Happy, Nervous or Surprised?
Classification of Human Affective States in Social Media

Munmun De Choudhury **Michael Gamon** **Scott Counts**
Microsoft Research, One Microsoft Way, Redmond, WA 98051, USA
{munmund, mgamon, counts}@microsoft.com

Abstract

Sentiment classification has been a well-investigated research area in the computational linguistics community. However, most of the research is primarily focused on detecting simply the polarity in text, often needing extensive manual labeling of ground truth. Additionally, little attention has been directed towards a finer analysis of human moods and affective states. Motivated by research in psychology, we propose and develop a classifier of several human affective states in social media. Starting with about 200 moods, we utilize mechanical turk studies to derive naturalistic signals from posts shared on Twitter about a variety of affects of individuals. This dataset is then deployed in an affect classification task with promising results. Our findings indicate that different types of affect involve different emotional content and usage styles; hence the performance of the classifier on various affects can differ considerably.

Introduction

Social media tools including Twitter have been gaining significant traction of late in emerging as platforms of human sentiment and affect expression. Sentiment and affect analysis can be useful in a number of scenarios, including marketing campaigns, monitoring responses to local and global happenings, and deciphering geographic and temporal mood trends. As social media tools become increasingly ubiquitous, such analyses of affect can also enable new information-seeking approaches; for instance, identifying search features given an affect attribute. Consequently, there is significant value to be derived from *predicting and classifying human affect* in social media.

In sentiment analysis research, two broad, general factors – typically labeled Positive Affect (PA) and Negative Affect (NA) – have emerged reliably as the dominant dimensions of emotional experience, as classification categories or as ways to measure public sentiment. However for a more elaborate understanding of emotional behavior of individuals in social media, it is imperative to account for not only these two general dimensions of affect and sentiment, but more distinguishable and fine-grained affective states. Such states would reflect the specific content, language and state of the individual sharing the content, i.e., the distinctive qualities of individuals' affects, beyond simply the valence (positivity/negativity) of the affect descriptors.

However classifying affective states in social media domains presents itself with several challenges. First, the variety of linguistic styles and emotional interpretations of millions of individuals makes it difficult to infer affective states concretely; at the same time constructing consistent features across shared content is challenging. Second, most standard machine learning techniques rely on availability of labeled data for training – an aspect often circumvented via manual labeling of ground truth. As we move to social media domains featuring enormous data, coupled with unavailability of ground truth, gathering appropriate training examples necessitates a scalable alternative approach.

Our major contribution lies in the development of an affect classifier of social media data, that does not rely on any hand-built list of features or words, except for the near 200 mood hashtags that we use as a supervision ground truth signal. We are motivated by findings in the psychology literature in inferring 11 different affective states of individuals in Twitter posts. For this purpose, we derive a mapping of affective states from explicit mood words used as hashtags at the end of posts, via a series of mechanical turk studies. These affect-labeled posts are then used in a maximum entropy classification framework to predict the affective state, given a post. Our experimental results indicate a wide variation in classifier performance across different affects – perhaps as a consequence of the diversity in usage patterns and linguistic styles across different affective states as well as the content sharing process.

Prior Work

Considerable prior research has been directed towards automatic classification of sentiment in online domains (Pang et al., 2002). These machine learning techniques need extensive manual labeling of text for creating ground truth. Some of these issues have been tackled by utilizing emoticons present in text as labels for sentiment (Davidov et al., 2010), although they tend to perform well mostly in the context of the two basic positive and negative affect classes. The closest attempt towards multiclass classification of sentiment has been on LiveJournal blog data, wherein the

mood tags associated with blog posts were used as ground truth labels (Mishne, 2005).

An alternative that circumvents the problems of machine learning techniques has been the use of generic sentiment lexicons such as WordNet, LIWC, and other lists (Esuli et al., 2006). Recently, there has been a growing interest in crowdsourcing techniques to manually rate polarity in Twitter posts (Diakopoulos et al., 2010). However these manually curated word lists are likely to be unreliable and scale poorly on noisy, topically-diverse Twitter data.

Finally, another problem with polarity-centric sentiment classifiers is that they typically encompass a *vague* notion of polarity that includes mood, emotion, and opinion; and lumps them all into two classes "positive" and "negative" (or "positive", "negative" and "neutral"), refer (Wiebe et al., 2004). In order to better make sense of emotional behavior on social media, we require a principled notion of "affect" – a central contribution of this work.

Affect in Social Media

Affect refers to the experience of feeling or emotion and is a key part of the process of an individual's interaction with environmental stimuli. The primary challenge in classifying affect lies in the unavailability of ground truth. The psychology literature indicates that there is an implicit relationship between the externally observed affect and the internal mood of an individual (Watson et al., 1988). When affect is detected by an individual (e.g., smile as an expression of *joviality*), it is characterized as an emotion or mood. In the rest of this section, we, therefore, discuss how we arrive at a representative list of mood-indicative words as well as affective states, and thereafter our mechanism of mapping moods to affects.

Representative Moods and Affects

We utilized five established sources to develop a mood lexicon that was eventually used to define affect classes:

1. ANEW: ANEW (Affective Norms for English Words) that provides a set of normative emotional ratings for ~2000 words in English (Bradley and Lang, 1999).
2. LIWC: For LIWC, we used sentiment-indicative categories like positive/negative emotions, anxiety, anger and sad (http://www.liwc.net/).
3. EARL: Emotion Annotation and Representation Language dataset that classifies 48 emotions (http://emotion-research.net/projects/humaine/earl).
4. A list of "basic emotions" provided by (Ortony and Turker, 1990), e.g., fear, contentment, disgust etc.
5. A list of moods provided by the blogging website LiveJournal (http://www.livejournal.com/).

However, this large ensemble of words is likely to contain several words that do not necessarily define a mood (e.g., *sleepy* is a state of a person, rather than a mood). To circumvent this issue, we first performed a mechanical turk study (http://aws.amazon.com/mturk/) to narrow our candidate set to truly mood-indicative words. In our task, each word had a 1 – 7 Likert scale (1 indicated not a mood at all, 7 meant definitely a mood). Only turkers from the U.S. and having an approval rating greater than 95% were allowed. Combining 12 different turkers' ratings, we constructed a list of those words where the median rating was at least 4, and the standard deviation was less than or equal to 1. Finally, based on feedback from two researchers, we performed one final filtering step on these words, eliminating moods that proved to be very ambiguous between true mood indicators and sarcasm or evaluative judgments. The final set of mood words contained 172 terms.

We then proceeded towards identifying representative affects. Although affect has been found to comprise both positive and negative dimensions (PANAS – positive and negative affect schedule (Watson et al., 1994)), we are interested in more fine-grained representation of human affect. Hence we utilize a source known as PANAS-X (Watson et al., 1994). PANAS-X defines 11 specific affects apart from the positive and negative dimensions – '*fear*', '*sadness*', '*guilt*', '*hostility*', '*joviality*', '*self-assurance*', '*attentiveness*', '*shyness*', '*fatigue*', '*surprise*', and '*serenity*'. We utilize these 11 affects in our classification process.

Inferring Mood to Affect Associations

Next, based on the mapping of moods to affects provided in the PANAS-X literature, we derived associations for 60% moods from our final lexicon of 172 words. For the remaining associations, we conducted a second turk study. Each turker was shown a set of 10 mood words and the set of 11 affects were listed with each. The turker was asked to select from the list the most appropriate affect that described the particular mood. We thus collected 12 ratings per mood. Finally, we combined the ratings per mood, and used the affect that received majority rating to correspond to it (Fleiss-Kappa for inter-rater reliability was 0.7).

Table 1. Associations of (sample) moods to five affects using the PANAS-X source and mechanical turk study.

Sample Moods	Associated Affect
Ecstatic, amused, festive, happy, jolly	*Joviality*
Afraid, defensive, terrified, nervous	*Fear*
Depressed, shattered, troubled, upset	*Sadness*
Shocked, bewildered, perplexed	*Surprise*
Calm, relieved, peaceful, gentle	*Serenity*

This way we collated a list of 172 moods where each mood corresponded to one type of affect (Table 1). Note that for the sake of simplicity, we consider that a mood can be associated with exactly one affect. The distribution of number of moods over affects is shown in Table 2.

Data Collection for Classification

For data collection, we utilized the Twitter Firehose that is made available to us via our company's contract with Twitter. We focused on one year's worth of Twitter posts posted in English from Nov 1, 2010 to Oct 31, 2011. From this ensemble, in the absence of ground truth affect labels on Twitter posts, we resorted to a method that could infer labels reasonably consistently and in a principled manner. We conjecture that posts containing moods as hashtags at the end are likely to capture the emotional state of the individual, in the limited context of the post. This is motivated by prior work where Twitter's hashtags and smileys were used as labels for sentiment classifiers (Davidov et al., 2010). For instance, *"#iphone4 is going to be available on verizon soon! #excited*" expresses the mood '*excited'*, which can subsequently be mapped to the affect *joviality* based on the association derived in the previous section.

Using this technique, we collected a large ground truth dataset where each post contained one of the 172 mood words as a hashtag at the end. We utilized the mapping obtained in previous section on the associations between the 172 moods and 11 affects, so that we ended up with a dataset of affect-labeled posts (6.8 million posts). Finally, we eliminated RT (retweet) posts, because there may be cases where a mood hashtag is added at the end to comment on the retweet – which is arguably different from the author associating a mood with the language they produce.

Table 2. Number of moods associated with the affects.

Affect	#moods	Affect	#moods
Joviality	30	*Fear*	14
Fatigue	19	*Guilt*	5
Hostility	17	*Surprise*	8
Sadness	38	*Shyness*	7
Serenity	12	*Attentiveness*	2
Self-assurance	20		

Classification and Experimental Findings

We use a classification setup that is standard in text classification as well as in sentiment classification. We represent Twitter posts as vectors of unigram and bigram features. Before feature extraction, the posts are lowercased, numbers are normalized into a canonical form, and URLs are removed. Finally the posts are tokenized. After feature extraction, features that occur fewer than five times are removed in a first step of feature reduction. We then randomly split the data into three folds for cross-validation. Features are subsequently reduced to the top 50K features in terms of log likelihood ratio, as a second feature reduction step. The classification algorithm is a standard maximum entropy classifier (Ratnaparkhi, 1998); we do not perform systematic parameter tuning, but select parameter values based on prior performance on various sentiment classification tasks. For each fold, we deploy this classifier to predict the affect labels of the test portion of the fold (33.3%), after training on the training portion (66.6%) of the fold.

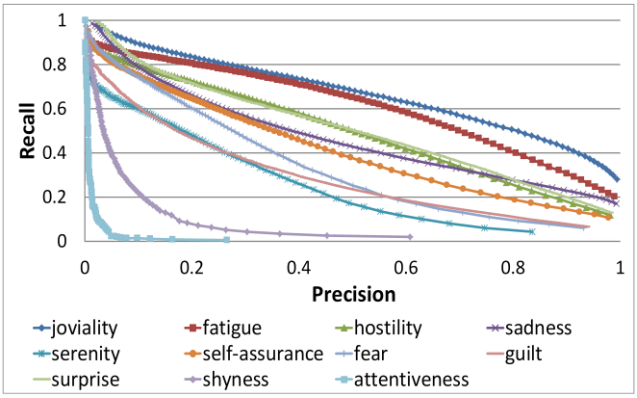

Figure 1. Precision-recall curve for the 11 class affect classification on Twitter.

We begin by discussing the performance of classifying the Twitter posts in our dataset into the 11 different affect classes. We report the mean precision and recall across the three folds of cross-validation in Figure 1. Our results show that the performance (precision/recall) of various affect classes differs widely. To better understand these differences, we report the mean F1 measures for the 11 affect classes in Table 3.

The best performances are observed for the affects *joviality*, *fatigue*, *hostility* and *sadness*, while the worst are for *guilt*, *shyness* and *attentiveness*. We also observe mediocre performances in the cases of *self-assurance* and *fear*.

Table 3. Mean F1 measures of 11 affect classes.

Affect class	Mean F1	Affect class	Mean F1
Joviality	0.4644	*Fear*	0.2319
Fatigue	0.4146	*Guilt*	0.1838
Hostility	0.3270	*Surprise*	0.3328
Sadness	0.2885	*Shyness*	0.0722
Serenity	0.1833	*Attentiveness*	0.0203
Self-assurance	0.2642		

Noting the mood distributions for the various affects in Table 2, it appears that the good performance can be explained by the fact that all of *joviality*, *fatigue*, *hostility* and *sadness* have a large number of moods – consequently their feature space may be less sparse, spanning a variety of topical and linguistic contexts in Twitter posts. On the other hand, the worst performing classes, e.g., *guilt*, *shyness* and *attentiveness*, are also the ones with fewer corresponding moods. Hence it is possible that their feature spaces are rather sparse due to the limited contexts they are typically used in on Twitter. Moreover a qualitative study of the posts that belong to these classes tend to indicate significant degrees of sarcasm or irony in them – e.g., for the *guilt* affect class: "*I hate when ppl read too deep into*

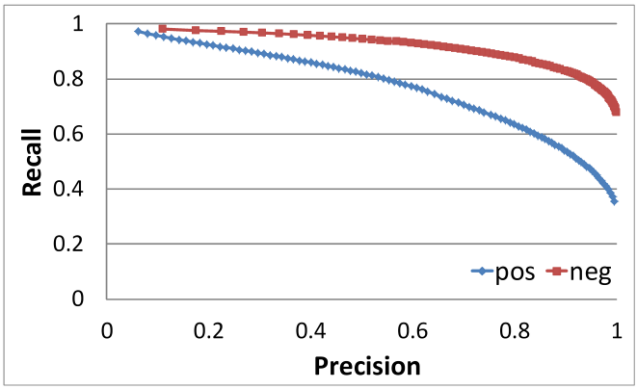

Figure 2. Precision-recall curve for two-class affect classification on Twitter (11 affects mapped to positive/negative affect).

ur tweet and think it's about them..... damn .. #guilty"; and for the *attentiveness* affect class: "*If a tomato is a fruit does that mean ketchup is a smoothie? #suspicious*". Due to such contextual mismatch between content and the labeled affect, the classifier performs worse for these classes.

However we tend to observe some exceptions in the cases of the affects *serenity* and *surprise* (Table 3) – *serenity*, despite having a moderate number of moods associated with it, tends to perform worse than others, while *surprise* in spite of having a very limited number of moods, performs relatively better. To investigate this, we conduct an experiment to better understand the relationship of these affects (in terms of post content) with respect to a "background" model of Twitter posts. We begin with a set of random posts without mood words: we call this the *background model of posts* – indicative of generic chatter on Twitter. For the affect classes *serenity* and *surprise*, we compute the Jensen-Shannon (J-S) divergence on unigram probabilities with respect to the background model. The J-S divergence for *serenity* is found to be 0.13 while that for *surprise* is 0.09. These numbers indicate that *surprise* has a usage pattern that is closer to the background model than *serenity* – consequently, despite having fewer moods, the feature space of *surprise* is not very sparse, helping its classification performance, compared to *serenity*.

What the classification results indicate in general is that, the manner in which the various affect classes are used on Twitter (via explicit mood hashtags) has a significant impact on the performance of the classifier. Moreover, it is well-established that different moods have different 'valence' and 'arousal' measures (e.g., *angry* and *frustrated* are both negative moods, but *angry* indicates higher arousal than *frustrated*). These differences make the context of affect manifestation widely diverse – in turn making affect classification in social media a challenging problem.

Because of these inherent differences in affect classes, we conduct a final experiment on the conventional polarity classification problem – PA (positive affect) and NA (negative affect). We map all of the 172 mood words into PA and NA, instead of the 11 affect classes. Using the same classifier as before, we show the precision-recall curves for the two-class affect classification in Figure 2. Our classifier yields good results in this case – the mean F1 for PA is 0.59, while that for NA is 0.78. This validates our methodology of using mood hashtags in posts as a mechanism to infer affect. It also indicates that while polarity classification (PA/NA) might be a relatively easier task, fine-grained affective states present with numerous challenges in light of classification – their diversity in terms of usage patterns, mood association, and language and style differences.

Conclusion

In this paper we proposed a novel way towards classifying different affective states of individuals in social media. Motivated by literature in psychology, we characterized human affect on Twitter via 11 classes, and used explicit mood words as hashtags at the end of posts to be supervising signals for inferring affect. We used this dataset in a maximum entropy classification framework. Our findings illustrated that different affective states have a wide range of usage patterns, as well as exhibit diversity in the linguistic context they are shared. We believe that investigating implicit factors – e.g., network structure and information sharing behavior in light of improving affect classification is one particularly interesting direction for future research.

References

Bradley, M.M., & Lang, P.J. (1999). Affective norms for English words (ANEW). Gainesville, FL. *The NIMH Center for the Study of Emotion and Attention*.

Davidov, D., Tsur, O., & Rappoport, A. (2010). Enhanced Sentiment Learning Using Twitter Hashtags and Smileys. *Proceedings of the 23rd International Conference on Computational Linguistics Posters*, (August), 241-249.

Diakopoulos, N., & Shamma, D. (2010). Characterizing debate performance via aggregated twitter sentiment. In *Proc. CHI 2010*. 1195-1198.

Esuli, A., & Sebastiani, F. (2006). SentiWordNet: A publicly available lexical resource for opinion mining. *Proceedings of LREC* (Vol. 6, p. 417–422).

Mishne, G. (2005). Experiments with mood classification in blog posts. In Style2005 -- the 1st Workshop on Stylistic Analysis Of Text For Information Access, at SIGIR 2005.

Ortony, A., & Turner, T. J. (1990). What's basic about basic emotions? *Psychological Review*, 97, 315-331.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. In *Proc. EMNLP 2002*, Vol. 10. 79-86.

Ratnaparkhi, Adwait (1993). Maximum Entropy Models for natural language ambiguity resolution. PhD thesis, University of Pennsylvania.

Watson, D., & Clark, L. A. (1994). The PANAS-X: Manual for the positive and negative affect schedule-Expanded Form. Iowa City: University of Iowa.

Wiebe, J., Wilson, T., Bruce, R, Bell, M., & Martin, M. (2004). Learning subjective language. *Computational Linguistics*, 30 (3).