2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust

Harnessing Twitter 'Big Data' for Automatic Emotion Identification

Wenbo Wang
Kno.e.sis Center
Wright State University
Dayton, OH 45435 USA
Email: wenbo@knoesis.org

Lu Chen
Kno.e.sis Center
Wright State University
Dayton, OH 45435 USA
Email: chen@knoesis.org

Krishnaprasad Thirunarayan
Kno.e.sis Center
Wright State University
Dayton, OH 45435 USA
Email: tkprasad@knoesis.org

Amit P. Sheth
Kno.e.sis Center
Wright State University
Dayton, OH 45435 USA
Email: amit@knoesis.org

***Abstract*—User generated content on Twitter (produced at an enormous rate of 340 million tweets per day) provides a rich source for gleaning people's emotions, which is necessary for deeper understanding of people's behaviors and actions. Extant studies on emotion identification lack comprehensive coverage of "emotional situations" because they use relatively small training datasets. To overcome this bottleneck, we have automatically created a large emotion-labeled dataset (of about 2.5 million tweets) by harnessing emotion-related hashtags available in the tweets. We have applied two different machine learning algorithms for emotion identification, to study the effectiveness of various feature combinations as well as the effect of the size of the training data on the emotion identification task. Our experiments demonstrate that a combination of unigrams, bigrams, sentiment/emotion-bearing words, and parts-of-speech information is most effective for gleaning emotions. The highest accuracy (65.57%) is achieved with a training data containing about 2 million tweets.**

I. Introduction

Emotion is both prevalent in and essential to all aspects of our lives. It influences our decision-making, affects our social relationships, shapes our daily behavior, even outlasts our memories. With the rapid growth of emotion-rich textual content, such as microblog posts, blog posts, and forum discussions, there is a great need and opportunity to develop automatic tools for identifying and analyzing people's emotions expressed in text.

Identifying the expressed emotions in text is very challenging for at least two reasons. First, emotions can be implicit and triggered by specific events or situations. Text describing an event or situation that causes the emotion can be devoid of explicit emotion-bearing words. Consider the examples in Table I: in Example 1, *fear* is inferred because of *"see a cop"*; in Example 2, *anger* is inferred because of *"mom compares me to my friends"*; and in Example 3, *embarrassment* is inferred because of *"hiccups in class"*. We can recognize *fear* emotion even though there is no explicit reference to words such as *"scare"* and *"panic"*. Second, gleaning distinction between different emotions purely on the basis of keywords can be very subtle. Examples 2 and 3 in Table I have similar sentence pattern and contain the same emotion-bearing word "hate", but they belong to different emotions (i.e., *anger* and *embarrassment*).

Most of current emotion identification research relies on manually annotated training data [1], [2]. Manual annotation of data by human experts is very labor-intensive and time-consuming. Moreover, in contrast with other annotation tasks such as entity or topic detection, a human annotator's judgement of emotion in text tends to be subjective and varied, and hence, less reliable. Consequently, most of existing emotion datasets are relatively small, of the order of thousands of entries, which fail to provide a comprehensive coverage of emotion-triggering events and situations.

TABLE I
Sample emotion texts

1	**Fear:** *"When I see a cop, no matter where I am or what I'm doing, I always feel like every law I've ever broken is stamped all over my body"*
2	**Anger:** *"I hate when my mom compares me to my friends."*
3	**Embarrassment:** *"I hate when I get the hiccups in class."*

While there is a lack of sufficient labeled data for emotion research, many social media services have entered the big data era. Twitter, the popular microblogging service, provides more than 340 million tweets per day[1] on a wide variety of topics, and a significant part of it is about "what is happening" in our daily lives expressed using emotion hashtags. For example, *"leaving for the hospital... #nervous"*, in the tweet the user "annotates" the tweet with hashtag *#nervous* to express *nervousness* emotion. **Can this Twitter 'big data' be harnessed to tackle the emotion identification problem?** Specifically, we address the following three questions:

1) Can we automatically create a large emotion dataset with high quality labels from Twitter by leveraging the emotion hashtags? If so, how?
2) What features can effectively improve the performance of supervised machine learning algorithms? Are they consistent with the findings of contemporary research?
3) Does big (millions instead of thousands) training data improve emotion identification accuracy? How much performance gain can be achieved by increasing the size of training data?

[1] http://blog.twitter.com/2012/03/twitter-turns-six.html

978-0-7695-4848-7/12 $26.00 © 2012 IEEE
DOI 10.1109/SocialCom-PASSAT.2012.119

To answer the above questions, we used 131 emotion hashtags as keywords and collected 5 million tweets for 7 emotion categories (*joy*, *sadness*, *anger*, *love*, *fear*, *thankfulness*, *surprise*) between Nov. 10th, 2011 and Dec. 22nd, 2011 (see Table II). To improve the quality of collected tweets, a set of heuristics were developed to retain relevant tweets, which contain the emotion hashtags that correctly annotate the expressed emotions. The evaluation of these heuristics shows that they can be used to create a high quality emotion dataset, in which 93.16% of the tweets are emotion-relevant. After applying these heuristics, we obtained an emotion tweet corpus containing about 2.5 million tweets[2]. To find the effective features for emotion identification, we explored a wide variety of features including n-grams, emotion lexicons, part-of-speech (POS), n-gram positions, etc. using two machine learning algorithms: LIBLINEAR [5] and Multinomial Naive Bayes (MNB) [18]. The best results were achieved by a combination of unigram, bigram, sentiment and emotion lexicon, and POS features. To investigate the contributions of large training data, we increased the size of training data from 1,000 to 2 million and achieved an absolute gain of 22.16% in accuracy.

II. Related Work

In this paper, we focus on supervised emotion identification literature and how to automatically collect training data rather than rule-based approaches [10], [15].

There are only a few efforts on supervised methods, partly due to the labor intensive nature of the manual labeling task. Alm et al. [1] present an empirical study of applying machine learning techniques to classify fairy tale sentences into different emotions. Aman and Szpakowicz [2] combine unigrams, emotion lexicons and a thesaurus as features to classify blog sentences into six basic emotion categories.

Automatically creating training data from weblogs has also been addressed. The approaches in [8], [13], [19] collect blog posts that have been assigned mood labels (e.g., amused, tired) by the blog writers. Tokuhisa et al. [16] collect 1.3 million sentences in Japanese by exploiting the sentence pattern *"I was ** that ..."*, in which "**" and "..." refer to an emotion word and the sentence reflecting the emotion, respectively. For example, *"I was disappointed that it suddenly started raining."*

Turning to harnessing the hashtag phenomenon on Twitter, Choudhury et al. [4] collect emotion tweets via emotion hashtags and analyze users' emotional states in social media through affective space (valence and activation). Both our work and [4] collect tweets via emotion hashtags, but identifying the writer's emotion from a tweet is not their focus. Mohammad [9] also collects emotion tweets via emotion hashtags. However our work uses a much larger dataset and a extensive list of features. To the best of our knowledge, the three questions we raised earlier are largely unexplored.

[2]The dataset is available for download from our project page http://knoesis.org/projects/emotion.

III. Collecting Labeled Emotion Tweets

In this section, we describe how we automatically created a labeled emotion dataset from Twitter. We first collected 7 sets of emotion words for 7 different emotions (e.g., word "annoying" for emotion *anger*) from existing psychology literature [12], and then utilized Twitter streaming API to collect tweets that have one of these emotion words in the form of a hashtag (e.g, #annoying). Each collected tweet was automatically labeled with one emotion according to its emotion hashtag, and the hashtag itself is removed from the tweet. For example, from an incoming tweet *"I hate when my mom compares me to my friends.**#annoying**"*, we obtain the following training example: *"I hate when my mom compares me to my friends."* labeled with *anger*, since it contains *"#annoying"* hashtag.

Our source of the emotion words is Shaver et. al.'s highly cited psychology paper [12], where the authors organize emotions into a hierarchy in which the first layer contains six basic emotions (i.e., *love*, *joy*, *surprise*, *anger*, *sadness* and *fear*), and the second layer contains 25 secondary emotions that are subcategories of the six basic emotions. Each secondary emotion has a list of emotion words. We expanded the list of emotion words by including their lexical variants, e.g., adding *"surprising"* and *"surprised"* for "surprise". In addition, we removed ambiguous words. For example, "glee" means *"great delight"* in the dictionary and is used to indicate the emotion *joy*, but it is also the name of a popular TV series in America. For each basic emotion, we also used the emotion words corresponding to its secondary emotions when collecting tweets. Besides the aforementioned six basic emotions, we added one more basic emotion, *thankfulness*, which is not covered by [12]. Table II shows the seven emotions, sample emotion hashtags, example tweets and the number of tweets in each category after filtering irrelevant ones.

Totally, we collected 5 million tweets. Before using these tweets as training examples, it is necessary to verify their quality, i.e., whether the emotion hashtags truly indicate the authors' emotional states. For this purpose, we randomly sampled a set of 400 tweets. Two annotators first independently annotated each tweet as relevant/irrelevant. A tweet was labeled as relevant if the emotion hashtag in the tweet reflects the writer's emotion. Otherwise, it was labeled as irrelevant. When there was a disagreement on the annotation of a tweet, the annotators collaborated to reach an agreement.

A set of filtering heuristics was developed on the aforementioned set of 400 tweets (development set). (1) We kept only the tweets with the emotion hashtags at the end. Based on our observation and corroborated by [4], if the emotion hashtag is not at the end of a tweet, it is less likely that the hashtag indicates the author's emotional state. (2) We discarded tweets which have less than five words, since they may not provide sufficient context to infer emotions. (3) We removed the tweets which contain URLs or quotations. A large amount of tweets with URLs are information-oriented, which do not convey emotions. Furthermore, we removed all the retweets, non-

TABLE II
EMOTION WORDS USED FOR COLLECTING TWEETS AND THE NUMBER OF COLLECTED TWEETS FOR EACH EMOTION (AFTER FILTERING)

Emotion	Hashtag Word Examples(#)	# of Tweets	Tweet Example
joy	excited, happy, elated, proud (36)	706,182	*"Omg I finally fit into one pair of my jeans from last year!! #excited"*
sadness	sorrow, unhappy, depressing, lonely (36)	616,471	*"im losing both of my semi-final games #depressing"*
anger	irritating, annoyed, frustrate, fury (23)	574,170	*"Ugh I have no money but payday tomorrow #Irritating"*
love	affection, lovin, loving, fondness (7)	301,759	*"iloveyou, just the way you are #love"*
fear	fear, panic, fright, worry, scare (22)	135,154	*"Calculus test today #studying #nervous"*
thankfulness	thankfulness, thankful (2)	131,340	*"The Maury show makes me realize my life isn't so bad. #thankful"*
surprise	surprised, astonished, unexpected (5)	23,906	*"Today's going a lot better than I thought on no sleep #surprised"*
TOTAL	**(131)**	**2,488,982**	

English tweets and tweets having more than 3 hashtags.

To evaluate the filtering heuristics, we randomly sampled another disjoint set of 400 tweets, annotated each tweet as relevant/irrelevant in the same manner as the first 400 tweets, and used it as the test dataset. we developed our heuristics on the development dataset and applied it to the test dataset. Since we are interested in creating high quality dataset, precision far outweighs recall. The precision on the development dataset were 95.08%, while the precision on the test dataset were 93.16%. Thus, our filtering heuristics were effective in removing irrelevant tweets. After applying the heuristics on all the collected tweets, we finally obtained a collection of 2,488,982 tweets. The distribution of tweets per emotion is summarized in Table II.

IV. EXPERIMENTAL SETUP

Datasets: Out of the 2,488,982 tweets in Table II, we randomly sampled 250,000 tweets as test dataset **Te**, reserved another randomly sampled 247,798 tweets as development dataset for tuning the algorithms, and used the remaining 1,991,184 tweets (denoted as **Tr**) for training. Note that the test, development and training datasets are disjoint. We divided **Tr** into eight subsets (denoted as **Tr1**, **Tr2**, ..., **Tr8**, respectively), each comprising 248,898 tweets. **Tr1** was used for exploring effective features, and all the eight subsets were used to examine the effect of increasing training data.

Data preprocessing: We lower-cased all the words; replaced user mentions (e.g., @ladygaga) with @user to anonymize users; replaced letters/punctuations that are repeated more than twice with the same two letters/punctuations (e.g., *cooool* $\rightarrow$ *cool*, *!!!!!* $\rightarrow$ *!!*); normalized some frequently used informal expressions (e.g., *ll* $\rightarrow$ *will*, *dnt* $\rightarrow$ *do not*); and stripped hash symbols (*#tomorrow* $\rightarrow$ *tomorrow*).

Machine learning classifiers: We selected LIBLINEAR [5] and Multinomial Naive Bayes (MNB) [18] to use, since they are very efficient even for handling millions of tweets. We employed Weka's implementations [7] for MNB. We used logistic regression branch for LIBLINEAR and default values for all parameters in both classifiers.

Evaluation metrics: The overall performance of individual classifier is measured by: $accuracy$=$\frac{\text{\# of correctly labeled tweets}}{\text{\# of all the tweets in the test dataset}}$. In the test dataset, let E be the set of tweets with emotion e, E' be the set of tweets which are classified as emotion e by the classifier, then we define the precision on emotion e as: $pre(e)$=$\frac{|E \cap E'|}{|E'|}$, the recall on e as: $rec(e)$=$\frac{|E \cap E'|}{|E|}$ and the F-measure on e as: $F\text{-}measure(e)$=$\frac{2*pre(e)*rec(e)}{pre(e)+rec(e)}$.

V. EXPLORING EFFECTIVE FEATURES

We explore features that are effective for emotion identification. We start with features that are known to be effective for text classification, especially, sentiment analysis. Since both sentiment and emotion are subjective, comparative study of the useful features for identifying them may provide better insights for emotion identification.

A. Features

N-gram: N-gram features are widely used in a variety of tasks, including emotion analysis [2], [16]. In our study, we experimented with unigrams (n=1), bigrams (n=2), trigrams (n=3) and their combinations. Punctuations (e.g., !!, ??) and emoticons (e.g., :P, <3, </3) were also included into the n-gram model. Neither stemming nor stop word eliminations was applied. We used only n-grams that appear in at least five different tweets. Like [11], we used a boolean feature for each n-gram, which is set to true if and only if the n-gram is present in the tweet.

N-gram Position: Similar to [11], we also hypothesize that the words located towards the end of a tweet are more important than other words, because people usually summarize or highlight their points in the end. For example, *"I hate it when stuff like that happens,.. ;/ thank god it worked out.<3 #thankful."*. Although *"hate"* appears in the first half of the tweet, the overall emotion is dominated by *"thank"* in the latter half. We encoded the position information into a feature by attaching a number (i.e, 1 or 2) to each n-gram to indicate whether it is in the first half or the second half of the tweet. For example, if a tweet has 10 unigrams, then the first 5 unigrams belong to the first half of the tweet and 1 is attached to them. We also experimented with dividing a tweet into three parts, but the results were not as good.

LIWC Dictionary: Linguistic Inquiry and Word Count[3] (LIWC) is a text analysis software which provides a dictionary covering about 4,500 words and word stems from more than 70 categories. We collected emotion words from the positive emotion category (408 words) and negative emotion category (499 words) in LIWC2007 dictionary. For each tweet, we counted the number of positive/negative words based on the

[3]http://www.liwc.net/

set of collected emotion words, and used the percentage of words that are positive and that are negative as features.

MPQA Lexicon: MPQA (Multi-Perspective Question Answering) Subjectivity Lexicon [17] provides prior polarities (i.e, positive, negative, neutral or both) of 8,211 words. In a similar way as obtaining LIWC features, we got the percentage of words with positive/negative polarity in a tweet as features using the MPQA lexicon.

WordNet-Affect: WordNet-Affect [14] is a lexical resource that provides a hierarchy of "affective domain labels", and 2874 synsets and 4787 words in WordNet are annotated with the emotion labels in the hierarchy. We collected the words from 32 direct subcategories of *positive-emotion* (e.g., joy, love, etc.), *negative-emotion* (e.g., anxiety, sadness etc.), *neutral-emotion* (e.g., apathy, etc.) and *ambiguous-emotion* (e.g., surprise, etc) defined in WordNet-Affect. For each tweet, we used 32 features, each of which represents the number of words from one of the 32 subcategories.

Part-of-Speech (POS): POS features have been proven effective in sentiment/emotion classification [3], [8]. We used LingPipe[4] for POS tagging, and trained the tagger on a POS annotated tweet corpus [6]. We calculated the percentage of words belonging to each POS[5] in a tweet as features.

Adjectives: In sentiment analysis, adjectives are usually considered as effective features since they can be good indicators of sentiment. Some research [11] shows that using adjectives alone produces results competitive with those obtained by using n-grams in sentiment classification of movie reviews. We want to verify whether this also holds for emotion identification.

B. Classification Results with Different Feature Combinations

Using the features described above, we trained LIBLINEAR and MNB classifiers on **Tr1**, and applied the classifiers to the test dataset **Te**. Table III shows the classification accuracy achieved by different feature combinations. Note that it is a multi-class classification and each tweet is classified into one of the seven emotion categories.

Adjectives: In one early sentiment classification study [11], Naive Bayes classifier is reported to achieve an accuracy of 77% using only adjective features, which is very close to the performance of using bigrams (77.3%), and not far from the accuracy resulting from using unigrams (81%). But the situation is different for emotion identification. Line 1 in Table III shows that the accuracy obtained by using only adjectives as features (34.74% and 35.03% with MNB and LIBLINEAR classifiers, respectively) is about 60% of that using unigrams (57.75% and 60.31% with MNB and LIBLINEAR classifiers, respectively). This suggests that emotions are expressed more implicitly compared to sentiments, and accurate results cannot be obtained with only emotion or sentiment bearing adjectives, but require situational information. Recall that the Example 1 in Table I in which the writer expresses *fear* emotion without explicitly saying "I am scared."

[4]http://alias-i.com/lingpipe/

[5]Refer to Table 1 in paper [6] for a complete list of POS tags

Unigram vs. Bigram vs. Trigram: For n-gram models, as we increase n (n=1,2,3), higher order n-grams are expected to better capture the contextual information. However, does this lead to better results? As shown in line 3-5 in Table III, the best accuracy of MNB classifier is achieved using bigram features and is 58.53%, followed by 57.75% with unigrams and 51.86% with trigrams. For LIBLINEAR classifier, the accuracy of using unigrams (60.31%) is better than that of using any n-grams (n=1,2,3) with MNB classifier, but as we increase the value of n, the accuracies keep decreasing to 57.68% for bigrams and 50.65% for trigrams.

In addition, we also experimented with combined n-gram features. Although unigrams, bigrams and trigrams are not conditionally independent of each other(, which violates the conditional-independence assumptions made by MNB), the MNB classifiers using the combinations beat the ones that use only unigrams, bigrams or trigrams. Specifically, combining unigrams and bigrams (line 5) increases the accuracy to 61.13%, and further incorporation of trigrams (line 6) decreases the accuracy slightly to 60.96%, which is still better than the accuracy for using one of them alone. We observed a similar pattern for LIBLINEAR classifiers. The best accuracy of 61.56% is obtained by LIBLINEAR classifiers using unigrams and bigrams, which is slightly better than the best accuracy achieved by MNB classifiers (61.13%).

Our experimental results show that combining unigrams and bigrams yields better performance than using unigrams alone. This is different from existing discoveries on sentiment classification [11], where using unigrams alone is better than applying either bigrams or a combination of unigrams and bigrams. It might suggest that bigrams are effective at capturing contextual information in our setting, leading to performance gain. Trigrams do not show such effectiveness. The previous research on emotion classification does not provide comparative study of different orders of n-grams. Aman and Szpakowicz [2] use only unigram features, and Tokuhisa et al. [16] use a combination of unigrams, bigrams and trigrams for classification, but without comparing it with the classifiers that use unigrams, bigrams or trigrams alone or other combinations.

N-gram Position: Contrary to our intuitions, the accuracy of both the classifiers (line 7) deteriorates with position information. In fact, injecting position information into n-grams is also found to decrease the performance in sentiment classification [11].

Sentiment/Emotion Lexicons and Part-of-Speech: Adding lexicon-based features and part-of-speech features does not greatly improve the performance (see lines 8-11 in Table III) in our setting of emotion identification. However, sentiment lexicon and part-of-speech have been used and shown to be effective in sentiment analysis. It further indicates that emotions are expressed more implicitly and subtly than sentiments. It is likely that people tend to use sentiment words (positive or negative) when they talk about their likes and dislikes, but may not verbalize and share their emotions through specific emotion-bearing words.

TABLE III
ACCURACIES OF MNB AND LIBLINEAR ON TR1 DATASET WITH DIFFERENT FEATURE SETS: BOOLEAN VALUE (PRESENCE) IS USED FOR ALL N-GRAM FEATURES; PERCENTAGES WERE USED FOR LIWC, MPQA AND POS FEATURES; FREQUENCY (COUNTS) WERE USED FOR WORDNET-AFFECT FEATURE

#	Features	Accuracy(%)	
		MNB	LIBLINEAR
1	adjective	34.74	35.03
2	n-gram(n=1)	57.75	60.31
3	n-gram(n=2)	58.53	57.68
4	n-gram(n=3)	51.86	50.65
5	n-gram(n=1,2)	61.13	61.56
6	n-gram(n=1,2,3)	60.96	61.55
7	n-gram(n=1,2),n-gram position	60.40	60.76
8	n-gram(n=1,2),LIWC	61.13	61.59
9	n-gram(n=1,2),MPQA	**61.15**	61.57
10	n-gram(n=1,2),WordNet-Affect	**61.15**	61.57
11	n-gram(n=1,2),POS	61.12	61.62
12	n-gram(n=1,2),LIWC,MPQA, WordNet-Affect,POS	**61.15**	**61.63**

VI. EFFECT OF INCREASING THE TRAINING DATA

We examine the effect of increasing the size of training dataset on the accuracy of LIBLINEAR and MNB classifiers. Since most extant emotion identification [2] is conducted on datasets of thousands of sentences, we expect to derive new insights and benefits of using large training data.

We started with thousands of tweets. We randomly sampled a set of 1,000 tweets, and another disjoint set of 9,000 tweets from Tr1, denoting them as Tr11 and Tr12. Finally we created a sequence of datasets with increasing sizes: Tr11, Tr11∪Tr12, Tr1, Tr1∪Tr2, Tr1∪Tr2∪Tr3, ..., Tr1∪Tr2∪...∪Tr8, in which each smaller dataset is contained in the larger following datasets (see Section IV).

We trained LIBLINEAR and MNB classifiers on each dataset in the sequence with unigram and bigram features[6]. Figure 1 shows the accuracies of applying the classifiers on test dataset Te.

Benefits of Increasing the Training Data: From Figure 1 we observe that as the training data increases from 1,000 to about 2 million, we get an absolute accuracy gain of (65.57%-43.41%=) 22.16% with LIBLINEAR classifiers. Specifically, accuracy grows by $(52.92\% - 43.41\% =)$ 9.51%, $(61.56\% - 52.92\% =)$ 8.64%, and $(65.57\% - 61.56\% =)$ 4.01% when the training data increases from 1,000 to 10,000 tweets, 10,000 to about 250K tweets, and 250K to about 2M tweets, respectively. This result demonstrates that learning from large training data can play an important role in emotion identification.

Table IV shows the performance of LIBLINEAR classifier (trained with all tweets in Tr) on each emotion category. We made the following discoveries. For the three most popular emotions – *joy*, *sadness* and *anger*, which account for 76.1% of all tweets, the classifier achieves precisions of over 62%, recalls of over 66%, and F-measures of over 64% for each of the three emotions. Performance declines can be seen on

[6]We also experimented with a few other feature combinations (line 6 and 12 in Table III), but the results do not differ much from using the combination of unigrams and bigrams.

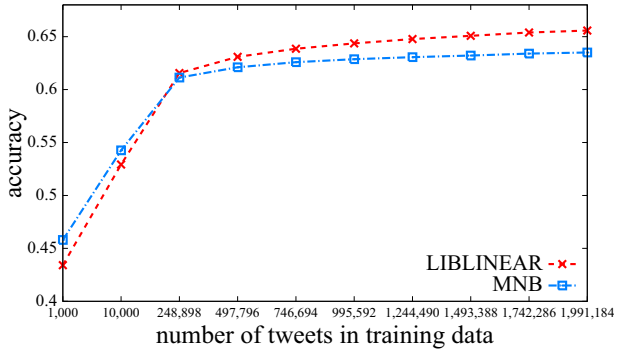

Fig. 1. Accuracies of LIBNEAR and MNB with varied sizes of training data

TABLE IV
DETAILED RESULT OF LIBLINEAR WITH THE LARGEST TRAINING DATA

Emotion	Precision(%)	Recall(%)	F-measure(%)
joy (28.5%)	67.6	**77.3**	**72.1**
sadness (24.6%)	62.6	66.8	64.7
anger (23.0%)	**69.8**	73.3	71.5
love (12.1%)	58.1	46.2	51.5
fear (5.6%)	59.7	34.7	43.9
thankfulness (5.3%)	66.6	50.0	57.1
surprise (1.0%)	44.7	8.2	13.9

three less popular emotions (i.e., *love*, *fear* and *thankfulness*), which consist of 23.0% of all the tweets in our dataset. The precisions of these three emotion categories are relatively high (with lowest precision of 58.1%) compared with the recalls, but because of the low recalls, the classifier achieves F-measures of only 51.5%, 43.9%, and 57.1% for *love*, *fear* and *thankfulness* categories, respectively. For the remaining minority emotion (i.e., *surprise*, with only 1.0% of all tweets), the classifier gets the lowest precision, recall, and F-measure because of the heavily imbalanced emotion distribution in the training data.

To get more insight from the result, we analyzed the confusion matrix to figure out what are the top misjudged cases. Out of the 86,071 misjudged tweets, there are 20,799 (24.2%) tweets misclassified between *sadness* and *anger* (i.e., either *sadness* tweets were misclassified as *anger* or vice versa), 13,400 (15.6%) tweets were misclassified between *love* and *joy*, and 11,709 (13.6%) tweets were misclassified between *joy* and *sadness*. This is in line with the fact that some emotion pairs (*anger* and *sadness*, *joy* and *love*) are naturally related to each other. Moreover, different people might have different emotions when facing similar events. For example, *"My phone bout to die too..uggghhhhh #annoyed"* vs. *"It's dark so I can't read, my phone is about to die so no music. #sad"*. It is interesting to see that emotions with opposite polarities, e.g., *joy* and *sadness*, may be expressed in the same tweet. E.g., *"i hate myself for being such a procrastinator. midnight shower then five hours of sleep. #joy"*.

VII. Discussion

We share our observations and experiences gleaning emotion from Twitter data.

Compared with manually annotated emotion datasets, the corpus of automatically collected tweets with emotion labels (i.e., hashtags) shows its advantages in several aspects. (i) The emotion hashtags of tweets are provided by their writers, which are more natural and reliable than the emotion labels of other datasets given by a few annotators. This is because writers are accurate about their own emotions, while the traditional way of annotating data requires annotators to infer the writers' emotions from text, which may not be accurate. (ii) It is very labor-intensive and time-consuming to manually annotate the data, which greatly limits the size of training datasets. Given that we collected 2.5 million tweets with high quality emotion hashtags in six weeks, over 20 million such tweets can be collected from Twitter per year. As we have shown, larger training data will lead to higher accuracy for emotion identification because it can provide a comprehensive coverage of emotional moments in our daily lives.

Further improvements can be made to improve the quality of the collected Twitter data along the following lines. (i) We cannot manually verify all the emotion tweets because of its large scale. Although we have removed some irrelevant tweets using heuristics, there are extraneously labeled tweets. In our case, after the filtering process, 93.16% of tweets are relevant on a test bed of 400 tweets (Section III), which can be improved using additional heuristics. (ii) Our dataset does not contain tweets with neutral labels, which is a common problem faced by automatically collecting training data for emotion analysis [16], [19]. (iii) The distribution of emotions in the Twitter dataset is imbalanced, e.g., only 1% of the tweets belong to *surprise*, and the classifiers do not perform well on less popular emotions. To further improve the performance, one possible way is to increase the amount of tweets for minority emotions by using more of their hashtags to collect tweets. In addition, undersampling tweets from majority emotions might provide an alternative approach.

VIII. Conclusion

We culled 2.5 million emotion tweets covering 7 emotion categories for automatic emotion identification. This is one of the largest datasets for automatic emotion identification. The experimental results show that the feature combination of unigrams, bigrams, existing sentiment and emotion lexicons, and part-of-speech achieves the best accuracy, although lexicon-based and part-of-speech features become less effective in identifying fine-grained emotions than in sentiment analysis. We achieved the highest accuracy of 65.57% with a training data containing about 2 million tweets.

As future work, we will explore how to automatically collect neutral tweets so that the system can support the detection of text without emotions. We plan to add more emotion hashtags to increase the number of tweets for less popular emotions (especially, *surprise* and *fear*), thereby reducing the imbalance in the dataset. We are developing transfer learning techniques so that our dataset can improve emotion identification in other domains (e.g., blog).

Acknowledgment

This research was supported by US National Science Foundation grant IIS-1111182: SoCS: Social Media Enhanced Organizational Sensemaking in Emergency Response. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] C. Alm, D. Roth, and R. Sproat, "Emotions from text: machine learning for text-based emotion prediction," in *Proceedings of HLT and EMNLP*. ACL, 2005, pp. 579–586.

[2] S. Aman and S. Szpakowicz, "Using roget's thesaurus for fine-grained emotion recognition," in *Proceedings of IJCNLP*, 2008, pp. 296–302.

[3] P. Chesley, B. Vincent, L. Xu, and R. K. Srihari, "Using verbs and adjectives to automatically classify blog sentiment," in *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, 2006, pp. 27–29.

[4] M. D. Choudhury, S. Counts, and M. Gamon, "Not all moods are created equal! exploring human emotional states in social media," in *Proceedings of ICWSM*, 2012.

[5] R. Fan, K. Chang, C. Hsieh, X. Wang, and C. Lin, "Liblinear: A library for large linear classification," *The Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[6] K. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, and N. A. Smith, "Part-of-speech tagging for twitter: annotation, features, and experiments," in *Proceedings of HLT:short papers*, ser. HLT '11. Stroudsburg, PA, USA: ACL, 2011, pp. 42–47.

[7] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten, "The weka data mining software: an update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[8] G. Mishne, "Experiments with mood classification in blog posts," in *Proceedings of ACM SIGIR 2005 Workshop on Stylistic Analysis of Text for Information Access*.

[9] S. Mohammad, "#emotional tweets," in *Proceedings of the Sixth International Workshop on Semantic Evaluation*. ACL, 7-8 June 2012, pp. 246–255.

[10] A. Neviarouskaya, H. Prendinger, and M. Ishizuka, "Affect analysis model: Novel rule-based approach to affect sensing from text," *Natural Language Engineering*, vol. 17, no. 1, pp. 95–135, 2011.

[11] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in *Proceedings of EMNLP*. ACL, 2002, pp. 79–86.

[12] P. Shaver, J. Schwartz, D. Kirson, and C. O'connor, "Emotion knowledge: Further exploration of a prototype approach." *Journal of personality and social psychology*, vol. 52, no. 6, pp. 1061–1086, 1987.

[13] C. Strapparava and R. Mihalcea, "Learning to identify emotions in text," in *Proceedings of the 2008 ACM symposium on Applied computing*. ACM, 2008, pp. 1556–1560.

[14] C. Strapparava and A. Valitutti, "Wordnet-affect: an affective extension of wordnet," in *Proceedings of LREC*, vol. 4. Citeseer, 2004, pp. 1083–1086.

[15] C. Strapparava and R. Mihalcea, "Semeval-2007 task 14: affective text," in *Proceedings of the 4th International Workshop on Semantic Evaluations*, ser. SemEval '07, 2007, pp. 70–74.

[16] R. Tokuhisa, K. Inui, and Y. Matsumoto, "Emotion classification using massive examples extracted from the web," in *Proceedings of COLING*. ACL, 2008, pp. 881–888.

[17] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in *Proceedings of HLT and EMNLP*. ACL, 2005, pp. 347–354.

[18] I. Witten, E. Frank, and M. Hall, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2011.

[19] C. Yang, K. Lin, and H. Chen, "Emotion classification using web blog corpora," in *IEEE/WIC/ACM International Conference on Web Intelligence*. IEEE, 2007, pp. 275–278.