Using Hashtags as Labels for Supervised Learning of Emotions in Twitter Messages

Maryam Hasan Emmanuel Agu Elke Rundensteiner

Department of Computer Science, Worcester Polytechnic Institute, Worcester, MA, United States
mhasan@wpi.edu, emmanuel@wpi.edu, rundenst@wpi.edu

ABSTRACT

Many college students experience depression or anxiety but do not seek help due to the social stigma associated with psychological counseling services. Automatic techniques to classify social media messages based on the emotions they express can assist in the early detection of students in need of counseling. Supervised machine learning methods yield accurate results but require training datasets of text messages that have been labelled with the classes of emotions they express. Manually labeling a large corpus of Twitter messages is labor-intensive, error prone and time-consuming. Hashtags are keywords inserted into social media messages by their authors. In this paper, we investigate using hashtags as emotion labels and evaluate them through two user studies, one with psychology experts and the other with the general crowd. The study showed that the labels created by general crowd was inconsistent and unreliable. However, the labels generated by experts matched with hashtag labels in over 87% of Twitter messages, which indicates that hashtags are indeed good emotion labels. Leveraging the concept of hashtags as emotion labels, we develop Emotex, a supervised learning approach that classifies Twitter messages into the emotion classes they express. We show that Emotex correctly classifies the emotions expressed in over 90% of text messages.

Keywords

Emotion Detection, Supervised Learning, Twitter Hashtags

1. INTRODUCTION

1.1 Background

The high rates of depression and anxiety among college students are well documented [7]. The American College Health Association 2012 survey detailed that 50% of students felt overwhelming anxiety, 30% felt so depressed that it was difficult to function, and 7% seriously considered suicide. Although 30-50% of students have diagnosable mental health conditions, only about 10% of students seek psychological support resources, due to the social stigma or lack of self-awareness. Untreated students experience lower productivity (GPA) and graduation rates, and in some cases loss of life.

Stealth methods of early detection are attractive so that students that possibly have emotional problems can be identified discreetly. Early detection is important because in many cases, serious cases of depression often start out mildly in the form of dysphoria. If caught and managed early, dysphoria can often be treated using simple treatments such as by increasing the physical activity levels or by improving the quality of sleep of patients.

Social networks such as Twitter allow individuals to express their opinions, feelings, and thoughts in the form of short text messages at any time of the day. These short messages (or tweets) explicitly or implicitly capture the emotional states (such as happiness, anxiety, and depression) of individuals as well as larger groups (such as the opinions of people in a certain country or affiliation) [2, 27].

1.2 Motivation

If the Twitter messages of communities (e.g., students in a college) can be automatically classified in real time, this method could be employed for a large variety of applications, ranging from studying emotions of populations, providing evidence for counseling services, and a wide spectrum of emotion management applications. Specifically, this technology could be used by counseling agencies to monitor and track a patient's emotional states, or to recognize anxiety or systemic stressors of populations. University counseling centers could be warned early about distressed students that may require further personal assessment. Counselors may then reach out to students flagged by such a system to confirm diagnoses and prevent deterioration. Widespread anxiety amongst students taking the same class could be automatically detected. Instead of treating only the students that seek counseling, the professor teaching the class could be approached proactively and presented with the evidence. Since Twitter is heavily used by young people, automatic methods to detect students who frequently express potentially harmful emotions on Twitter promises to be effective. While a holistic detection method would also incorporate other correlated emotion indicators such as poor sleep and lack of physical activities, in this work we focus on the analysis of text messages to detect emotional states.

Census bureaus and polling organizations could also use automatic emotion classification to estimate the percentage of people in a community experiencing certain emotions. Mood is an important indicator of well-being, which is typically measured using self-reports and surveys [11, 6]. People are asked to fill out questionnaires about their life and their day-to-day emotions. Collecting these questionnaires is not only time consuming, tedious, but also error-prone. Instead, a system that automatically detects emotions from text messages could cost-effectively detect what people feel about their lives from twitter messages. For example, the system could recognize:

- percentage of people with high levels of life satisfaction
- percentage of people who feel happy and cheerful versus those with depression
- percentage of people who feel calm and peaceful versus those who feel anxious

1.3 Challenges of Manual Labeling

Our overall goal is to classify each text message into several classes of emotion. Existing approaches can be grouped into two main groups: lexical methods and machine learning approaches [15]. Lexical methods classify the emotion expressed in a text based on the occurrence of certain words. Lexical methods are based on shallow word-level analysis and usually ignore many semantic features [15, 18] (e.g., they can fail to consider negation). Moreover they rely on an emotion lexicon, which is difficult to construct a comprehensive set of emotion keywords.

We adopt instead a supervised machine learning approach, which learns a classification model from human-labeled messages (i.e., training dataset), and then classifies unlabeled messages using that model. While supervised learning methods achieve high accuracy, they require a large corpus of texts that have been classified into the emotion classes they express (i.e., labeled data) [28].

Prior work on text emotion identification has mostly utilized manually annotated data. Crowdsourcing, in which humans manually infer and annotate each message with the emotions it expresses, is a popular approach for labeling data [4, 5, 20]. Crowdsourcing tools such as Amazon's mechanical Turk [1]facilitate access to large numbers of manual data labelers and annotators. Manually labeling Twitter messages with the emotions they express faces numerous challenges including:

1. *Tedious:* Manual annotation of emotion data by human experts is very tedious, labor-intensive and time-consuming.
2. *Semantic ambiguity:* Human emotions as well as the texts expressing them are ambiguous, which makes it difficult to accurately infer the author's emotional state.
3. *Casual style:* Twitter messages are written in a casual style, thus contain many grammatical and spelling errors along with slang words.
4. *Numerous topics and emotional states:* The large breadth of topics discussed on Twitter makes it challenging to manually create a comprehensive collection of labeled data that covers all emotional states.
5. *Inconsistent annotators:* In general, human annotators may not be reliable. A human annotator's judgement of the emotions in a text message is likely to be subjective and inconsistent. Consequently, different annotators tend to classify the same text message into different emotion classes, as confirmed by our user study in Section 2.

1.4 Proposed Approach: Use Hashtags as Emotion Labels

Due to the above described issues with manual labeling, researchers have started to investigate automatic methods for labeling training data. For example, Go et al. [12] and Pak and Paroubek [19] used Western-style emoticons as labels to classify Twitter messages as having either positive or negative sentiment [12, 19].

Authors of Twitter messages frequently mark keywords or topics by prefixing them with a hashtag (#) symbol. Originally, embedded hashtags were intended to make Twitter messages more searchable. Using the Twitter API we can automatically filter and query tweets by query terms or by hashtags. We now observe that, in many cases these hashtag keywords may also correspond to the author's own classification of the main topics of their Twitter message. We conjecture that hashtag keywords written by authors indicate the main emotion (emotion labels) expressed by the Twitter messages. For example, a tweet with the hashtag "#depressed" can be interpreted as expressing a depressed emotion, while a tweet containing the hashtag "#excited" as expressing excitement. The tweet "I need a holiday :-(#stressed", with the tag #stressed expresses a stressed mood. The tweet "Feelings hurt tonight! #sad" expresses sadness, while the tweet "Home made chicken soup is the best #happy" indicates happiness. *The key intuition here is that hashtags provide direct access to the author's emotional state, which is more accurate and trustworthy than the interpretations of a third-party annotator.*

The usage of hashtags in tweets is very common. For example, a study of a sample of 0.6 million tweets by Wang *et al.* [28] showed that 14.6% of tweets in their sample had at least one hashtag.

We build a large corpus of text messages for supervised learning by using embedded hashtags to label the emotions expressed by each text message. This approach requires no effort for manual labeling. It yields a completely automatic scheme for labeling a massive repository of Twitter messages. This approach could equally be applied in other Twitter mining applications where labeling is required.

However, there are challenges in using hashtags as emotion labels. Many tweets contain more than one hashtag,

[1] https://www.mturk.com/mturk/welcome

and some may even contain hashtags expressing opposing emotion classes. For example in the tweet "Got a job interview today with At&t... #nervous #excited.", hashtag #excited shows happiness, while tag #nervous shows stress. These tweets should be detected and removed from training dataset. In summary, we make the following major contributions in this paper.

- We propose a new labeling approach that uses hashtags to automatically label Twitter messages with the emotions they express.
- We validate the effectiveness of using hashtags as emotion labels by comparing hashtag labels with the labels assigned by humans (psychology experts as well as crowdsourced psychology novices).
- We compare the performance of supervised learning algorithms trained using proposed labeling approach, using different emotion lexicons.

2. EVALUATING THE EFFICACY OF HASHTAG LABELS

Our key research question is to determine whether human annotators would categorize texts into the same emotion classes selected by automatic labeling using hashtags. To answer this question, we performed two user studies in which two different classes of people participated. First experts in psychology (counselors and psychology graduate students) and then psychology novices (the crowd) were asked to classify texts into emotion classes.

2.1 Emotion Classes

To define emotion classes, we utilized the Circumplex model of affect [23], which defines emotions in a two-dimensional circular space with valence (i.e., pleasure) and arousal (i.e., activation) as axes. Based on this model we defined four classes of emotion namely *Happy-Active*, *Happy-Inactive*, *Unhappy-Active*, and *Unhappy-Inactive*, as shown in Figure 1. We chose the Circumplex model because it concisely describes emotions in a discrete two-dimensional space making it suitable for computational approaches in emotional research.

2.2 Comparing Hashtags with Crowdsourced Labels

This user study compares the accuracy of emotion labels that are generated automatically using hashtags with labels generated by non-expert annotators (the crowd).

2.2.1 *Experimental Methodology*

For this study, we selected a group of psychology novices as non-expert annotators. We design the study by randomly selecting 120 tweets (more precisely, 30 tweets from each emotion class) from our collected labeled tweets. See Section 4.1 for more details about our collected labeled tweet dataset. The tweets are shuffled to make their order random. Any embedded hashtags were removed from these 120 tweets. Then participants were asked to indicate the emotion expressed in each message, by selecting the pleasure level (high for happy or low for unhappy), and the arousal level (high for active or low for inactive).

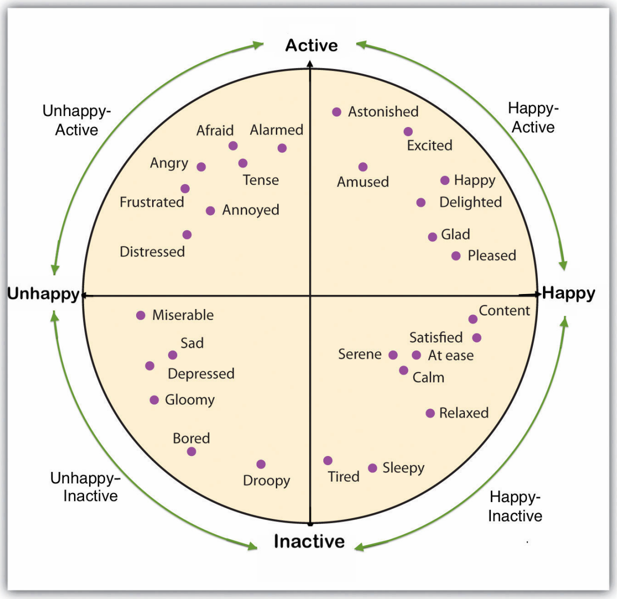

Figure 1: Circumplex Model of Affect including 28 affect words by J. A. Russell, 1980. [23]

Since the goal of this user study is to explore the labels made by non-experts, we recruited subjects from the student body of WPI (Worcester Polytechnic Institute) including students in an introductory psychology class at WPI. Our user study was run online using the Qualtrics [2] survey system between March and May 2014. 59 students participated, and 49 students completed the survey.

2.2.2 *Results and Analysis*

The interpretation of emotions expressed in texts tends to be subjective and diverse. As expected, inconsistencies occurred in the answers, such that in some cases different participants categorized the same text into different classes. Our analysis thus measures the level of agreement among the study participants.

First, we measure to what degree the annotators agreed on the level of pleasure or activation of each tweet. For this, we utilized Fleiss-Kappa [9], which is a popular statistical measure of the level of agreement between a fixed number of labelers in classifying subjects into known categories. Our Fleiss-Kappa measure of inter-labeler agreement for the pleasure level of tweets was 0.67, which corresponds to a substantial agreement. This value for the activation level was 0.25 which corresponds to a low level of agreement. In summary, although the annotators substantially agreed on the level of pleasure, there was a relatively low agreement among them for the level of activation. This conclusion can be explained by the fact that authors of text messages tend to express pleasure in explicit and unambiguous terms, which thus makes it easy to identify. For example, the tweet "Final weeks is going to be a death of me!" shows sadness. However it does't clearly indicate the level of arousal (i.e., activation).

[2] http://www.qualtrics.com

Figure 2 shows the pair-wise agreement between labelers on the level of activation and pleasure of tweets. The horizontal axis presents tweets and the vertical axis presents Pi,the extent to which labelers agree for the i-th tweet. This value computes how many labeler-labeler pairs are in agreement relative to the number of all possible pairs. As Figure 2 shows the Pi values are mostly less than one. 71% of Pi values of activation level are less than 0.7, which indicates low level of agreement. The figure also shows that pleasure has larger Pi values (i.e., higher levels of agreement) than activation.

The result of this study indicates that the labels created by non-experts to classify emotions expressed in Twitter messages are not sufficiently reliable. Thus casts doubt on the use of the crowd (such as via Amazon Mechanical Turk), for this particular task of emotion classification.

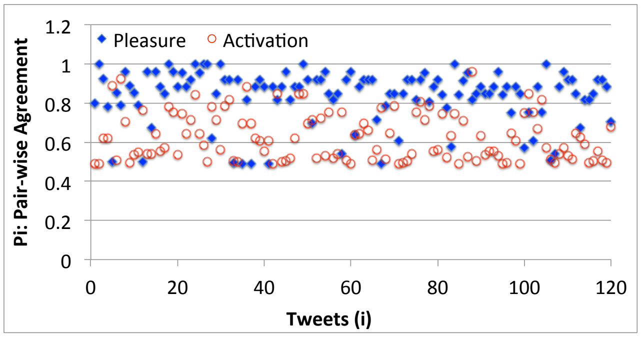

Figure 2: Pair-wise agreement between crowd labelers on level of activation and pleasure of tweets

2.3 Comparing Hashtags with Expert Labels

As Figure 2 indicates the level of agreement among crowd labelers is not adequate to consider them as ground truth especially for the activation level of messages.Instead we seek help from domain experts for labeling. We asked three experts to manually label 120 tweets (same tweets that had been utilized in Section 2.2). One of the experts is the director of counseling at WPI Student Development and Counseling Centre. The other two experts are graduate students in psychology who have been trained to identify emotions.

Figure 3 shows the pair-wise agreement between experts on level of activation and pleasure of tweets. The Fleiss-Kappa measure of agreement between experts for pleasure level of tweets is 0.84 which constitutes a perfect agreement. This value for activation level is 0.64 which shows a substantial agreement.

Table 1 lists the Fleiss-Kappa values of crowd labelers versus expert labelers. The agreement between experts is much higher than the agreement between crowd labelers. These results indicate that emotion labeling by trained experts is more reliable and can be utilized as the ground truth data. In contrast to expert labels, labeling by novices shows high inconsistencies. However, this leads to the challenge that if experts were asked for labeling, crowdsourcing would be expensive.

We now utilize the expert labels to evaluate the accuracy of hashtags. Table 2 lists the accuracy of hashtags based on expert labels. Comparing the hashtag labels with the expert labels for the 120 tweets indicates that the hashtag labels are the same as the expert labels in 100 tweets. There are 15 tweets that their hashtag labels are different from the expert labels. Also there is no consensus among experts about 4 tweets. Therefore, In over 87% of the cases, emotions indicated by hashtags embedded in tweets accurately captured the author's emotion indicated by ground truth (i.e., expert labels).

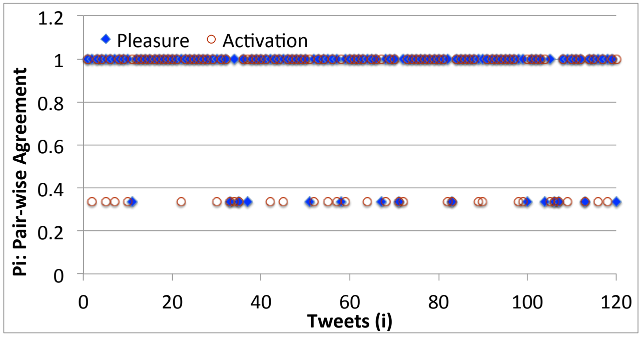

Figure 3: Pair-wise agreement between expert labelers on level of activation and pleasure of tweets

Labeler	Pleasure level	Activation level
Crowd Labeler	0.67	0.25
Expert Labeler	0.84	0.64

Table 1: Comparing Fleiss-Kappa values of crowd and expert labelers

Table 3 presents the tweets in which the hashtag label conflicts with the expert label. As the table shows most of the mismatches between hashtags and expert labels belong to the arousal level of tweets (i.e., active or inactive).

Expert	**Counseling Director**	**Trained Expert1**	**Trained Expert2**	**Experts Consensus**
Accuracy	81%	81%	84%	**87%**

Table 2: Accuracy of hashtag labels based on expert labels

3. EMOTEX: A SUPERVISED CLASSIFIER USING HASHTAG AS EMOTION LABEL

Since our user studies showed that hashtags are good indicators (labels) of the emotions in Twitter messages, we then set out to design a supervised learning approach that leverages this concept. A supervised learning approach learns a model from a large corpus of labeled messages, called training data. It then classifies unlabeled messages using the learned model. Supervised learning methods represent each message by an n-dimensional vector:

$$F = (f_1, ..., f_n) \in R^n$$

of numerical values. Clearly, the selection of relevant features to consider in the learning process plays an important role in such approaches. The features that describe the emotion expressed in text could for instance be unigrams (i.e., single words), bigrams (i.e., two consecutive words), punctuation features, emoticon features, or part-of-speech tags.

Hashtag Label	Expert Consensus	Tweet
Unhappy-Inactive	Unhappy-Active	I don't understand how people can stay in at weekends, it's half I'm already on the verge of suicide.
Happy-Inactive	Happy-Active	evening with Louis Armstrong...... means evening with great music
Happy-Active	Happy-Inactive	Finished writing my last paper for molecular genetics and now watching Deck the Halls
Unhappy-Inactive	Happy-Inactive	I have no life just on the Internet eating saltines and Nutella!!
Happy-Inactive	Unhappy-Inactive	I'm gonna do some life thinking. See you when the sun comes up
Unhappy-Inactive	Unhappy-Active	Finals week is going to be the death of me.
Happy-Inactive	Unhappy-Active	Im so sleepy but my brain obviously isnt tired at all as it keeps having these annoying dreams
Unhappy-Active	Happy-Active	Serving my Italian food to a special group tonite. Fingers crossed.
Unhappy-Inactive	Unhappy-Active	I'd like to give a big screw you to the river rats. I use to love it so much. Now it's just a pain.
Happy-Inactive	Happy-Active	Wonderful hot shower to finish off the night!
Happy-Inactive	Happy-Active	Morning on the mountain!
Unhappy-Inactive	Unhappy-Active	We live in a time where we're so busy competing instead of helping each other. genuine hearts are hard to find
Unhappy-Inactive	Unhappy-Active	Poor things will never be rich or famous. I would kill myself before I would want to have your sad, pathetic lives.
Unhappy-Inactive	Happy-Active	Miss you already!
Unhappy-Active	Unhappy-Inactive	I probably shouldn't have procrastinated the whole vacation on doing my homework
Unhappy-Inactive	Unhappy-Active	Think its quiet sad that someone can be unhappy enough to make up lies about other peoples families.

Table 3: Tweets with hashtag label different than expert label

Machine learning methods perform a deep analysis of text and are able to consider many features of the text. However, the feature space may have a very high number of dimensions. Single words are the most common features to be included in the feature vector. However, with the large breadth of topics discussed on Twitter, the number of words tends to be extremely large. Thus, the feature vector of each tweet would be very large and sparse (i.e., most features would have a value of zero).

To overcome this problem, we select an emotion lexicon as the set of unigram features. We have used and compared different emotion lexicons in our system, including ANEW lexicon (Affective Norms for English Words) [3], LIWC dictionary (Linguistic Inquiry and Word Count) [21], and AFINN [29]. Please see Section 4 for our detailed results. LIWC [21] contains a dictionary of several thousands words and prefixes, broken down into psychological categories. We use emotion-indicative categories including positive emotions, negative emotions, anxiety, anger, sadness, and negations. ANEW lexicon [3] contains 2477 English affect words. The interesting aspect of this lexicon is that each word is rated for its valence and arousal in a 1-9 scale. Each word has been labeled by several human labelers and the mean rating and standard deviation are given. ANEW and other word lists had been developed before the advent of microblog tools. AFINN [29] was constructed to include a new word list specifically for microblogs. Compared to ANEW, the AFINN word list contains more words including obscene words.

Beyond unigrams as baseline features, we also explored using emoticons and punctuations as features for emotion classification in text messages. To tackle the problem of negated phrases such as "not sad" or "not happy", we explicitly defined negation as a separate feature. We selected the list of negated phrases from the LIWC dictionary.

3.1 Automatic Labeling in Emotex

As mentioned before in Section 2.1, we defined four classes of emotions based on the level of valence and arousal. In order to label each message as one of the defined four emotion classes we needed to find the arousal and valence level of each message. As validated in Section 2, hashtags embedded by the authors of messages accurately indicate their emotions. We thus need to determine the arousal and valence of the hashtags at the end of the message. To achieve this, we use the ANEW lexicon which provides arousal and valence level of affect words (See Section 3).

One advantage of this approach is that the process of collecting labeled data based on emotion can be relatively easy since it avoids manual annotations. Twitter has an API that can be used to automatically collect tweets and filter them by query terms or by hashtags. This gives us the ability to collect a large number of tweets with various emotion hashtags, which can then serve as labeled data.

Another major advantage of this approach is that it gives us direct access to the author's own intended emotional state, without relying on the possibly inconsistent and inaccurate interpretations of third-party annotators.

4. EMOTEX EVALUATION

In this section, we present results of the experiments showing how well Emotex works.

4.1 Experimental Data

We used the Twitter dataset that we had collected in our previous work [13]. The tweets were collected for three weeks from December 26, 2013 to January 15, 2014 using the Twitter Streaming service. To improve the quality of the collected tweets, a set of heuristics rules were developed to eliminate noisy tweets (e.g., tweets with hashtags belong to opposing emotion classes). The collected tweets contained the emotion hashtags that indicate a distinct emotion. Table 4 lists the number of collected tweets before and after pre-processing. As it shows the number of tweets decreased by about 19% when removing noisy tweets during preprocessing.

For this experiment we labeled the tweets using proposed approach in Section 3.1. Then the hashtags were removed in order to force the classifier algorithm to learn from the other features of the tweets itself.

Class	#Tweets before Preprocessing	#Tweets after Preprocessing
Happy-Active	39600	34000
Happy-Inactive	41000	29200
Unhappy-Active	44000	37000
Unhappy-Inactive	40700	33900
Total	165300	134100

Table 4: Number of tweets collected as labeled data

The histogram in Figure 4 represents the distribution of happy and unhappy classes of tweets that we labeled using hashtags, during and after the new year vacation. It shows that the number of happy tweets during the vacation are higher than the number of happy tweets after vacation by about 12%, as expected. However the number of happy tweets didn't change significantly (only 1%) between one week after new year and two weeks after it.

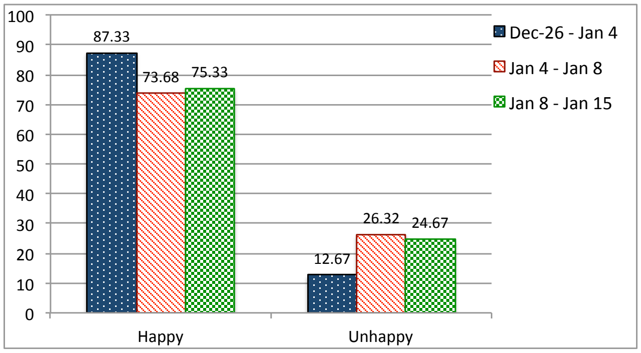

Figure 4: Distribution of happy emotions in collected tweets during and after the new year vacation.

Similarly, the histogram in Figure 5 represents the distribution of *active* and *inactive* classes of labeled tweets. It shows that the number of active tweets during the vacation are higher than the number of active tweets after vacation by about 3%.

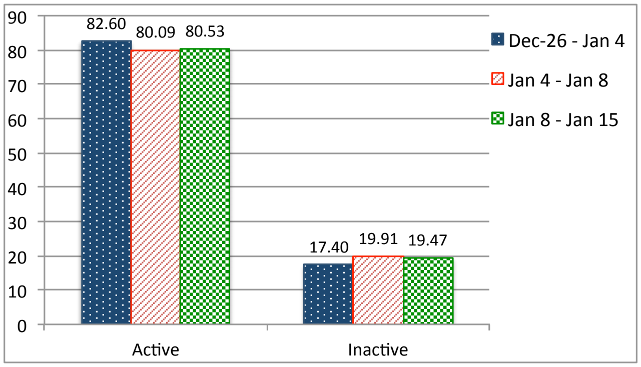

Figure 5: Distribution of the active emotions in collected data during and after the new year vacation.

As mentioned in Section 3, Emotex explores the usage of different features including unigrams, emoticons, punctuations, and negations. Table 5 lists the distribution of features present in the collected data after preprocessing.

Features	Number of tweets containing a feature	Percentage of tweets containing a feature
Happy-icon	5800	4.3%
Sad-icon	1320	1%
Angry-icon	1020	0.7%
Sleepy-icon	270	0.2%
Negation	9050	6.7%
Punctuation	19450	14.5%

Table 5: Distribution of Features in the collected data

4.2 Evaluating Emotex For Emotion Classification

Emotex Evaluation Using Different Machine Learning Models. Using labeled tweets, two well-known classification models namely SVM and K-Nearest Neighbor were created. KNN and SVM run fast and provide high accuracy. We used the WEKA [10] machine learning framework for KNN. We used the SVM-light [14] software with a linear kernel to learn the SVM classifier. SVM-light runs faster than SMO in WEKA.

For this evaluation we utilized LIWC lexicon [21]. Figure 6 presents the accuracy of SVM for each class using different features, based on 3-fold cross validation. Similarly, Figure 7 presents the accuracy of KNN for each class using different features. In both methods, the Happy-Active class achieved the highest accuracy and the Happy-Inactive class achieved the lowest accuracy. SVM achieved the total accuracy of above 89% using all features, and KNN achieved the highest total accuracy of above 90% using all features.

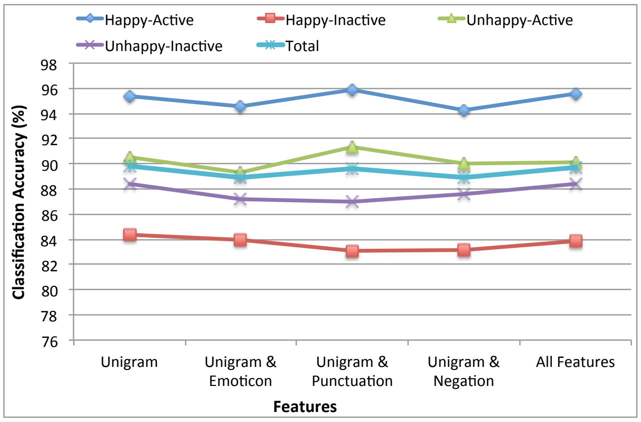

Figure 6: Classification accuracy of SVM for each class, using different features

Emotex Evaluation Using Different Emotion Lexicons. In order to evaluate the effect of different lexicons, we compared results achieved using three different lexicons including ANEW [3], LIWC [21], and AFINN [29]. Figure 8 presents the classification accuracy of SVM using all the features, when different lexicons are applied. As it shows, SVM achieved the highest accuracy using LIWC lexicon.

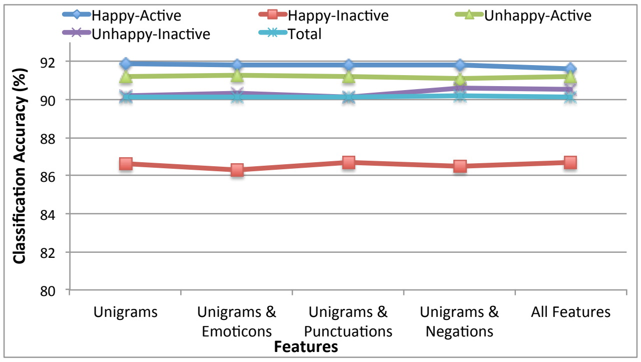

Figure 7: Classification accuracy of KNN for each class, using different features

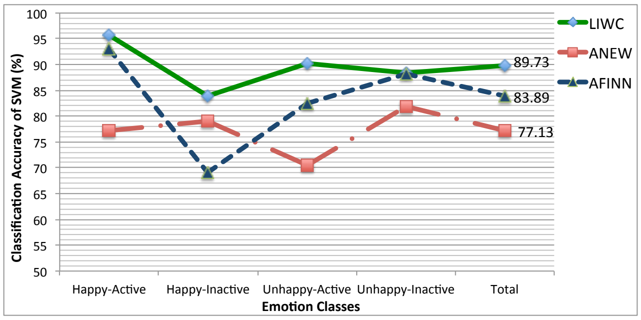

Figure 8: Classification accuracy of Emotex, using different lexicons

5. RELATED WORK ON EMOTION ANALYSIS IN TEXT

This section briefly surveys prior work on classifying emotion in text. Recently, researchers have explored social media such as Twitter to investigate the potential use of social media to detect depressive disorders. Park *et al.* [20] ran some studies to capture the depressive mood of users in Twitter. They studied 69 individuals to understand how their depressive states are reflected in their tweets. They found that people post about their depression and even their treatment on social media. Their results showed that participants with depression exhibited an increased usage of words related to negative emotions and anger in their tweets [20].

Another effort for emotion analysis on Twitter data accomplished by Bollen and his colleagues [2]. They tried to find a relationship between overall public mood and social, economic and other major events. They extracted 6 dimensions of mood (tension, depression, anger, vigor, fatigue, confusion) using an extended version of POMS (Profile of Mood States), a psychometric instrument. They found that social, political, cultural and economic events have a significant and immediate effect on the dimensions of public mood.

Most research on textual emotion recognition is based on building and employing emotion lexicons [16, 24, 17, 18, 25].

The lexical-based approach has been previously studied in the context of emotion classification. Ma *et al.* [16] searched WordNet for emotional words for all 6 emotional types defined by Ekman [8]. They then assigned weights to those words according to the proportion of Synsets with emotional association that those words belong to. Strapparava and Mihalcea [25] constructed a large lexicon annotated for six basic emotions: anger, disgust, fear, joy, sadness and surprise. They used linguistic information from WordNet Affect [26].

In another work, Choudhury *et al* [5] identified a lexicon of more than 200 moods frequently observed on Twitter. Inspired by the circumflex model, they measured the valence and arousal of each mood using mechanical turk and psychology literature sources. Then, they collected posts which had at least one of the moods in their mood lexicon as indicated by a hashtag at the end of a post.

Lexical methods are fast and intuitive as they are based on shallow word-level analysis. However, these methods can recognize only surface and predetermined features of the text. They usually ignore semantic features [15, 18] (e.g., they can fail to consider negation).

Some researchers applied supervised learning methods to identify emotions in text. For example, Choudhury *et al* [4] also detected depressive disorders by measuring behavioral attributes including social engagement, emotion, language and linguistic styles, ego network, and mentions of antidepressant medication. Then they leveraged these behavioral features to build a statistical classifier that estimates the risk of depression. They crowdsourced data from Twitter users who have been diagnosed with mental disorders. Their models showed an accuracy of 70% in predicting depression.

Purver *et al* tried to train supervised classifiers for emotion detection in Twitter messages, using automatically labeled data [22]. They used the 6 basic emotions identified by Ekman [8] including happiness, sadness, anger, fear, surprise and disgust. They used a collection of Twitter messages, all marked with emoticons or hashtags corresponding to one of six emotion classes, as their labeled data. Their method did better for some emotions (happiness, sadness and anger), than others (fear, surprise and disgust). Their overall accuracies (60%) were much lower than our accuracy.

6. CONCLUSION

We have proposed and evaluated the use of hashtags to automatically label a large corpus of Twitter messages with the types of emotions they express. To define the emotional states of users, we utilized the Circumplex model of human affect[23].

We validated hashtag-based labeling versus human labeling, by running an online user study in which psychology experts as well as novices were asked to label a random sample set of tweets. The Fleiss-Kappa results showed slightly high agreement among participants for valence level, however very low agreement for arousal level (See Table 2). These results confirm that human labeling of emotion using crowd sourcing is subjective, inconsistent, and thus unreliable. Human labeling by psychology experts showed higher agreement and could be useful but would be expensive. The expert labeling results showed above 87% accuracy for hashtag labels. These results confirm that hashtags of tweets are reliable features for automatic emotion labeling.

Based on the proposed hashtag method for labeling emotions in twitter messages, a system, Emotex was developed to classify Twitter messages. The supervised classifiers trained on labeled tweets, were able to achieve above 90% accuracy for multi-class emotion detection, while demonstrating robustness across different learning algorithms.

7. REFERENCES

[1] L. Barbosa and J. Feng. Robust sentiment detection on twitter from biased and noisy data. In *Proceedings of the 23rd ACL: Posters*, pages 36–44. Association for Computational Linguistics, 2010.

[2] J. Bollen, H. Mao, and A. Pepe. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *ICWSM'13*, 2011.

[3] M. M. Bradley and P. J. Lang. Affective norms for english words (anew): Instruction manual and affective ratings. Technical report, Citeseer, 1999.

[4] M. D. Choudhury, M. Gamon, S. Counts, and E. Horvitz. Predicting depression via social media. In *ICWSM*. The AAAI Press, 2013.

[5] M. De Choudhury, S. Counts, and M. Gamon. Not all moods are created equal! exploring human emotional states in social media. In *ICWSM*, 2012.

[6] S. O. Ed Diener Ed Diener. Subjective well-being: The science of happiness and life satisfaction.

[7] D. Eisenberg, M. F. Downs, E. Golberstein, and K. Zivin. Stigma and help seeking for mental health among college students. *Medical Care Research and Review*, 66(5):522–541, 2009.

[8] P. Ekman. Basic emotions. *Handbook of cognition and emotion*, 98:45–60, 1999.

[9] J. L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

[10] E. Frank, M. Hall, G. Holmes, R. Kirkby, B. Pfahringer, I. H. Witten, and L. Trigg. Weka-A Machine Learning Workbench for Data Mining. In O. Maimon and L. Rokach, editors, *Data Mining and Knowledge Discovery Handbook*, chapter 66, pages 1269–1277. Springer US, Boston, MA, 2010.

[11] B. S. Frey and A. Stutzer. What can economists learn from happiness research? *Journal of Economic literature*, 40(2):402–435, 2002.

[12] A. Go, R. Bhayani, and L. Huang. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, pages 1–12, 2009.

[13] M. Hasan, E. Rundensteiner, and E. Agu. Emotex: Detecting emotions in twitter messages. In *SocialCom-Stanford, May 2014*. ASE, 2014.

[14] T. Joachims. Making Large-Scale SVM Learning Practical. In B. Schölkopf, C. J. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, Cambridge, MA, USA, 1999. MIT Press.

[15] H. Liu, H. Lieberman, and T. Selker. A model of textual affect sensing using real-world knowledge. In *Proceedings of the 8th international conference on Intelligent user interfaces*, pages 125–132. ACM, 2003.

[16] C. Ma, H. Prendinger, and M. Ishizuka. Emotion estimation and reasoning based on affective textual interaction. In *Affective computing and intelligent interaction*, pages 622–628. Springer, 2005.

[17] A. Neviarouskaya, H. Prendinger, and M. Ishizuka. Textual affect sensing for sociable and expressive online communication. In *Affective Computing and Intelligent Interaction*, pages 218–229. Springer, 2007.

[18] A. Neviarouskaya, H. Prendinger, and M. Ishizuka. Affect analysis model: novel rule-based approach to affect sensing from text. *Natural Language Engineering*, 17(1):95–135, 2011.

[19] A. Pak and P. Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may 2010. European Language Resources Association (ELRA).

[20] M. Park, C. Cha, and M. Cha. Depressive moods of users portrayed in twitter. In *Proc. of the ACM SIGKDD Workshop on Healthcare Informatics, HI-KDD*, 2012.

[21] J. W. Pennebaker, M. E. Francis, and R. J. Booth. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, page 71, 2001.

[22] M. Purver and S. Battersby. Experimenting with distant supervision for emotion classification. In *Proceedings of the 13th EACL*, pages 482–491. Association for Computational Linguistics, 2012.

[23] J. A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39:1161–1178, 1980.

[24] C. Strapparava and R. Mihalcea. Semeval-2007 task 14: Affective text. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 70–74. Association for Computational Linguistics, 2007.

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

[26] C. Strapparava and A. Valitutti. Wordnet affect: an affective extension of wordnet. In *Proceedings of 4th International Conference on Language Resources and Evaluation, LREC*, volume 4, pages 1083–1086, May 2004.

[27] M. Thelwall, K. Buckley, and G. Paltoglou. Sentiment in twitter events. *Journal of the American Society for Information Science and Technology*, 62(2):406–418, 2011.

[28] X. Wang, F. Wei, X. Liu, M. Zhou, and M. Zhang. 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*, pages 1031–1040. ACM, 2011.

[29] F. Årup Nielsen. A new anew: evaluation of a word list for sentiment analysis in microblogs. In M. Rowe, M. Stankovic, A.-S. Dadzie, and M. Hardey, editors, *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages*, volume 718 of *CEUR Workshop Proceedings*, pages 93–98, May 2011.