In the 4th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis (WASSA 2013)

Bootstrapped Learning of Emotion Hashtags #hashtags4you

Ashequl Qadir
School of Computing
University of Utah
Salt Lake City, UT 84112, USA
asheq@cs.utah.edu

Ellen Riloff
School of Computing
University of Utah
Salt Lake City, UT 84112, USA
riloff@cs.utah.edu

Abstract

We present a bootstrapping algorithm to automatically learn hashtags that convey emotion. Using the bootstrapping framework, we learn lists of emotion hashtags from unlabeled tweets. Our approach starts with a small number of seed hashtags for each emotion, which we use to automatically label tweets as initial training data. We then train emotion classifiers and use them to identify and score candidate emotion hashtags. We select the hashtags with the highest scores, use them to automatically harvest new tweets from Twitter, and repeat the bootstrapping process. We show that the learned hashtag lists help to improve emotion classification performance compared to an N-gram classifier, obtaining 8% micro-average and 9% macro-average improvements in F-measure.

1 Introduction

The increasing popularity of social media has given birth to new genres of text that have been the focus of NLP research for applications such as event discovery (Benson et al., 2011), election outcome prediction (Tumasjan et al., 2011; Bermingham and Smeaton, 2011), user profile classification (De Choudhury et al., 2012), conversation modeling (Ritter et al., 2010), consumer insight discovery (Chamlertwat et al., 2012), etc. A hallmark of social media is that people tend to share their personal feelings, often in publicly visible forums. As a result, social media has also been the focus of NLP research on sentiment analysis (Kouloumpis et al., 2011), emotion classification and lexicon generation (Mohammad, 2012), and sarcasm detection (Davidov et al., 2010). Identifying emotion in social media text could be beneficial for many application areas, for example to help companies understand how people feel about their products, to assist governments in recognizing growing anger or fear associated with an event, and to help media outlets understand the public's emotional response toward controversial issues or international affairs.

Twitter, a micro-blogging platform, is particularly well-known for its use by people who like to instantly express thoughts within a limited length of 140 characters. These status updates, known as tweets, are often emotional. Hashtags are a distinctive characteristic of tweets, which are a community-created convention for providing meta-information about a tweet. Hashtags are created by adding the '*#*' symbol as a prefix to a word or a multi-word phrase that consists of concatenated words without whitespace (e.g., *#welovehashtags*). People use hashtags in many ways, for example to represent the topic of a tweet (e.g., *#graduation*), to convey additional information (e.g., *#mybirthdaytoday*), or to express an emotion (e.g., *#pissedoff*).

The usage of hashtags in tweets is common, as reflected in the study of a sample of 0.6 million tweets by Wang et al. (2011) which found that 14.6% of tweets in their sample had at least one hashtag. In tweets that express emotion, it is common to find hashtags representing the emotion felt by the tweeter, such as *"the new iphone is a waste of money! nothing new! #angry"* denoting anger or *"buying a new sweater for my mom for her birthday! #loveyoumom"* denoting affection.

Identifying the emotion conveyed by a hashtag has not yet been studied by the natural language processing community. The goal of our research is to automatically identify hashtags that express one of five emotions: *affection*, *anger/rage*, *fear/anxiety*, *joy*, or *sadness/disappointment*. The learned hashtags are then used to recognize tweets that express one of these emotions. We use a bootstrapping approach that begins with 5 seed hashtags for each emotion class and iteratively learns more hashtags from unlabeled tweets. We show that the learned hashtags can accurately identify tweets that convey emotion and yield additional coverage beyond the recall of an N-gram classifier.

The rest of the paper is divided into the following sections. In Section 2, we present a brief overview of previous research related to emotion classification in social media and the use of hashtags. In Section 3, we describe our bootstrapping approach for learning lists of emotion hashtags. In Section 4 we discuss the data collection process and our experimental design. In Section 5, we present the results of our experiments. Finally, we conclude by summarizing our findings and presenting directions for future work.

2 Related Work

Recognizing emotions in social media texts has grown popular among researchers in recent years. Roberts et al. (2012) investigated feature sets to classify emotions in Twitter and presented an analysis of different linguistic styles people use to express emotions. The research of Kim et al. (2012a) is focused on discovering emotion influencing patterns to classify emotions in social network conversations. Esmin et al. (2012) presented a 3-level hierarchical emotion classification approach by differentiating between emotion vs. non-emotion text, positive vs. negative emotion, and then classified different emotions. Yang et al. (2007b) investigated sentence contexts to classify emotions in blogs at the document level. Some researchers have also worked on analyzing the correlation of emotions with topics and trends. Kim et al. (2012b) analyzed correlations between topics and emotions in Twitter using topic modeling. Gilbert and Karahalios (2010) analyzed correlation of anxiety, worry and fear with downward trends in the stock market. Bollen et al. (2011) modeled public mood and emotion by creating six-dimensional mood vectors to correlate with popular events that happened in the timeframe of the dataset.

On the other hand, researchers have recently started to pay attention to the hashtags of tweets, but mostly to use them to collect labeled data. Davidov et al. (2010) used *#sarcasm* to collect sarcastic tweets from twitter. Choudhury et al. (2012) used hashtags of 172 mood words to collect training data to find associations between mood and human affective states, and trained classifiers with unigram and bigram features to classify these states. Purver and Battersby (2012) used emotion class name hashtags and emoticons as distant supervision in emotion classification. Mohammad (2012) also used emotion class names as hashtags to collect labeled data from Twitter, and used these tweets to generate emotion lexicons. Wang et al. (2012) used a selection of emotion hashtags as the means to acquire labeled data from twitter, and found that a combination of unigrams, bigrams, sentiment/emotion-bearing words, and parts-of-speech information to be the most effective in classifying emotions. A study by Wang et al. (2012) also shows that hashtags can be used to create a high quality emotion dataset. They found about 93.16% of the tweets having emotion hashtags were relevant to the corresponding emotion.

However, none of this work investigated the use of emotion hashtag lists to help classify emotions in tweets. In cases where hashtags were used to collect training data, the hashtags were manually selected for each emotion class. In many cases, only the name of the emotion classes were used for this purpose. The work most closely related to our research focus is the work of Wang et al. (2011) where they investigated several graph based algorithms to collectively classify hashtag sentiments. However, their work is focused on classifying hashtags of positive and negative sentiment polarities, and they made use of sentiment polarity of the individual tweets to classify hashtag sentiments. On the contrary, we learn emotion hashtags and use the learned hashtag lists to classify emotion tweets. To the best of our knowledge, we are the first to present a bootstrapped learning framework to automatically learn emotion hashtags from unlabeled data.

3 Learning Emotion Hashtags via Bootstrapping

3.1 Motivation

The hashtags that people use in tweets are often very creative. While it is common to use just single word hashtags (e.g., *#angry)*, many hashtags are multi-word phrases (e.g., *#LoveHimSoMuch*). People also use elongated[1] forms of words (e.g., *#yaaaaay*, *#goawaaay*) to put emphasis on their emotional state. In addition, words are often spelled creatively by replacing a word with a number or replacing some characters with phonetically similar characters (e.g., *#only4you*, *#YoureDaBest*). While many of these hashtags convey emotions, these stylistic variations in the use of hashtags make it very difficult to create a repository of emotion hashtags manually. While emotion word lexicons exist (Yang et al., 2007a; Mohammad, 2012), and adding a '#' symbol as a prefix to these lexicon entries could potentially give us lists of emotion hashtags, it would be unlikely to find multi-word phrases or stylistic variations frequently used in tweets. This drives our motivation to automatically learn hashtags that are commonly used to express emotion in tweets.

3.2 Emotion Classes

For this research, we selected 5 prominent emotion classes that are frequent in tweets: *Affection*, *Anger/Rage*, *Fear/Anxiety*, *Joy* and *Sadness/Disappointment*. We started by analyzing Parrott's (Parrott, 2001) emotion taxonomy and how these emotions are expressed in tweets. We also wanted to ensure that the selected emotion classes would have minimal overlap with each other. We took Parrott's primary emotion *Joy* and *Fear*[2] directly. We merged Parrott's secondary emotion *Affection* and *Lust* into our *Affection* class and merged Parrott's secondary emotion *Sadness* and *Disappointment* into our *Sadness/Disappointment* class, since these emotions are often difficult to distinguish from each other. Lastly, we mapped Parrott's secondary emotion *Rage* to our *Anger/Rage* class directly. There were other emotions in Parrott's taxonomy such as *Surprise*, *Neglect*, etc. that we did not use for this research. In addition to the five emotion classes, we used a *None of the Above* class for tweets that do not carry any emotion or that carry an emotion other than one of our five emotion classes.

[1]This feature has also been found to have a strong association with sentiment polarities (Brody and Diakopoulos, 2011)

[2]we renamed the *Fear* class as *Fear/Anxiety*

3.3 Overview of Bootstrapping Framework

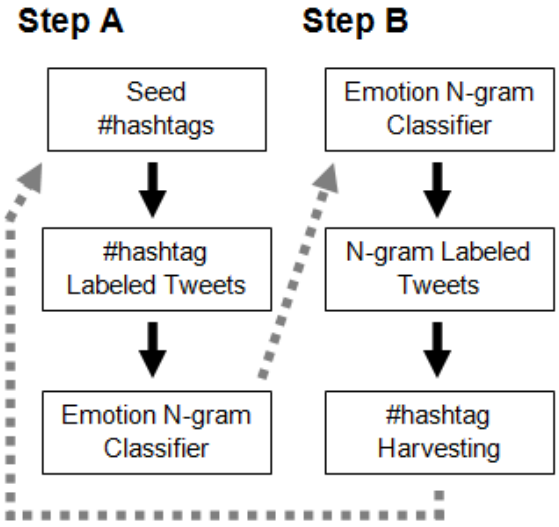

Figure 1: Bootstrapping Architecture

Figure 1 presents the framework of our bootstrapping algorithm for learning emotion hashtags. The algorithm runs in two steps. In the first step, the bootstrapping process begins with five manually defined "seed" hashtags for each emotion class. For each seed hashtag, we search Twitter for tweets that contain the hashtag and label these tweets with the emotion class associated with the hashtag. We use these labeled tweets to train a supervised N-gram classifier for every emotion $e \in E$, where E is the set of emotion classes we are classifying.

In the next step, the emotion classifiers are applied to a large pool of unlabeled tweets and we collect the tweets that are labeled by the classifiers. From these labeled tweets, we extract the hashtags found in these tweets to create a candidate pool of emotion hashtags. The hashtags in the candidate pool are then scored and ranked and we select the most highly ranked hashtags to add to a hashtag repository for each emotion class.

Finally, we then search for tweets that contain the learned hashtags in a pool of unlabeled tweets and label each of these with the appropriate emotion class. These newly labeled tweets are added to the

set of training instances. The emotion classifiers are retrained using the larger set of training instances, and the bootstrapping process continues.

3.4 Seeding

For each of the 5 emotion classes, we manually selected 5 seed hashtags that we determined to be strongly representative of the emotion. Before collecting the initial training tweets containing the seed hashtags, we manually searched in Twitter to ensure that these seed hashtags are frequently used by tweeters. Table 1 presents our seed hashtags.

Emotion Classes	Seed Hashtags
AFFECTION	*#loveyou*, *#sweetheart*, *#bff* *#romantic*, *#soulmate*
ANGER & RAGE	*#angry*, *#mad*, *#hateyou* *#pissedoff*, *#furious*
FEAR & ANXIETY	*#afraid*, *#petrified*, *#scared* *#anxious*, *#worried*
JOY	*#happy*, *#excited*, *#yay* *#blessed*, *#thrilled*
SADNESS & DISAPPOINT-MENT	*#sad*, *#depressed* *#disappointed*, *#unhappy* *#foreveralone*

Table 1: Seed Emotion Hashtags

3.5 N-gram Tweet Classifier

The tweets acquired using the seed hashtags are used as training instances to create emotion classifiers with supervised learning. We first pre-process the training instances by tokenizing the tweets with a freely available tokenizer for Twitter (Owoputi et al., 2013). Although it is not uncommon to express emotion states in tweets with capitalized characters inside words, the unique writing styles of the tweeters often create many variations of the same words and hashtags. We, therefore, normalized case to ensure generalization.

We trained one logistic regression classifier for each emotion class. We chose logistic regression as the classification algorithm because it produces probabilities along with each prediction that we later use to assign scores to candidate emotion hashtags. As features, we used unigrams to represent all of the words and hashtags in a tweet, but we removed the seed hashtags that were used to select the tweets (or the classifier would simply learn to recognize the seed hashtags). Our hypothesis is that the seed hashtag will not be the only emotion indicator in a tweet, most of the time. The goal is for the classifier to learn to recognize words and/or additional hashtags that are also indicative of the emotion. Additionally, we removed from the feature set any user mentions (by looking for words with '@' prefix). We also removed any word or hashtag from the feature set that appeared only once in the training data.

For emotion e, we used the tweets containing seed hashtags for e as the positive training instances and the tweets containing hashtags for the other emotions as negative instances. However, we also needed to provide negative training instances that do not belong to any of the 5 emotion classes. For this purpose, we added 100,000 randomly collected tweets to the training data. While it is possible that some of these tweets are actually positive instances for e, our hope is that the vast majority of them will not belong to emotion e.

We experimented with feature options such as bigrams, unigrams with the '#' symbol stripped off from hashtags, etc., but the combination of unigrams and hashtags as features worked the best. We used the freely available java version of the LIBLINEAR (Fan et al., 2008) package with its default parameter settings for logistic regression.

3.6 Learning Emotion Hashtags

The next step is to learn emotion hashtags. We apply the emotion classifiers to a pool of unlabeled tweets and collect all of the tweets that the classifier can label. For each emotion $e \in E$, we first create a candidate pool of emotion hashtags H_e, by collecting all of the hashtags in the labeled tweets for emotion e. To limit the size of the candidate pool, we discarded hashtags with just one character or more than 20 characters, and imposed a frequency threshold of 10. We then score these hashtags to select the top N emotion hashtags we feel most confident about.

To score each candidate hashtag $h \in H_e$, we compute the average of the probabilities assigned by the logistic regression classifier to all the tweets containing hashtag h. We expect the classifier to assign higher probabilities only to tweets it feels confident about. Therefore, if h conveys e, we expect that

the average probability of all the tweets containing h will also be high. We select the top 10 emotion hashtags for each emotion class e, and add them to our list of learned hashtags for e.

3.7 Adding New Training Instances for Bootstrapping

To facilitate the next stage of bootstrapping, we collect all tweets from the unlabeled data that contain hashtag h and label them with the emotion associated with h. By adding more training instances, we expect to provide the classifiers with new tweets that will contain a potentially more diverse set of words that the classifiers can consider in the next stage of the bootstrapping.

When the new tweets are added to the training set, we remove the hashtags from them that we used for labelling to avoid bias, and the bootstrapping process continues. We ran the bootstrapped learning for 100 iterations. Since we learned 10 hashtags during each iteration, we ended up with emotion hashtag lists consisting of 1000 hashtags for each emotion.

4 Experimental Setup

4.1 Data Collection

To collect our initial training data, we searched Twitter for the seed hashtags mentioned in Section 3.4 using Twitter's Search API[3] over a period of time. To ensure that the collected tweets are written in English, we used a freely available language recognizer trained for tweets (Carter et al., 2013). We filtered out tweets that were marked as re-tweets using *#rt* or beginning with *"rt"*[4] because re-tweets are in many cases exactly the same or very similar to the original. We also filtered out any tweet containing a URL because if such a tweet contains emotion, it is possible that the emotion indicator may be present only on the linked website (e.g., a link to a comic strip followed by an emotion hashtag). After these filtering steps, we ended up with a seed labeled training dataset of 325,343 tweets.

In addition to the seed labeled data, we collected random tweets using Twitter's Streaming API[5] over a period of time to use as our pool of unlabeled tweets. Like the training data, we filtered out re-tweets and tweets containing a URL as well as tweets containing any of the seed hashtags. Since our research focus is on learning emotion hashtags, we also filtered out any tweet that did not have at least one hashtag. After filtering, we ended up with roughly 2.3 million unlabeled tweets.

4.2 Test Data

Since manual annotation is time consuming, to ensure that many tweets in our test data have at least one of our 5 emotions, we manually selected 25 topic keywords/phrases[6] that we considered to be strongly associated with emotions, but not necessarily any specific emotion. We then searched in Twitter for any of these topic phrases and their corresponding hashtags. These 25 topic phrases are: *Prom, Exam, Graduation, Marriage, Divorce, Husband, Wife, Boyfriend, Girlfriend, Job, Hire, Laid Off, Retirement, Win, Lose, Accident, Failure, Success, Spider, Loud Noise, Chest Pain, Storm, Home Alone, No Sleep* and *Interview*. Since the purpose of collecting these tweets is to evaluate the quality and coverage of the emotion hashtags that we learn, we filtered out any tweet that did not have at least one hashtag (other than the topic hashtag).

To annotate tweets with respect to emotion, two annotators were given definitions of the 5 emotion classes from Collins English Dictionary[7], Parrott's (Parrott, 2001) emotion taxonomy of these 5 emotions and additional annotation guidelines. The annotators were instructed to label each tweet with up to two emotions. The instructions specified that the emotion must be felt by the tweeter at the time the tweet was written. After several trials and discussions, the annotators reached a satisfactory agreement level of 0.79 Kappa (κ) (Carletta, 1996). The annotation disagreements in these 500 tweets were then adjudicated, and each annotator labeled an additional 2,500 tweets. Altogether this gave us an emotion annotated dataset of 5,500 tweets. We randomly separated out 1,000 tweets from this collection as a tuning set, and used the remaining 4,500 tweets as evaluation data.

In Table 2, we present the emotion distribution in

[3] https://dev.twitter.com/docs/api/1/get/search

[4] a typical convention to mark a tweet as a re-tweet

[5] https://dev.twitter.com/docs/streaming-apis

[6] This data collection process is similar to the emotion tweet dataset creation by Roberts et al. (2012)

[7] http://www.collinsdictionary.com/

tweets that were labeled using the seed hashtags in the second column. In the next column, we present the emotion distribution in the tweets that were annotated for evaluation by the human annotators.

Emotion	Tweets with Seed Hashtags	Evaluation Tweets
AFFECTION	14.38%	6.42%
ANGER/RAGE	14.01%	8.91%
FEAR/ANXIETY	11.42%	13.16%
JOY	37.47%	22.33%
SADNESS/ DISAPPOINTMENT	23.69%	12.45%
NONE OF THE ABOVE	-	42.38%

Table 2: Distribution of emotions in tweets with seed hashtags and evaluation tweets

4.3 Evaluating Emotion Hashtags

For comparison, we trained logistic regression classifiers with word unigrams and hashtags as features for each emotion class, and performed 10-fold cross-validation on the evaluation data. As a second baseline for comparison, we added bigrams to the feature set of the classifiers.

To decide on the optimum size of the lists for each emotion class, we performed list lookup on the tuning data that we had set aside before evaluation. For any hashtag in a tweet in the tuning dataset, we looked up that hashtag in our learned lists, and if found, assigned the corresponding emotion as the label for that tweet. We did this experiment starting with only seeds in our lists, and incrementally increased the sizes of the lists by 50 hashtags at each experiment. We decided on the optimum size based on the best F-measure obtained for each emotion class. In Table 3, we show the list sizes we found to achieve the best F-measure for each emotion class in the tuning dataset.

Emotion	List Sizes
AFFECTION	500
ANGER/RAGE	1000
FEAR/ANXIETY	850
JOY	1000
SADNESS/DISAPPOINTMENT	400

Table 3: Optimum list sizes decided from tuning dataset

To use the learned lists of emotion hashtags for classifying emotions in tweets, we first used them as features for the logistic regression classifiers. We created 5 list features with binary values, one for each emotion class. Whenever a tweet in the evaluation data contained a hashtag from one of the learned emotion hashtags lists, we set the value of that list feature to be 1, and 0 otherwise. We used these 5 new features in addition to the word unigrams and hashtag features, and evaluated the classification performance of the logistic regression classifiers in a 10-fold cross-validation setup by calculating precision, recall and F-measure.

Since the more confident hashtags are added to the lists at the beginning stages of bootstrapping, we also tried creating subsets from each list by grouping hashtags together that were learned after each 5 iterations of bootstrapping (50 hashtags in each subset). We then created 20 list subset features for each emotion with binary values, yielding 100 additional features in total. We also evaluated this feature representation of the hashtag lists in a 10-fold cross-validation setup.

As a different approach, we also used the lists independently from the logistic regression classifiers. For any hashtag in the evaluation tweets, we looked up the hashtag in our learned lists. If the hashtag was found, we assigned the corresponding emotion class label to the tweet containing the hashtag. Lastly, we combined the list lookup decisions with the decisions of the baseline logistic regression classifiers by taking a union of the decisions, i.e., if either assigned an emotion to a tweet, we assigned that emotion as the label for the tweet. We present the results of these different approaches in Section 5.

5 Results and Analysis

Table 4 shows the precision, recall and F-measure of the N-gram classifier as well as several different utilizations of the learned hashtag lists. The first and the second row in Table 4 correspond to the results for the baseline unigram classifier (UC) alone and when bigrams are added to the feature set. These baseline classifiers had low recall for most emotion classes, suggesting that the N-grams and hashtags are not adequate as features to recognize the emotion classes.

Results of using the hashtag lists as 5 additional features for the classifier are shown in the third row

Evaluation	Affection	Anger Rage	Fear Anxiety	Joy	Sadness Disappointment
	P R F	P R F	P R F	P R F	P R F
Baseline Classifiers					
Unigram Classifier (UC)	67 43 52	51 19 28	63 33 43	65 48 55	57 29 39
UC + Bigram Features	70 38 50	52 15 23	64 29 40	65 45 53	57 25 34
Baseline Classifier with List Features					
UC + List Features	71 49 58	56 28 37	67 41 51	66 50 57	61 34 44
UC + List Subset Features	73 45 56	58 23 33	69 38 49	66 48 55	61 32 42
List Lookup					
Seed Lookup	**94** 06 11	**75** 01 03	**100** 06 11	**93** 04 08	**81** 02 05
List Lookup	73 40 52	59 25 35	61 36 45	70 16 26	80 17 28
Baseline Classifier with List Lookup					
UC ∪ Seed Lookup	68 45 54	52 21 30	63 33 44	66 49 56	58 31 40
UC ∪ List Lookup	63 **60** **61**	52 **38** **44**	56 **53** **54**	64 **54** **59**	59 **38** **46**

Table 4: Emotion classification result (P = Precision, R = Recall, F = F-measure)

of Table 4. The hashtag lists consistently improve precision and recall across all five emotions. Compared to the unigram classifier, F-measure improved by 6% for AFFECTION, by 9% for ANGER/RAGE, by 8% for FEAR/ANXIETY, by 2% for JOY, and by 5% for SADNESS/DISAPPOINTMENT. The next row presents the results when the list subset features were used. Using this feature representation as opposed to using each list as a whole shows precision recall tradeoff as the classifier learns to rely on the subsets of hashtags that are good, resulting in improved precision for several emotion classes, but recognizes emotions in fewer tweets, which resulted in less recall.

The fifth and the sixth rows of Table 4 show results of list lookup only. As expected, seed lookup recognizes emotions in tweets with high precision, but does not recognize the emotions in many tweets because the seed lists have only 5 hashtags per emotion class. Comparatively, using learned hashtag lists shows substantial improvement in recall as the learned lists contain a lot more emotion hashtags than the initial seeds.

Finally, the last two rows of Table 4 show classification performance of taking the union of the decisions made by the unigram classifier and the decisions made by matching against just the seed hashtags or the lists of learned hashtags. The union with the seed hashtags lookup shows consistent improvement across all emotion classes compared to the unigram baseline but the improvements are small. The union with the lookup in the learned lists of emotion hashtags shows substantial recall gains. This approach improves recall over the unigram baseline by 17% for AFFECTION, 19% for ANGER/RAGE, 20% for FEAR/ANXIETY, 6% for JOY, and 9% for SADNESS/DISAPPOINTMENT. At the same time, we observe that despite this large recall gain, precision is about the same or just a little lower. As a result, we observe an overall F-measure improvement of 9% for AFFECTION, 16% for ANGER/RAGE, 11% for FEAR/ANXIETY, 4% for JOY, and 7% for SADNESS/DISAPPOINTMENT.

Evaluation	Micro Average	Macro Average
	P R F	P R F
Baseline Classifiers		
Unigram Classifier (UC)	62 37 46	61 34 44
UC + Bigram Features	63 33 43	62 30 41
Baseline Classifier with List Features		
UC + List Features	65 42 51	64 40 49
UC + List Subset Features	66 39 49	65 37 48
List Lookup		
Seed Lookup	**93** 04 08	**89** 04 08
List Lookup	67 24 35	68 27 38
Baseline Classifier with List Lookup		
UC ∪ Seed Lookup	63 38 47	61 36 45
UC ∪ List Lookup	60 **49** **54**	59 **49** **53**

Table 5: Micro and Macro averages

Table 5 shows the overall performance improvement of the classifiers, averaged across all five emotion classes, measured as micro and macro aver-

AFFECTION	ANGER RAGE	FEAR ANXIETY	JOY	SADNESS DISAPPOINT-MENT
#youthebest	#godie	#hatespiders	#thankinggod	#catlady
#yourthebest	#donttalktome	#freakedout	#thankyoulord	#buttrue
#hyc	#fuckyourself	#creepedout	#thankful	#singleprobs
#yourethebest	#getoutofmylife	#sinister	#superexcited	#singleproblems
#alwaysandforever	#irritated	#wimp	#tripleblessed	#lonelytweet
#missyou	#pieceofshit	#shittingmyself	#24hours	#lonely
#loveyoumore	#ruinedmyday	#frightened	#ecstatic	#crushed
#loveyoulots	#notfriends	#paranoid	#happyme	#lonerproblems
#thanksforeverything	#yourgross	#haunted	#lifesgood	#unloved
#flyhigh	#madtweet	#phobia	#can'twait	#friendless
#comehomesoon	#stupidbitch	#shittingbricks	#grateful	#singlepringle
#yougotthis	#sofuckingannoying	#hateneedles	#goodmood	#brokenheart
#missyoutoo	#annoyed	#biggestfear	#superhappy	#singleforever
#youdabest	#fuming	#worstfear	#missedthem	#nosociallife
#otherhalf	#wankers	#concerned	#greatmood	#teamnofriends
#youramazing	#asshole	#waitinggame	#studio	#foreverugly
#cutiepie	#dontbothermewhen	#mama	#tgfl	#nofriends
#bestfriendforever	#fu	#prayforme	#exicted	#leftout
#alwayshereforyou	#fuckyou	#nightmares	#smiles	#singleforlife
#howimetmybestfriend	#yousuck	#baddriver	#liein	#:'(

Table 6: Top 20 hashtags learned for each emotion class

age precision, recall and F-measure scores. We see both types of feature representations of the hashtag lists improve precision and recall across all emotion classes over the N-gram classifier baselines. Using the union of the classifier and list lookup, we see a 12% recall gain with only 2% precision drop in micro-average over the unigram baseline, and 15% recall gain with only 2% precision drop in macro-average. As a result, we see an overall 8% micro-average F-measure improvement and 9% macro-average F-measure improvement.

In Table 6, we show the top 20 hashtags learned in each emotion class by our bootstrapped learning. While many of these hashtags express emotion, we also notice a few hashtags representing reasons (e.g., *#baddriver* in FEAR/ANXIETY) that are strongly associated with the corresponding emotion, as well as common misspellings (e.g., *#exicted* in JOY).

6 Conclusions

In this research we have presented a bootstrapped learning framework to automatically learn emotion hashtags. Our approach makes use of supervision from seed hashtag labeled tweets, and through a bootstrapping process, iteratively learns emotion hashtags. We have experimented with several approaches to use the lists of emotion hashtags for emotion classification and have found that the hashtag lists consistently improve emotion classification performance in tweets. In future research, since our bootstrapped learning approach does not rely on any language specific techniques, we plan to learn emotion hashtags in other prominent languages such as Spanish, Portuguese, etc.

7 Acknowledgments

This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center (DoI / NBC) contract number D12PC00285. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBE, or the U.S. Government.

References

Edward Benson, Aria Haghighi, and Regina Barzilay. 2011. Event discovery in social media feeds. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 389–398.

Adam Bermingham and Alan Smeaton. 2011. On using twitter to monitor political sentiment and predict election results. In *Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2011)*, pages 2–10.

Johan Bollen, Huina Mao, and Alberto Pepe. 2011. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *Proceedings of the Fifth International Conference on Weblogs and Social Media*.

Samuel Brody and Nicholas Diakopoulos. 2011. Cooooooooooooooollllllllllllll!!!!!!!!!!!!!!: using word lengthening to detect sentiment in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 562–570.

Jean Carletta. 1996. Assessing agreement on classification tasks: the kappa statistic. *Comput. Linguist.*, 22:249–254, June.

S. Carter, W. Weerkamp, and E. Tsagkias. 2013. Microblog language identification: Overcoming the limitations of short, unedited and idiomatic text. *Language Resources and Evaluation Journal*, 47(1).

Wilas Chamlertwat, Pattarasinee Bhattarakosol, Tippakorn Rungkasiri, and Choochart Haruechaiyasak. 2012. Discovering consumer insight from twitter via sentiment analysis. *Journal of Universal Computer Science*, 18(8):973–992, apr.

Munmun De Choudhury, Michael Gamon, and Scott Counts. 2012. Happy, nervous or surprised? classification of human affective states in social media. In *Proceedings of the Sixth International Conference on Weblogs and Social Media*.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Semi-supervised recognition of sarcastic sentences in twitter and amazon. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, CoNLL '10, pages 107–116.

Munmun De Choudhury, Nicholas Diakopoulos, and Mor Naaman. 2012. Unfolding the event landscape on twitter: classification and exploration of user categories. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, CSCW '12, pages 241–244.

Ahmed Ali Abdalla Esmin, Roberto L. De Oliveira Jr., and Stan Matwin. 2012. Hierarchical classification approach to emotion recognition in twitter. In *Proceedings of the 11th International Conference on Machine Learning and Applications, ICMLA, Boca Raton, FL, USA, December 12-15, 2012. Volume 2*, pages 381–385. IEEE.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. Liblinear: A library for large linear classification. *J. Mach. Learn. Res.*, 9:1871–1874, June.

Eric Gilbert and Karrie Karahalios. 2010. Widespread worry and the stock market. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*.

Suin Kim, JinYeong Bak, and Alice Oh. 2012a. Discovering emotion influence patterns in online social network conversations. *SIGWEB Newsl.*, (Autumn):3:1–3:6, September.

Suin Kim, JinYeong Bak, and Alice Oh. 2012b. Do you feel what i feel? social aspects of emotions in twitter conversations. In *International AAAI Conference on Weblogs and Social Media*.

Efthymios Kouloumpis, Theresa Wilson, and Johanna Moore. 2011. Twitter sentiment analysis: The good the bad and the omg! In *Proceedings of the Fifth International Conference on Weblogs and Social Media*.

Saif Mohammad. 2012. #emotional tweets. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics*, pages 246–255.

Olutobi Owoputi, Brendan OConnor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A. Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL-2013)*.

W. Gerrod Parrott, editor. 2001. *Emotions in Social Psychology*. Psychology Press.

Matthew Purver and Stuart Battersby. 2012. Experimenting with distant supervision for emotion classification. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 482–491.

Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Unsupervised modeling of twitter conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 172–180.

Kirk Roberts, Michael A. Roach, Joseph Johnson, Josh Guthrie, and Sanda M. Harabagiu. 2012. Empatweet: Annotating and detecting emotions on twitter. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 3806–3813. ACL Anthology Identifier: L12-1059.

Andranik Tumasjan, Timm O. Sprenger, Philipp G. Sandner, and Isabell M. Welpe. 2011. Election forecasts with twitter: How 140 characters reflect the political landscape. *Social Science Computer Review*, 29(4):402–418, November.

Xiaolong Wang, Furu Wei, Xiaohua Liu, Ming Zhou, and Ming Zhang. 2011. Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, CIKM '11, pages 1031–1040.

Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P. Sheth. 2012. Harnessing twitter "big data" for automatic emotion identification. In *Proceedings of the 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust*, SOCIALCOM-PASSAT '12, pages 587–592.

Changhua Yang, Kevin Hsin-Yih Lin, and Hsin-Hsi Chen. 2007a. Building emotion lexicon from weblog corpora. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 133–136.

Changhua Yang, Kevin Hsin-Yih Lin, and Hsin-Hsi Chen. 2007b. Emotion classification using web blog corpora. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, WI '07, pages 275–278.