode Standard 6.0 and ISO/IEC 10646 (Universal Character Set) and subsequent support in newer operating systems and mobile phones. Despite their increasing prevalence, emoji have not been used in previous work[6] In the absence of existing categorizations, we labeled

[3] A recent study [27] adapted Ekman's classification to define an emotional ontology and a set of emotional oppositions very similar to those in Plutchik's Wheel.

[4] http://en.wikipedia.org/wiki/List_of_emoticons

[5] http://en.wikipedia.org/wiki/Emoji

[6] Of the 38.9 million emotional tweets in our dataset, 7% include emoji from our lexicon and 7.8% contain emoticons from our lexicon.

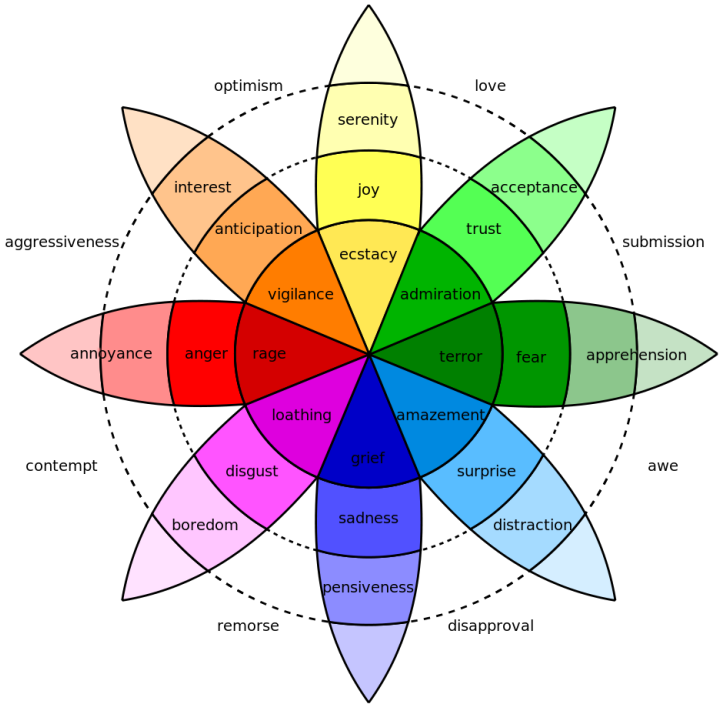

Fig. 1. Plutchik's Wheel of Emotions (Image from Wikimedia Commons)

70 emoji (consisting of facial expressions and a few additional symbols such as hearts, kissing lips, etc.) with the eight Plutchik primary emotion categories.

Our initial approach to determining the hashtags to be included in the lexicon used the eight primary emotion names defined by Plutchik (ANGER, DISGUST, FEAR, JOY, SADNESS, SURPRISE, TRUST, ANTICIPATION) as seed words, and added the WordNet 3.1 synsets and hyponyms of each name in order to create a set of terms for each emotion. This resulted in a large set of over 60 terms for each emotion. We later abandoned this method because the WordNet terms–and to a lesser extent the emotion names themselves–occurred infrequently in the data. Rather, users tend to use shorter, more colloquial hashtags instead of the words in WordNet synsets; for example, users prefer the tag `#ew` to a longer and more formal term `#disgusted`. We therefore turned to the data itself to determine a set of hashtags that reflect actual user behavior. Using a list of the most frequent hashtags in our training set, we identified those that are likely to be emotional labels (e.g., `#happytweet, #ugh, #yuck, #fml`). This method of determining a set of relevant hashtags maximized our ability to col-

lect a large number of labeled tweets, since the hashtags were guaranteed to appear frequently in our dataset. It also provides a more representative sampling of typical tweets in terms of word use and content and avoids selecting for unusual tweets containing infrequent hashtags. We also filtered out ambiguous tags such as `#sad`, which, in addition to occurring in its sense of "experiencing or showing sorrow or unhappiness" (WordNet3.1 sense 1) occurs frequently in the sense of "bad; unfortunate" (WordNet3.1 sense 3–e.g., "`Christina Aguilera used to have the best body.. then she got fat. #sad`"), which is closer to DISGUST.

We next assigned one of Plutchik's eight primary emotions to each of the selected hashtags. In cases where a hashtag seems indicative of one of the primary dyad (combined) emotions, the hashtag was associated with both of the primary emotions that comprise it–for example, `#love` was assigned to both JOY and TRUST, which combine into the complex emotion "love" according to Plutchik. Ultimately, we assigned 56 hashtags to Plutchik's eight primary emotion classes.[7]

3.3 Data Collection and Preparation

The data used in this study consists of microblog messages ("tweets") collected in real-time from the Twitter Streaming API service[8], which provides a 1-2% random sample of all tweets produced during the connection. We use the streaming API rather than sampling on specific query terms to avoid bias introduced by limiting the collection to tweets containing specific search terms, and to obtain a more representative sample of language from the average twitter user. Data collection was continuous over the period November 9 through November 30, 2012, thus eliminating any bias due to the influence of time of day or day of the week. Because our goal is to provide real-time monitoring of emotional trends in the Untied States, we limited the data to tweets produced by users within the US by imposing latitude and longitude constraints on the extracted messages in addition to specifying a country parameter. We also filtered for English language messages through the language parameter. The resulting dataset consists of 38.9 million tweets.

We extracted a dataset D_k consisting of 5.9 million tweets from the 38.9 million tweet dataset containing any of the emotional tokens in our lexicon and labeled each with the corresponding emotion. Tweets with multiple emotional tokens were assigned a label for each of the associated emotion classes. We included tweets with labels appearing both within (i.e., as a part of the message, as in "`I am so #angry about that!`") and at the end of the tweet; it has been suggested that in-line labels are less reliable indicators for sarcasm [14], but examination of our data does not support this observation for emotions. Tweets containing one or more emotional tokens from both classes of an opposing binary

[7] The complete emotional lexicon used in this study is available at http://www.emotitweets.com.

[8] See http://dev.twitter.com/docs/streaming-api

pair were discarded, since the emotional content was considered to be undecidable based on Plutchik's assumption of exclusivity of opposite emotions. Table 1 shows the distribution of labels for each emotion in the initial dataset.

Label Type	Joy	Sadness	Anticipation	Surprise
Hashtag	54,172	29,325	24,008	35,871
Emoticon	1,692,711	352,527	128,287	68,478
Emoji	735,023	275,861	24,133	26,363
Hashtag+Emoticon	1,741,767	379,571	152,005	104,120
Hashtag+Emoji	786,594	303,490	48,069	62,052
Emoticon+Emoji	2,419,383	625,398	152,277	94,765
All	2,465,884	650,771	175,923	130,220
Label Type	Anger	Fear	Disgust	Trust
Hashtag	31,109	25,066	25,724	30,501
Emoticon	101,939	128,287	101,842	454,768
Emoji	196,936	344,978	287,583	847,695
Hashtag+Emoticon	132,736	152,931	127,343	483,781
Hashtag+Emoji	226,565	368,792	312,381	874,633
Emoticon+Emoji	297,888	472,773	388,197	1,298,420
All	327,208	496,160	412,777	1,323,897

Table 1. Distribution of emotional labels in D_k.

The data were tokenized and normalized as follows: following [1, 13, 21], we replaced usernames (names prepended with "@") with the token USERNAME and web addresses (e.g. http://t.co/zDO9b7xD) with the token URL, and replaced repetitions of more than two letters consecutively (e.g. "cuteee", "cuteeeee", etc.) with only two, on the assumption that the number of repeating letters was arbitrary. Because we are interested in the emotions of the authors of tweets, quoted text was excluded as it may represent a retweet or someone else's opinion.

We compiled a training dataset D_t consisting of subsets corresponding to each of the four binary emotion pairs: D_t^1 (joy/sadness), D_t^2 (anticipation/surprise), D_t^3 (anger/fear), D_t^4 (trust/disgust). We used the labels appearing in our emotional lexicon to group tweets from D_k into emotion classes within the appropriate D_t^n set, then removed them so that classification would rely solely on language and non-emotional hashtags. The dataset for each binary emotion pair was normalized so that there were equal numbers of tweets for each member of the pair. As such, the total number of tweets for each emotion pair differed in proportion to the number appearing in the D_k tweet dataset, in which occurrences of JOY far outweigh those for other emotions (see Figure 2). The resulting training set contained approximately three million tweets, with each emotion pair (in equal numbers for each emotion in a pair) represented as shown in Table 2.

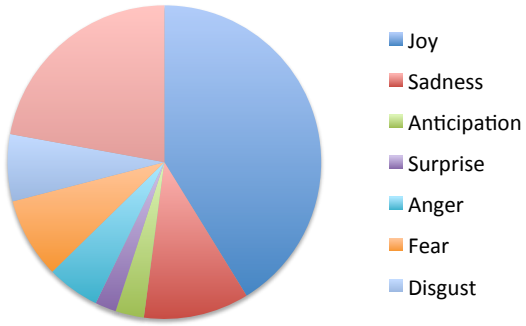

Fig. 2. Emotion proportions based on labels in D_k.

Training set	Size
D_t^1 (joy/sadness)	1,301,542
D_t^2 (anticipation/surprise)	260,440
D_t^3 (anger/fear)	654,416
D_t^4 (trust/disgust)	825,554
D_t	3,041,952

Table 2. Distribution of tweets for each of four datasets D_t^n.

4 Experiments

We performed two experiments: (1) a cross-validation of emotional class assignments to the different label types, to investigate their correspondence; and (2) evaluation of binary classifiers trained with various combinations of label types on a small manually-labeled dataset of emotional tweets. In all experiments, classification was performed using Naïve Bayes (NB) and Maximum Entropy (ME) from the Natural Language Toolkit (NLTK) 2.0.4[9]. Because of the size of the input, all experiments were run using concurrent algorithms on a machine with 158 2.2 GHz AMD Opteron 6174 12-core CPUs. Because of the long running times for training, our experiments include only unigrams as features; however, previous studies [13, 24] have shown that classifiers trained on unigrams outperform those trained on additional phenomena such as bigrams and part-of-speech information.

4.1 Experiment 1: Cross-validation of emotional labels

Our approach relies on the assumption that sets of hashtags, emoticons, and emoji associated with the same emotion are indeed indicative of the same underlying phenomenon. To validate this assumption, we tested the ability of each

[9] http://www.nltk.org

label to predict the emotion(s) signaled by the other labels. Separate binary classifiers for each label convention were trained on each dataset $D_t^n, 0 < n \leq 4$, using the set of emotion labels for that convention as noisy labels. We evaluated against the same 12 subsets (3 label types, 4 binaries) of D_t^n.

Accuracies are given in Table 3. Experiments used both Naïve Bayes and Maximum Entropy for classification. Ten-fold cross-validation was used for within-label tests. Full sets were used for all other tests. The values in the table show the highest accuracies; with a few exceptions, the accuracies returned by the two different classifiers were no more than a few percentage points apart. All accuracies are significantly higher than chance according to χ^2 tests, except one, shown with strikethrough, which is in fact significantly *lower* than chance. We suspect this is because we have only two emoji for anticipation in our lexicon, although further investigation is needed to determine the actual cause.

The results show that many of the classifiers trained on data labeled using one label type can distinguish classes that were labeled with one or both of the other two, suggesting that the emotion assignments are relatively reliable, or at least consistent, among the three label types. The clear exception is the lack of ability for hashtags to predict emoticons and emoji for ANTICIPATION; this likely results from the fact that very few emoticons and emoji can be considered indicative of anticipation (which has no obvious facial depiction), whereas hashtags such as `#cantwait` and `#excited` unambiguously signal this emotion. ANTICIPATION is one of the two primary emotions that is included in Plutchik's scheme but not in Ekman's[10], the scheme most commonly used in previous studies, which means that there exists no established set of labels for this emotion nor comparative data from other work. We therefore repeated our experiments with a variety of different ANTICIPATION emoji and emoticons, but these variants did not improve our results and in some cases actually worsened them. At the least, our results suggest that hashtags are likely a better source for automatic labeling of this emotion in tweets.

In general, accuracies are more consistently high for JOY and SADNESS than the other emotions. This result is similar to that reported in [24], where cross-label testing for classifiers trained on emoticons and hashtags performed relatively well for distinguishing JOY ("happy"), SADNESS ("sad"), and ANGER as compared with the other three emotions in their study[11], although their accuracies overall were much lower than ours (60-65% range).

4.2 Experiment 2: Classifier evaluation

Evaluation was performed using a manually labeled set D_m of 420 tweets that is disjoint from either D_e and D_t, consisting of 400 emotional tweets annotated for at least one emotion from at least one emotion pair, and 20 neutral tweets with

[10] The lack of a pictorial representation for ANTICIPATION may in fact account for its absence in Ekman's emotion scheme, which is based on facial expressions.

[11] [24] uses Ekman's six emotions.

Test		Train		
Label	Emotion	Hashtag	Emoticon	Emoji
Hashtag	Joy	73.8%	95.2%	94.5%
	Sadness	92.1%	*98.8%*	*99.3%*
	Anticipation	82.8%	*62.8%*	*34.5%*
	Surprise	*86.5%*	81.5%	61.1%
	Anger	*74.7%*	26.3%	*98.6%*
	Fear	78.1%	77.8%	79.8%
	Disgust	*82.2%*	*87.0%*	*92.0%*
	Trust	*85.2%*	92.8%	96.1%
Emoticon	Joy	93.3%	78.3%	89.0%
	Sadness	98.5%	*83.0%*	*95.4%*
	Anticipation	15.0%	*81.4%*	*92.6%*
	Surprise	93.8%	70.7%	88.5%
	Anger	36.9%	*63.6%*	*47.0%*
	Fear	51.7%	89.6%	56.7%
	Disgust	58.9%	*85.2%*	*58.3%*
	Trust	82.4%	82.1%	89.7%
Emoji	Joy	81.3%	85.0%	75.8%
	Sadness	98.0%	*96.9%*	*83.3%*
	Anticipation	~~*5.8%*~~	*97.5%*	*80.8%*
	Surprise	*90.1%*	59.1%	65.2%
	Anger	86.2%	46.6%	*81.9%*
	Fear	*87.0%*	*42.1%*	70.8%
	Disgust	*83.6%*	*91.4%*	*81.6%*
	Trust	*88.7%*	74.1%	77.5%

Table 3. Highest accuracies for cross-validation of emotional labels on datasets D_t^n. Values in italics used a Naïve Bayes classifier, non-italics used Maximum Entropy.

no emotion from any pair. Because the collection of tweets was random, the distribution of emotion classes in D_m is roughly proportional to their representation in D_k and D_e.

Annotation was performed by two annotators. The annotation procedure presented a randomly selected tweet to the annotator together with five annotation options. For example, "joy/sadness" is presented as follows:

```
omg I freaking love sweet potatoes! Literally ate one today!

[1] joy
[2] sadness
[3] neutral
[4] don't know for this emotion pair
[5] don't know for any pair (leave tweet out of dataset)
```

Options 4 and 5 allow the annotator to identify tweets that are difficult to understand and/or rate, either for a particular emotion or any emotion. Each tweet was annotated for all four emotional binary pairs. In cases where the annotator identified the tweet as "neutral" (option 3) for all four emotion pairs, the tweet was labeled NEUTRAL (non-emotional).

Evaluation was performed for classifiers trained using each label type as well as all possible combinations of labels. The accuracies for this experiment are given in Table 4. All values were found to be significantly different from chance based on χ^2 tests, except one (shown with strikethrough).

Train	Test					
	Joy/Sadness			Anticipation/Surprise		
	ME	NB	Size	ME	NB	Size
Hashtag	86.3%	73.8%	58,650	66.1%	60.3%	48,016
Emoticon	89.1%	84.8%	705,054	68.8%	69.8%	136,956
Emoji	88.7%	80.1%	551,722	64.0%	67.2%	48,266
Hashtag+Emoticon	**91.0%**	84.0%	759,142	73.0%	**75.7%**	208,240
Hashtag+Emoji	88.7%	80.1%	606,980	72.0%	72.5%	96,138
Emoticon+Emoji	90.2%	83.6%	1,250,796	65.1%	70.9%	189,530
All	90.6%	85.5%	1,301,542	71.4%	**75.7%**	260,440

Train	Test					
	Anger/Fear			Disgust/Trust		
	ME	NB	Size	ME	NB	Size
Hashtag	78.5%	74.6%	50,132	90.6%	86.6%	51,448
Emoticon	58.5%	~~49.2%~~	203,878	85.1%	87.1%	203,684
Emoji	80.8%	78.5%	393,872	90.1%	82.2%	575,166
Hashtag+Emoticon	70.0%	62.3%	265,472	89.1%	88.1%	254,686
Hashtag+Emoji	80.8%	79.2%	453,130	90.6%	84.2%	624,762
Emoticon+Emoji	**84.6%**	80.8%	595,776	89.1%	85.1%	776,394
All	83.1%	82.3%	654,416	**91.1%**	84.7%	825,554

Table 4. Evaluation results from Experiment 2. Values in bold are the highest scores for each emotion pair. Strikethrough identifies values that are not statistically significant.

Experiment 2 yields accuracies between 75% and 91%[12] for tests on manually labeled data, which exceed those reported in similar studies [4,24,28]. The results indicate that combining all three label types as distant labels yields the highest accuracies, or accuracies within (roughly) a percentage point of the highest. The remaining values are relatively consistent and reveal no pattern indicating that a particular label combination out-performs the others. The only anomaly in the results is the low accuracies for emoticons on "anger/fear", but this may be due to the difficulty of depicting fear with an emoticon (emoji provide a somewhat better depiction), making that pair particularly difficult to distinguish for emoticons alone.

We attribute our stronger results both to the use of binary classifiers, which reduces the complexity of the classification task, and to the inclusion of emoji as well as hashtags and emoticons as (noisy) labels for creating the training set. Inclusion of emoji provides more labeled data, and with more label types and more data, the classifier is less likely to be led down the wrong path by certain correlations between one label type and another (e.g., if the emoticon `:(` were used to indicate surprise by a large portion of writers), thus effectively giving us "ensemble noisy labeling".

Experiments 1 and 2 together give us some confidence the various labels actually signal the emotions we are assuming they do. That is, while the results from Experiment 1 verify the cross-label consistency of emotion assignments, they do not provide evidence that the assigned emotions correspond with human judgement. The strong results from Experiment 2, which uses manually labeled data, shows that the emotions associated with the labels are also reasonably consistent with independent human judgements, providing evidence that the associations made in the emotion lexicon are valid.

5 Next Steps

Our goal is to use the binary classifiers for the four emotion pairs to emulate a single multi-way classifier that identifies emotions in tweets. In fact, this combination of classifiers would identify up to four emotions (i.e., at most one from each pair of mutually exclusive emotions) in a tweet, which is appropriate since annotators identified multiple emotions in a large percentage of tweets in our manually labeled dataset. However, we also need to distinguish tweets containing no emotional content (which is the vast majority of tweets) from those containing an emotion from one or more of the four pairs. To address this, we have begun experimenting with four *neutral* binary classifiers, one for each emotion pair, that distinguishes tweets containing either of the emotions in that pair from those that do not, that is, tweets that include any of the six remaining emotions in Plutchik's system or contain no emotion at all. In turn, the combination of a classifier for one of the four emotion pairs with its corresponding neutral classifier would emulate a single three-way classifier that identifies each

[12] Accuracies fall between 85% and 91% if we eliminate the problematic "anticipation/-surprise" class.

tweet as containing one of the emotions in the pair or as emotionally neutral; subsequently combining the three-way classifiers for each of the four emotion pairs as shown in Figure 3 emulates a more complex multi-way classifier that identifies *all* of the emotions present in a tweet *or* labels it as non-emotional.

To train a neutral classifier for each emotion pair, we can use the results from the classifiers with the highest accuracies from Experiment 2. Since these classifiers return one emotion of a binary pair for any tweet, even when neither is present, we assume that results with lower probabilities reflect situations where the tweet actually contains neither emotion or contains no emotion at all. Based on this assumption, we determine the optimum cutoff probabilities for each emotion–that is, the value below which probabilities reported by the relevant emotional binary classifier identify tweets that do not contain one emotion from the pair or are emotionally neutral–by iterating over all possible probabilities to determine the one that best predicts the results in the manually labeled dataset. Once this process is complete, the cutoff values with maximum accuracy are retained for classification.

We have so far applied this procedure to create a first set of neutral classifiers for each emotion pair. We performed a two-fold cross-validation of these classifiers using the manually labeled dataset D_m. The accuracies are given in Table 5. Accuracies for JOY and SADNESS, and to a slightly lesser extent ANGER and FEAR, are reasonable, suggesting that it may be possible to develop a reliable multi-way classifier, at least for these emotion pairs.

Emotion binary	Accuracy
Joy/Sadness	82.9%
Anticipation/Surprise	44.6%
Anger/Fear	74.7%
Disgust/Trust	61.1%

Table 5. Accuracies for determining neutrals using optimized probabilities.

Our next steps are to improve the performance of the four binary emotion classifiers as well as the neutral classifiers, and then begin experimenting with the combined classifier configured as shown in Figure 3. Although our initial results for distinguishing neutrals are encouraging, we will need a larger test set with a greater proportion of emotionally neutral tweets to establish more definitive results.

6 Conclusion

The approach outlined in this paper shows that Plutchik's set of four pairs of opposing emotions provides a viable basis for developing binary emotion classifiers for Twitter data that can match or exceed results from previous studies. In

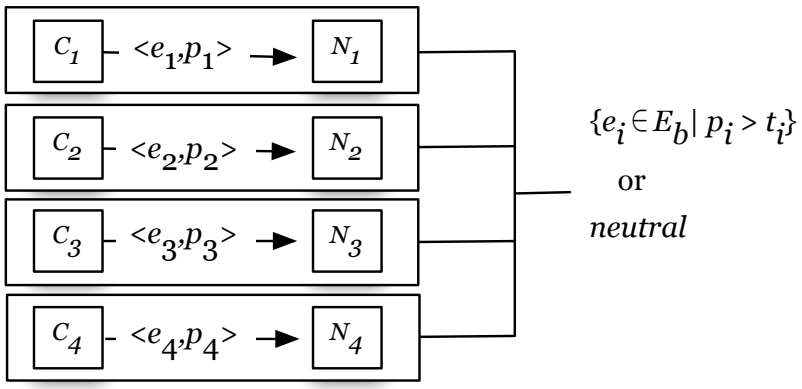

Fig. 3. Combined classifier that returns 1 to 4 emotion(s) from E_b (the set of binary emotions) that are present in a tweet, or *neutral* if the tweet has no emotion, with $e_i \in E_b$, the set of eight binary emotions; p_i the probability for e_i returned by classifier C_i; and t_i the optimal probability threshold for e_i; where $0 < i \leq 4$.

addition to emotions and hashtags, which have been used in similar work, we include emoji as emotional labels and show that they may be even more reliable emotion indicators than their pictorial cousins, emoticons. We have shown how the binary emotion classifiers can be combined to emulate a single multi-way classifier, thus avoiding the increased complexity (and corresponding weaker results) of multi-way classification; and how by further combining these classifiers with a combination of binary "neutral" classifiers, we not only emulate a multi-way emotion classifier but also isolate the particular emotions present in a given tweet. Our results on emotion label prediction suggest that our approach can produce reliable classifiers, and we therefore plan to attempt to improve on the work reported here by testing on much larger manually-annotated datasets and experimenting with the combined classifier described in Section 5.

References

1. Agarwal, A., Xie, B., Vovsha, I., Rambow, O., Passonneau, R.: Sentiment Analysis of Twitter Data. In: Proceedings of the Workshop on Language in Social Media (LSM 2011). pp. 30–38. Association for Computational Linguistics, Portland, Oregon (Jun 2011)
2. Alm, C.O., Roth, D., Sproat, R.: Emotions from text: Machine learning for text-based emotion prediction. In: Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing. pp. 579–586. Association for Computational Linguistics, Vancouver, British Columbia, Canada (October 2005)
3. Ansari, S.: Automatic emotion tone detection in twitter (2010)

4. Balabantaray, R.C., Mohammad, M., Sharma, N.: Multi-class twitter emotion classification: A new approach. International Journal of Applied Information Systems 4(1), 48–53 (2012)
5. Bollen, J., Pepe, A., Mao, H.: Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. CoRR (2009)
6. Choudhury, M.D., Counts, S., Gamon, M.: Not all moods are created equal! exploring human emotional states in social media. In: ICWSM (2012)
7. Chuang, Z.J., Wu, C.H.: Multi-modal emotion recognition from speech and text. International Journal of Computational Linguistics and Chinese Language Processing 9(2), 45–62 (2004)
8. Danisman, T., Alpkocak, A.: Emotion classification of audio signals using ensemble of support vector machines. In: PIT. pp. 205–216 (2008)
9. Derks, D., Bos, A.E.R., Von Grumbkow, J.: Emoticons and online message interpretation. Social Science Computer Review 26(3), 379–388 (2008)
10. Dodds, P.S., Harris, K.D., Kloumann, I.M., Bliss, C.A., Danforth, C.M.: Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. CoRR (2011)
11. Ekman, P.: Universals and cultural differences in facial expressions of emotions. Nebraska Symposium on Motivation 19, 207–283 (1972)
12. Ekman, P.: An argument for basic emotions. Cognition & Emotion 6(3-4), 169–200 (May 1992)
13. Go, A., Bhayani, R., Huang, L.: Twitter sentiment classification using distant supervision. Processing pp. 1–6 (2009)
14. González-Ibáñez, R., Muresan, S., Wacholder, N.: Identifying sarcasm in twitter: A closer look. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. pp. 581–586. Association for Computational Linguistics, Portland, Oregon, USA (June 2011)
15. Lee, C.M., Narayanan, S.S.: Toward detecting emotions in spoken dialogs. IEEE Transactions on Speech and Audio Processing pp. 293–303 (2005)
16. Mihalcea, R., Liu, H.: A corpus-based approach to finding happiness. In: AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs. pp. 139–144. AAAI (2006)
17. Mintz, M., Bills, S., Snow, R., Jurafsky, D.: Distant supervision for relation extraction without labeled data. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. pp. 1003–1011. Association for Computational Linguistics (2009)
18. 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. pp. 105–114. Association for Computational Linguistics, Portland, OR, USA (June 2011)
19. Mohammad, S.: #Emotional Tweets. In: *SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012). pp. 246–255. Association for Computational Linguistics, Montréal, Canada (7-8 June 2012)
20. Neviarouskaya, A., Prendinger, H., Ishizuka, M.: Analysis of affect expressed through the evolving language of online communication. In: Chin, D.N., Zhou, M.X., Lau, T.A., Puerta, A.R. (eds.) Proceedings of the 12th International Conference on Intelligent User Interfaces. pp. 278–281 (2009)

21. Pak, A., Paroubek, P.: Twitter as a corpus for sentiment analysis and opinion mining. In: Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., Tapias, D. (eds.) Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). European Language Resources Association (ELRA), Valletta, Malta (2010)
22. Plutchik, R.: Emotion: Theory, research, and experience. In: Theories of Emotion, vol. 1. Academic Press, New York, NY, USA (1980)
23. Provine, R., Spencer, R., Mandell, D.: Emotional Expression Online: Emoticons punctuate website text messages. Journal of Language and Social Psychology 26(3), 299–307 (2007)
24. Purver, M., Battersby, S.: 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. Association for Computational Linguistics, Avignon, France (April 2012)
25. Read, J.: Recognising Affect in Text using Pointwise-Mutual Information. Master's thesis, University of Sussex (2004)
26. Read, J.: Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In: Proceedings of the ACL Student Research Workshop. pp. 43–48. Association for Computational Linguistics, Stroudsburg, PA, USA (2005)
27. Roberts, K., Roach, M.A., Johnson, J., Guthrie, J., Harabagiu, S.M.: EmpaTweet: Annotating and Detecting Emotions on Twitter. In: Calzolari, N., Choukri, K., Declerck, T., Do?an, M.U., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S. (eds.) Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). European Language Resources Association (ELRA), Istanbul, Turkey (2012)
28. Seol, Y.S., Kim, D.J., Kim, H.W.: Emotion Recognition from Text Using Knowledge-based ANN. In: Proceedings of ITC-CSCC (2009)
29. Tanaka, Y., Takamura, H., Okumura, M.: Extraction and classification of facemarks. In: Proceedings of the 10th international conference on Intelligent user interfaces. pp. 28–34. IUI '05, ACM, New York, NY, USA (2005)
30. Wang, W., Chen, L., Thirunarayan, K., Sheth, A.: Harnessing Twitter Big Data for Automatic Emotion Identification. In: International Conference on Social Computing (2012)