#Emotional Tweets

Saif M. Mohammad
Emerging Technologies
National Research Council Canada
Ottawa, Ontario, Canada K1A 0R6
saif.mohammad@nrc-cnrc.gc.ca

Abstract

Detecting emotions in microblogs and social media posts has applications for industry, health, and security. However, there exists no microblog corpus with instances labeled for emotions for developing supervised systems. In this paper, we describe how we created such a corpus from Twitter posts using emotion-word hashtags. We conduct experiments to show that the self-labeled hashtag annotations are consistent and match with the annotations of trained judges. We also show how the Twitter emotion corpus can be used to improve emotion classification accuracy in a different domain. Finally, we extract a word–emotion association lexicon from this Twitter corpus, and show that it leads to significantly better results than the manually crafted WordNet Affect lexicon in an emotion classification task.[1]

1 Introduction

We use language not just to convey facts, but also our emotions. Automatically identifying emotions expressed in text has a number of applications, including customer relation management (Bougie et al., 2003), determining popularity of products and governments (Mohammad and Yang, 2011), and improving human-computer interaction (Velásquez, 1997; Ravaja et al., 2006).

Twitter is an online social networking and microblogging service where users post and read messages that are up to 140 characters long. The messages are called *tweets*.

[1]Email the author to obtain a copy of the hash-tagged tweets or the emotion lexicon: saif.mohammad@nrc-cnrc.gc.ca.

Often a tweet may include one or more words immediately preceded with a hash symbol (#). These words are called *hashtags*. Hashtags serve many purposes, but most notably they are used to indicate the topic. Often these words add to the information in the tweet: for example, hashtags indicating the tone of the message or their internal emotions.

From the perspective of one consuming tweets, hashtags play a role in search: Twitter allows people to search tweets not only through words in the tweets, but also through hashtagged words. Consider the tweet below:

> *We are fighting for the 99% that have been left behind. #OWS #anger*

A number of people tweeting about the Occupy Wall Street movement added the hashtag *#OWS* to their tweets. This allowed people searching for tweets about the movement to access them simply by searching for the *#OWS* hashtag. In this particular instance, the tweeter (one who tweets) has also added an emotion-word hashtag *#anger*, possibly to convey that he or she is angry.

Currently there are more than 200 million Twitter accounts, 180 thousand tweets posted every day, and 18 thousand Twitter search queries every second. Socio-linguistic researchers point out that Twitter is primarily a means for people to converse with other individuals, groups, and the world in general (Boyd et al., 2010). As tweets are freely accessible to all, the conversations can take on non-traditional forms such as discussions developing through many voices rather than just two interlocuters. For example, the use of Twitter and Facebook has been credited with

*First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 246–255,
Montréal, Canada, June 7-8, 2012.

providing momentum to the 2011 Arab Spring and Occupy Wall Street movements (Skinner, 2011; Ray, 2011). Understanding how such conversations develop, how people influence one another through emotional expressions, and how news is shared to elicit certain emotional reactions, are just some of the compelling reasons to develop better models for the emotion analysis of social media.

Supervised methods for emotion detection tend to perform better than unsupervised ones. They use ngram features such as unigrams and bigrams (individual words and two-word sequences) (Aman and Szpakowicz, 2007; Neviarouskaya et al., 2009; Mohammad, 2012b). However, these methods require labeled data where utterances are marked with the emotion they express. Manual annotation is time-intensive and costly. Thus only a small amount of such text exists. Further, supervised algorithms that rely on ngram features tend to classify accurately only if trained on data from the same domain as the target sentences (Mohammad, 2012b). Thus even the limited amount of existing emotion-labeled data is unsuitable for use in microblog analysis.

In this paper, we show how we automatically created a large dataset of more than 20,000 emotion-labeled tweets using hashtags. We compiled labeled data for six emotions—joy, sadness, anger, fear, disgust, and surprise—argued to be the most basic (Ekman, 1992). We will refer to our dataset as the Twitter Emotion Corpus (TEC). We show through experiments that even though the tweets and hashtags cover a diverse array of topics and were generated by thousands of different individuals (possibly with very different educational and socio-economic backgrounds), the emotion annotations are consistent and match the intuitions of trained judges. We also show how we used the TEC to improve emotion detection in a domain very different from social media.

Finally, we describe how we generated a large lexicon of ngrams and associated emotions from TEC. This emotion lexicon can be used in many applications, including highlighting words and phrases in a piece of text to quickly convey regions of affect. We show that the lexicon leads to significantly better results than that obtained using the manually crafted WordNet Affect lexicon in an emotion classification task.

2 Related Work

Emotion analysis can be applied to all kinds of text, but certain domains and modes of communication tend to have more overt expressions of emotions than others. Genereux and Evans (2006), Mihalcea and Liu (2006), and Neviarouskaya et al. (2009) analyzed web-logs. Alm et al. (2005) and Francisco and Gervás (2006) worked on fairy tales. Boucouvalas (2002), John et al. (2006), and Mohammad (2012a) explored emotions in novels. Zhe and Boucouvalas (2002), Holzman and Pottenger (2003), and Ma et al. (2005) annotated chat messages for emotions. Liu et al. (2003) and Mohammad and Yang (2011) worked on email data. Kim et al. (2009) analyzed sadness in posts reacting to news of Michael Jackson's death. Tumasjan et al. (2010) study Twitter as a forum for political deliberation.

Much of this work focuses on six Ekman emotions. There is less work on complex emotions, for example, work by Pearl and Steyvers (2010) that focuses on politeness, rudeness, embarrassment, formality, persuasion, deception, confidence, and disbelief. Bolen et al. (2009) measured tension, depression, anger, vigor, fatigue, and confusion in tweets. One of the advantages of our work is that we can easily collect tweets with hashtags for many emotions, well beyond the basic six.

Go et al. (2009) and González-Ibáñez et al. (2011) noted that sometimes people use the hashtag *#sarcasm* to indicate that their tweet is sarcastic. They collected tweets with hashtags of *#sarcasm* and *#sarcastic* to create a dataset of sarcastic tweets. We follow their ideas and collect tweets with hashtags pertaining to different emotions. Additionally, we present several experiments to validate that the emotion labels in the corpus are consistent and match intuitions of trained judges.

3 Existing Emotion-Labeled Text

The SemEval-2007 Affective Text corpus has newspaper headlines labeled with the six Ekman emotions by six annotators (Strapparava and Mihalcea, 2007). More precisely, for each headline–emotion pair, the annotators gave scores from 0 to 100 indicating how strongly the headline expressed the emotion. The inter-annotator agreement as determined by calculating the Pearson's product moment corre-

emotion	# of instances	% of instances	r
anger	132	13.2	0.50
disgust	43	4.3	0.45
fear	247	24.7	0.64
joy	344	34.4	0.60
sadness	283	28.3	0.68
surprise	253	25.3	0.36
		simple average	0.54
		frequency-based average	0.43

Table 1: Inter-annotator agreement (Pearson's correlation) amongst 6 annotators on the 1000-headlines dataset.

lation (r) between the scores given by each annotator and the average of the other five annotators is shown in Table 1. For our experiments, we considered scores greater than 25 to indicate that the headline expresses the corresponding emotion.

The dataset was created for an unsupervised competition, and consisted of 250 headlines of trial data and 1000 headlines of test data. We will refer to them as the 250-headlines and the 1000-headlines datasets respectively. However, the data has also been used in a supervised setting through (1) ten-fold cross-validation on the 1000-headlines dataset and (2) using the 1000 headlines as training data and testing on the 250-headlines dataset (Chaffar and Inkpen, 2011).

Other datasets with sentence-level annotations of emotions include about 4000 sentences from blogs, compiled by Aman and Szpakowicz (2007); 1000 sentences from stories on topics such as education and health, compiled by Neviarouskaya et al. (2009); and about 4000 sentences from fairy tales, annotated by Alm and Sproat (2005).

4 Creating the Twitter Emotion Corpus

Sometimes people use hashtags to notify others of the emotions associated with the message they are tweeting. Table 2 shows a few examples. On reading just the message before the hashtags, most people will agree that the tweeter #1 is sad, tweeter #2 is happy, and tweeter #3 is angry.

However, there also exist tweets such as the fourth example, where reading just the message before the hashtag does not convey the emotions of the tweeter. Here, the hashtag provides information not present (implicitly or explicitly) in the rest of the message.

1. *Feeling left out... #sadness*
2. *My amazing memory saves the day again! #joy*
3. *Some jerk stole my photo on tumblr. #anger*
4. *Mika used my photo on tumblr. #anger*
5. *School is very boring today :/ #joy*
6. *to me.... YOU are ur only #fear*

Table 2: Example tweets with emotion-words hashtags.

There are also tweets, such as those shown in examples 5 and 6, that do not seem to express the emotions stated in the hashtags. This may occur for many reasons including the use of sarcasm or irony. Additional context is required to understand the full emotional import of many tweets. Tweets tend to be very short, and often have spelling mistakes, short forms, and various other properties that make such text difficult to process by natural language systems. Further, it is probable, that only a small portion of emotional tweets are hashtagged with emotion words.

Our goal in this paper is to determine if we can successfully use emotion-word hashtags as emotion labels despite the many challenges outlined above:

- Can we create a large corpus of emotion-labeled hashtags?
- Are the emotion annotations consistent, despite the large number of annotators, despite no control over their socio-economic and cultural background, despite the many ways in which hashtags are used, and despite the many idiosyncracies of tweets?
- Do the hashtag annotations match with the intuitions of trained judges?

We chose to collect tweets with hashtags corresponding to the six Ekman emotions: *#anger, #disgust, #fear, #happy, #sadness,* and *#surprise*.

Eisenstein et al. (2010) collected about 380,000 tweets[2] from Twitter's official API.[3] Similarly, Go et al. (2009) collected 1.6 million tweets.[4] However, these datasets had less than 50 tweets that contained emotion-word hashtags. Therefore, we abandoned the search-in-corpora approach in favor of the one described below.

[2]http://www.ark.cs.cmu.edu/GeoText

[3]https://dev.twitter.com/docs/streaming-api

[4]https://sites.google.com/site/twittersentimenthelp

4.1 Hashtag-based Search on the Twitter Search API

The Archivist[5] is a free online service that helps users extract tweets using Twitter's Search API.[6] For any given query, Archivist first obtains up to 1500 tweets from the previous seven days. Subsequently, it polls the Twitter Search API every few hours to obtain newer tweets that match the query. We supplied Archivist with the six hashtag queries corresponding to the Ekman emotions, and collected about 50,000 tweets from those posted between November 15, 2011 and December 6, 2011.

We discarded tweets that had fewer than three valid English words. We used the *Roget Thesaurus* as the lexicon of English words.[7] This helped filter out most, if not all, of the non-English tweets that had English emotion hashtags. It also eliminated very short phrases, and some expressions with very bad spelling. We discarded tweets with the prefix "Rt", "RT", and "rt", which indicate that the messages that follow are re-tweets (re-postings of tweets sent earlier by somebody else). Like González-Ibáñez et al. (2011), we removed tweets that did not have the hashtag of interest at the end of the message. It has been suggested that middle-of-tweet hashtags may not be good labels of the tweets.[8] Finally, we were left with about 21,000 tweets, which formed the Twitter Emotion Corpus (TEC).

4.2 Distribution of emotion-word hashtags

Table 3 presents some details of the TEC. Observe that the distribution of emotions in the TEC is very different from the distribution of emotions in the 1000-headlines corpus (see Table 1). There are more messages tagged with the hashtag *#joy* than any of the other Ekman emotions.

Synonyms can often be used to express the same concept or emotion. Thus it is possible that the true distribution of hashtags corresponding to emotions is different from what is shown in Table 3. In the future, we intend to collect tweets with synonyms of *joy, sadness, fear,* etc., as well.

[5]http://archivist.visitmix.com

[6]https://dev.twitter.com/docs/using-search

[7]Roget's Thesaurus: www.gutenberg.org/ebooks/10681

[8]End-of-message hashtags are also much more common than hashtags at other positions.

hashtag	# of instances	% of instances
#anger	1,555	7.4
#disgust	761	3.6
#fear	2,816	13.4
#joy	8,240	39.1
#sadness	3,830	18.2
#surprise	3,849	18.3
Total tweets	21,051	100.0
# of tweeters	19,059	

Table 3: Details of the Twitter Emotion Corpus.

5 Consistency and Usefulness of Emotion Hashtagged Tweets

As noted earlier, even with trained judges, emotion annotation obtains only a modest inter-annotator agreement (see Table 1). As shown in Table 3, the TEC has about 21,000 tweets from about 19,000 different people. If TEC were to be treated as manually annotated data (which in one sense, it is), then it is data created by a very large number of judges, and most judges have annotated just one instance. Therefore, an important question is to determine whether the hashtag annotations of the tens of thousands of tweeters are consistent with one another. It will also be worth determining if this large amount of emotion-tagged Twitter data can help improve emotion detection in sentences from other domains.

To answer these questions, we conducted two automatic emotion classification experiments described in the two sub-sections below. For these experiments, we created binary classifiers for each of the six emotions using Weka (Hall et al., 2009).[9] For example, the *Fear–NotFear* classifier determined whether a sentence expressed fear or not. Note that, for these experiments, we treated the emotion hashtags as class labels and removed them from the tweets. Thus a classifier has to determine that a tweet expresses anger, for example, without having access to the hashtag *#anger*.

We chose Support Vector Machines (SVM) with Sequential Minimal Optimization (Platt, 1999) as the machine learning algorithm because of its successful application in various research problems. We used binary features that captured the presence or absence of unigrams and bigrams.

[9]http://www.cs.waikato.ac.nz/ml/weka

Label (X)	#gold	#right	#guesses	P	R	F
I. System using ngrams with freq. > 1						
anger	132	35	71	49.3	26.5	34.5
disgust	43	8	19	42.1	18.6	25.8
fear	247	108	170	63.5	43.7	51.8
joy	344	155	287	54.0	45.1	49.1
sadness	283	104	198	52.5	36.7	43.2
surprise	253	74	167	44.3	29.2	35.2
ALL LABELS	1302	484	912	53.1	37.2	**43.7**
II. System using all ngrams (no filtering)						
ALL LABELS	1302	371	546	67.9	28.5	40.1
III. System that guesses randomly						
ALL LABELS	1302	651	3000	21.7	50.0	30.3

Table 4: Cross-validation results on the 1000-headlines dataset. *#gold* is the number of headlines expressing a particular emotion. *#right* is the number these instances the classifier correctly marked as expressing the emotion. *#guesses* is the number of instances marked as expressing an emotion by the classifier.

In order to set a suitable benchmark for experiments with the TEC corpus, we first applied the classifiers to the SemEval-2007 Affective Text corpus. We executed ten-fold cross-validation on the 1000-headlines dataset. We experimented with using all ngrams, as well as training on only those ngrams that occurred more than once.

The rows under I in Table 4 give a breakdown of results obtained by the *EmotionX–NotEmotionX* classifiers. when they ignored single-occurrence ngrams (where X is one of the six basic emotions). *#gold* is the number of headlines expressing a particular emotion X. *#right* is the number of instances that the classifier correctly marked as expressing X. *#guesses* is the number of instances marked as expressing X by the classifier. Precision (P) and recall (R) are calculated as shown below:

$$P = \frac{\#right}{\#guesses} * 100 \quad (1)$$

$$R = \frac{\#right}{\#gold} * 100 \quad (2)$$

F is the balanced F-score. The ALL LABELS row shows the sums of *#gold*, *#right*, and *#guesses*.

The II and III rows in the table show overall results obtained by a system that uses all ngrams and by a system that guesses randomly.[10] We do not show a breakdown of results by emotions for II and III due to space constraints.

It is not surprising that the emotion classes with the most training instances and the highest inter-annotator agreement (joy, sadness, and fear) are also the classes on which the classifiers perform best (see Table 1).

The F-score of 40.1 obtained using all ngrams is close to 39.6 obtained by Chaffar and Inkpen (2011)—a sanity check for our baseline system. Ignoring words that occur only once in the training data seems beneficial. All classification results shown ahead are for the cases when ngrams that occurred only once were filtered out.

5.1 Experiment I: Can a classifier learn to predict emotion hashtags?

We applied the binary classifiers described above to the TEC. Table 5 shows ten-fold cross-validation results. Observe that even though the TEC was created from tens of thousands of users, the automatic classifiers are able to predict the emotions (hashtags) with F-scores much higher than the random baseline, and also higher than those obtained on the 1000-headlines corpus. Note also that this is despite the fact that the random baseline for the 1000-headlines corpus ($F = 30.3$) is higher than the random baseline for the TEC ($F = 21.7$). The results suggest that emotion hashtags assigned to tweets are consistent to a degree such that they can be used for detecting emotion hashtags in other tweets.

Note that expectedly the *Joy–NotJoy* classifier

[10] A system that randomly guesses whether an instance is expressing an emotion X or not will get half of the *#gold* instances right. Further, the system will mark half of all the instances as expressing emotion X. For ALL LABELS, $\#right = \frac{\#gold}{2}$, and $\#guesses = \frac{\#instances*6}{2}$.

Label	#gold	#right	#guesses	P	R	F
I. System using ngrams with freq. > 1						
anger	1555	347	931	37.3	22.31	27.9
disgust	761	102	332	30.7	13.4	18.7
fear	2816	1236	2073	59.6	43.9	50.6
joy	8240	4980	7715	64.5	60.4	62.4
sadness	3830	1377	3286	41.9	36.0	38.7
surprise	3849	1559	3083	50.6	40.5	45.0
ALL LABELS	21051	9601	17420	55.1	45.6	**49.9**
II. System that guesses randomly						
ALL LABELS	21051	10525	63,153	16.7	50.0	21.7

Table 5: Cross-validation results on the TEC. The highest F-score is shown in bold.

gets the best results as it has the highest number of training instances. The *Sadness–NotSadness* classifier performed relatively poorly considering the amount of training instances available, whereas the *Fear-NotFear* classifier performed relatively well. It is possible that people use less overt cues in tweets when they are explicitly giving it a sadness hashtag.

5.2 Experiment II: Can TEC improve emotion classification in a new domain?

As mentioned earlier, supervised algorithms perform well when training and test data are from the same domain. However, certain domain adaptation algorithms may be used to combine training data in the target domain with large amounts of training data from a different source domain.

The Daumé (2007) approach involves the transformation of the original training instance feature vector into a new space made up of three copies of the original vector. The three copies correspond to the target domain, the source domain, and the general domain. If X represents an original feature vector from the *target domain*, then it is transformed into XOX, where O is a zero vector. If X represents original feature vector from the *source domain*, then it is transformed into OXX. This data is given to the learning algorithm, which learns information specific to the target domain, specific to the source domain, as well as information that applies to both domains. The test instance feature vector (which is from the target domain) is transformed to XOX. Therefore, the classifier applies information specific to the target domain as well as information common to both the target and source domains, but not information specific only to the source domain.

In this section, we describe experiments on using the Twitter Emotion Corpus for emotion classification in the newspaper headlines domain. We applied our binary emotion classifiers on unseen test data from the newspaper headlines domain—the 250-headlines dataset—using each of the following as a training corpus:

- Target-domain data: the 1000-headlines data.
- Source-domain data: the TEC.
- Target and Source data: A joint corpus of the 1000-headlines dataset and the TEC.

Additionally, when using the 'Target and Source' data, we also tested the domain adaptation algorithm proposed in Daumé (2007). Since the *EmotionX* class (the positive class) has markedly fewer instances than the *NotEmotionX* class, we assigned higher weight to instances of the positive class during training.[11] The rows under I in Table 6 give the results. (Row II results are for the experiment described in Section 6, and can be ignored for now.)

We see that the macro-averaged F-score when using target-domain data (row I.a.) is identical to the score obtained by the random baseline (row III). However, observe that the precision of the ngram system is higher than the random system, and its recall is lower. This suggests that the test data has many n-grams not previously seen in the training data. Observe that as expected, using source-domain data produces much lower scores (row I.b.) than when using target-domain training data (row I.a.).

Using both target- and source-domain data produced significantly better results (row I.c.1.) than

[11] For example, for the *anger–NotAnger* classifier, if 10 out of 110 instances have the label anger, then they are each given a weight of 10, whereas the rest are given a weight of 1.

	# of features	P	R	F
I. System using ngrams in training data:				
a. the 1000-headlines text (target domain)	1,181	40.2	32.1	35.7
b. the TEC (source domain)	32,954	29.9	26.1	27.9
c. the 1000-headlines text and the TEC (target and source)				
c.1. no domain adaptation	33,902	41.7	35.5	38.3
c.2. with domain adaptation	101,706	46.0	35.5	**40.1**
II. System using ngrams in 1000-headlines and:				
a. the TEC lexicon	1,181 + 6	44.4	35.3	39.3
b. the WordNet Affect lexicon	1,181 + 6	39.7	30.5	34.5
c. the NRC emotion lexicon	1,181 + 10	46.7	38.6	**42.2**
III. System that guesses randomly	-	27.8	50.0	35.7

Table 6: Results on the 250-headlines dataset. The highest F-scores in I and II are shown in bold.

using target-domain data alone (I.a.). Applying the domain adaptation technique described in Daumé (2007), obtained even better results (row I.c.2.). (We used the Fisher Exact Test and a confidence interval of 95% for all precision and recall significance testing reported in this paper.) The use of TEC improved both precision and recall over just using the target-domain text. This shows that the Twitter Emotion Corpus can be leveraged, preferably with a suitable domain adaptation algorithm, to improve emotion classification results even on datasets from a different domain. It is also a validation of the premise that the self-labeled emotion hashtags are consistent, at least to some degree, with the emotion labels given by trained human judges.

6 Creating the TEC Emotion Lexicon

Word–emotion association lexicons are lists of words and associated emotions. For example, the word *victory* may be associated with the emotions of joy and relief. These emotion lexicons have many applications, including automatically highlighting words and phrases to quickly convey regions of affect in a piece of text. Mohammad (2012b) shows that these lexicon features can significantly improve classifier performance over and above that obtained using ngrams alone.

WordNet Affect (Strapparava and Valitutti, 2004) includes 1536 words with associations to the six Ekman emotions.[12] Mohammad and colleagues compiled emotion annotations for about 14,000 words by crowdsourcing to Mechanical Turk (Mohammad and Turney, 2012; Mohammad and Yang, 2011).[13] This lexicon, referred to as the NRC emotion lexicon, has annotations for eight emotions (six of Ekman, trust, and anticipation) as well as for positive and negative sentiment.[14] Here, we show how we created an ngram–emotion association lexicon from emotion-labeled sentences in the 1000-headlines dataset and the TEC.

6.1 Method

Given a dataset of sentences and associated emotion labels, we compute the *Strength of Association (SoA)* between an n-gram n and an emotion e to be:

$$SoA(n, e) = PMI(n, e) - PMI(n, \neg e) \quad (3)$$

where PMI is the pointwise mutual information.

$$PMI(n, e) = \log \frac{freq(n, e)}{freq(n) * freq(e)} \quad (4)$$

where $freq(n, e)$ is the number of times n occurs in a sentence with label e. $freq(n)$ and $freq(e)$ are the frequencies of n and e in the labeled corpus.

$$PMI(n, \neg e) = \log \frac{freq(n, \neg e)}{freq(n) * freq(\neg e)} \quad (5)$$

where $freq(n, \neg e)$ is the number of times n occurs in a sentence that does not have the label e. $freq(\neg e)$ is the number of sentences that do not have the label e. Thus, equation 4 is simplified to:

$$SoA(n, e) = \log \frac{freq(n, e) * freq(\neg e)}{freq(e) * freq(n, \neg e)} \quad (6)$$

[12] http://wndomains.fbk.eu/wnaffect.html

[13] http://www.purl.org/net/saif.mohammad/research

[14] Plutchik (1985) proposed a model of 8 basic emotions.

Emotion lexicon	# of word types
1000-headlines lexicon	152
TEC lexicon	11,418
WordNet Affect lexicon	1,536
NRC emotion lexicon	14,000

Table 7: Number of word types in emotion lexicons.

Since PMI is known to be a poor estimator of association for low-frequency events, we ignored ngrams that occurred less than five times.

If an n-gram has a stronger tendency to occur in a sentence with a particular emotion label, than in a sentence that does not have that label, then that ngram–emotion pair will have an SoA score that is greater than zero.

6.2 Emotion lexicons created from the 1000-headlines dataset and the TEC

We calculated SoA scores for the unigrams and bigrams in the TEC with the six basic emotions. All ngram–emotion pairs that obtained scores greater than zero were extracted to form the TEC emotion lexicon. We repeated these steps for the 1000-headlines dataset as well. Table 7 shows the number of word types in the two automatically generated and the two manually created lexicons. Observe that the 1000-headlines dataset produces very few entries, whereas the large size of the TEC enables the creation of a substantial emotion lexicon.

6.3 Evaluating the TEC lexicon

We evaluate the TEC lexicon by using it for classifying emotions in a setting similar to the one discussed in the previous section. The test set is the 250-headlines dataset. The training set is the 1000-headlines dataset. We used binary features that captured the presence or absence of unigrams and bigrams just as before. Additionally, we also used integer-valued affect features that captured the number of word tokens in a sentence associated with different emotions labels in the TEC emotion lexicon and the WordNet Affect lexicon. For example, if a sentence has two joy words and one surprise word, then the joy feature has value 2, surprise has value 1, and all remaining affect features have value 0.[15]

We know from the results in Table 6 (I.a. and I.c) that using the Twitter Emotion Corpus in addition to the 1000-headlines training data significantly improves results. Now we investigate if the TEC lexicon, which is created from TEC, can similarly improve performance. The rows under II in Table 6 give the results.

Observe that even though the TEC lexicon is a derivative of the TEC that includes fewer unigrams and bigrams, the classifiers using the TEC lexicon produces an F-score (row II.a.) significantly higher than in the scenarios of I.a. and almost as high as in I.c.2. This shows that the TEC lexicon successfully captures the word–emotion associations that are latent in the Twitter Emotion Corpus. We also find that the the classifiers perform significantly better when using the TEC lexicon (row II.a.) than when using the WordNet Affect lexicon (row II.b.), but not as well as when using the NRC emotion lexicon (row II.c.). The strong results of the NRC emotion lexicon are probably because of its size and because it was created by direct annotation of words for emotions, which required significant time and effort. On the other hand, the TEC lexicon can be easily improved further by compiling an even larger set of tweets using synonyms and morphological variants of the emotion words used thus far.

7 Conclusions and Future Work

We compiled a large corpus of tweets and associated emotions using emotion-word hashtags. Even though the corpus has tweets from several thousand people, we showed that the self-labeled hashtag annotations are consistent. We also showed how the Twitter emotion corpus can be combined with labeled data from a different target domain to improve classification accuracy. This experiment was especially telling since it showed that self-labeled emotion hashtags correspond well with annotations of trained human judges. Finally we extracted a large word–emotion association lexicon from the Twitter emotion corpus. Our future work includes collecting tweets with hashtags for various other emotions and also hashtags that are near-synonyms of the basic emotion terms described in this paper.

Acknowledgments

We thank Tara Small and Peter Turney for helpful discussions. For Archivist, we thank its creators.

[15] Normalizing by sentence length did not give better results.

References

Cecilia O. Alm and Richard Sproat, 2005. *Emotional sequencing and development in fairy tales*, pages 668–674. Springer.

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005. Emotions from text: Machine learning for text-based emotion prediction. In *Proceedings of the Joint Conference on HLT–EMNLP*, Vancouver, Canada.

Saima Aman and Stan Szpakowicz. 2007. Identifying expressions of emotion in text. In Vclav Matoušek and Pavel Mautner, editors, *Text, Speech and Dialogue*, volume 4629 of *Lecture Notes in Computer Science*, pages 196–205. Springer Berlin / Heidelberg.

Johan Bollen, Alberto Pepe, and Huina Mao. 2009. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. *CoRR*.

Anthony C. Boucouvalas. 2002. Real time text-to-emotion engine for expressive internet communication. *Emerging Communication: Studies on New Technologies and Practices in Communication*, 5:305–318.

J. R. G. Bougie, R. Pieters, and M. Zeelenberg. 2003. Angry customers don't come back, they get back: The experience and behavioral implications of anger and dissatisfaction in services. Open access publications from tilburg university, Tilburg University.

Danah Boyd, Scott Golder, and Gilad Lotan. 2010. Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. volume 0, pages 1–10, Los Alamitos, CA, USA. IEEE Computer Society.

Soumaya Chaffar and Diana Inkpen. 2011. Using a heterogeneous dataset for emotion analysis in text. In *Canadian Conference on AI*, pages 62–67.

Hal Daumé. 2007. Frustratingly easy domain adaptation. In *Conference of the Association for Computational Linguistics (ACL)*, Prague, Czech Republic.

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2010. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1277–1287, Stroudsburg, PA. Association for Computational Linguistics.

Paul Ekman. 1992. An argument for basic emotions. *Cognition and Emotion*, 6(3):169–200.

Virginia Francisco and Pablo Gervás. 2006. Automated mark up of affective information in english texts. In Petr Sojka, Ivan Kopecek, and Karel Pala, editors, *Text, Speech and Dialogue*, volume 4188 of *Lecture Notes in Computer Science*, pages 375–382. Springer Berlin / Heidelberg.

Michel Genereux and Roger P. Evans. 2006. Distinguishing affective states in weblogs. In *AAAI-2006 Spring Symposium on Computational Approaches to Analysing Weblogs*, pages 27–29, Stanford, California.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. In *Final Projects from CS224N for Spring 2008–2009 at The Stanford Natural Language Processing Group*.

Roberto González-Ibáñez, Smaranda Muresan, and Nina Wacholder. 2011. Identifying sarcasm in twitter: a closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, pages 581–586, Portland, Oregon.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The weka data mining software: an update. *SIGKDD*, 11:10–18.

Lars E. Holzman and William M. Pottenger. 2003. Classification of emotions in internet chat: An application of machine learning using speech phonemes. Technical report, Leigh University.

David John, Anthony C. Boucouvalas, and Zhe Xu. 2006. Representing emotional momentum within expressive internet communication. In *Proceedings of the 24th IASTED international conference on Internet and multimedia systems and applications*, pages 183–188, Anaheim, CA. ACTA Press.

Elsa Kim, Sam Gilbert, Michael J. Edwards, and Erhardt Graeff. 2009. Detecting sadness in 140 characters: Sentiment analysis of mourning michael jackson on twitter.

Hugo Liu, Henry Lieberman, and Ted Selker. 2003. A model of textual affect sensing using real-world knowledge. In *Proceedings of the 8th international conference on Intelligent user interfaces*, IUI '03, pages 125–132, New York, NY. ACM.

Chunling Ma, Helmut Prendinger, and Mitsuru Ishizuka. 2005. Emotion estimation and reasoning based on affective textual interaction. In J. Tao and R. W. Picard, editors, *First International Conference on Affective Computing and Intelligent Interaction (ACII-2005)*, pages 622–628, Beijing, China.

Rada Mihalcea and Hugo Liu. 2006. A corpus-based approach to finding happiness. In *AAAI-2006 Spring Symposium on Computational Approaches to Analysing Weblogs*, pages 139–144. AAAI Press.

Saif M. Mohammad and Peter D. Turney. 2012. Crowdsourcing a word–emotion association lexicon. *To Appear in Computational Intelligence*.

Saif M. Mohammad and Tony Yang. 2011. Tracking Sentiment in Mail: How Genders Differ on Emotional Axes. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011)*, pages 70–79, Portland, Oregon. Association for Computational Linguistics.

Saif M. Mohammad. 2012a. From once upon a time to happily ever after: Tracking emotions in mail and books. *To Appear in Decision Support Systems*.

Saif M. Mohammad. 2012b. Portable features for emotion classification. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2012)*, Montreal, Canada. Association for Computational Linguistics.

Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka. 2009. Compositionality principle in recognition of fine-grained emotions from text. In *Proceedings of the Proceedings of the Third International Conference on Weblogs and Social Media (ICWSM-09)*, pages 278–281, San Jose, California.

Lisa Pearl and Mark Steyvers. 2010. Identifying emotions, intentions, and attitudes in text using a game with a purpose. In *Proceedings of the NAACL-HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, Los Angeles, California.

John Platt. 1999. Using analytic qp and sparseness to speed training of support vector machines. In *In Neural Info. Processing Systems 11*, pages 557–563. MIT Press.

Robert Plutchik. 1985. On emotion: The chicken-and-egg problem revisited. *Motivation and Emotion*, 9(2):197–200.

Niklas Ravaja, Timo Saari, Marko Turpeinen, Jari Laarni, Mikko Salminen, and Matias Kivikangas. 2006. Spatial presence and emotions during video game playing: Does it matter with whom you play? *Presence: Teleoperators and Virtual Environments*, 15(4):381–392.

Tapas Ray. 2011. The 'story' of digital excess in revolutions of the arab spring. *Journal of Media Practice*, 12(2):189–196.

Julia Skinner. 2011. Social media and revolution: The arab spring and the occupy movement as seen through three information studies paradigms. *Sprouts: Working Papers on Information Systems*, 11(169).

Carlo Strapparava and Rada Mihalcea. 2007. Semeval-2007 task 14: Affective text. In *Proceedings of SemEval-2007*, pages 70–74, Prague, Czech Republic.

Carlo Strapparava and Alessandro Valitutti. 2004. Wordnet-Affect: An affective extension of WordNet. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC-2004)*, pages 1083–1086, Lisbon, Portugal.

Andranik Tumasjan, Timm O Sprenger, Philipp G Sandner, and Isabell M Welpe. 2010. Predicting elections with twitter : What 140 characters reveal about political sentiment. *Word Journal Of The International Linguistic Association*, pages 178–185.

Juan D. Velásquez. 1997. Modeling emotions and other motivations in synthetic agents. In *Proceedings of the fourteenth national conference on artificial intelligence and ninth conference on Innovative applications of artificial intelligence*, AAAI'97/IAAI'97, pages 10–15. AAAI Press.

Xu Zhe and A Boucouvalas, 2002. *Text-to-Emotion Engine for Real Time Internet CommunicationText-to-Emotion Engine for Real Time Internet Communication*, pages 164–168.