Distinguishing affective states in weblog posts

Michel Généreux and Roger Evans
Natural Language Technology Group (NLTG)
University of Brighton, United Kingdom
{M.Genereux,R.P.Evans}@brighton.ac.uk

Abstract

This short paper reports on initial experiments on the use of binary classifiers to distinguish affective states in weblog posts. Using a corpus of English weblog posts, annotated for mood by their authors, we trained support vector machine binary classifiers, and show that a typology of affective states proposed by Scherer's et al is a good starting point for more refined analysis.

Introduction

We are investigating the subjective use of language in text and the automatic classification of texts according to their subjective characteristics, or 'affect'. Our approach is to view affective states (such as 'happy', 'angry') as regions of Osgood's Evaluation-Activation (EA) space (Osgood, Suci, & Tannenbaum 1957), and draws on recent work in psychology to construct a typology of such affective states (Scherer, Dan, & Flykt 2006), Our overall aim is to determine the extent to which such a typology can be validated and applied to the task of text classification using automatic methods. In this paper we describe some initial experiments aimed at validating a basic two dimensional classification of weblog posts, using Support Vector Machine (SVM) binary classifiers. The domain of weblogs is particularly well-suited for this task given its highly subjective nature and the availability of data[1], including data which as been author-annotated for 'mood'.

Research into binary automatic sentiment classification can now almost reach 90% accuracy under specific conditions (Mullen & Collier 2004). These results allow us to move on to multi-class classification of affective states. Recent attempts to do this for weblog posts have shown modest, but consistent improvements over a 50% baseline, only slightly worse than human performance (Mishne 2005). One important milestone to be achieved is the elaboration of a typology of affective states. Work in psychology is currently looking into this aspect (Scherer, Dan, & Flykt 2006), and some results can already be re-used, such as Scherer et al.'s typology of affects shown in Figure 1. In this model, sentiment classification is simply a matter of distinguishing between the left and right halves of the space. The experiments we describe here seek to go beyond this basic distinction in two ways: first we divide the space into four quadrants, delineated by the vertical and horizontal axes; second, we explore the signficance of the spacial aspect of the diagram – do points close together behave more similarly than points far apart?

Experimental method

We have collected from Livejournal[2] a total of 346723 (mood-annotated by authors) weblog posts in English, from which almost half can be clearly identified as belonging to one of the four quadrant. In the following experiments we used a subset of this corpus (15662, 54940, 49779 and 35634 documents in quadrants Q1 to Q4 respectively, numbered clockwise from the top-right quadrant in Figure 1).

Our aim was to investigate to what extent the typology can serve as a basis for locating particular affects in the space. Our first hypothesis is that the classification of two disjoint (sets of) moods should yield a classification accuracy significantly above a baseline of 50%. Our second hypothesis is that the more geometrically distant two given moods are according to the typology, the more accurate a classifier is for these two moods, since each mood does not share many features with the other. In other words, the classification *accuracy* and the geometric *distance* between moods exhibits a positive correlation. We conducted a series of experiments using machine learning to classify weblog posts according to their mood, each class corresponding to one particular quadrant or an individual mood: annoyed, frustrated (in Q1); anxious, sad (in Q2); calm, hopeful(in Q3); excited, amused (in Q4). Using Support Vector Machines (Joachims 2001) with a few basic features (unigrams, POS and stems) and random samples of our corpus (1000 testing examples, 2000 and 4000 training examples for individual moods, plus 8000 and 16000 for each quadrant), we tested the binary classification. The set of features used varies for each of these tasks, they were selected by inspection of each (distinct) training data set. Two simple criteria were used for selecting features: no stop words and a number of occurences of at least

[1]For example, a 1.6GB corpus is now freely available for research at `http://ilps.science.uva.nl/Resources/blogmods/`

[2]`http://www.livejournal.com`

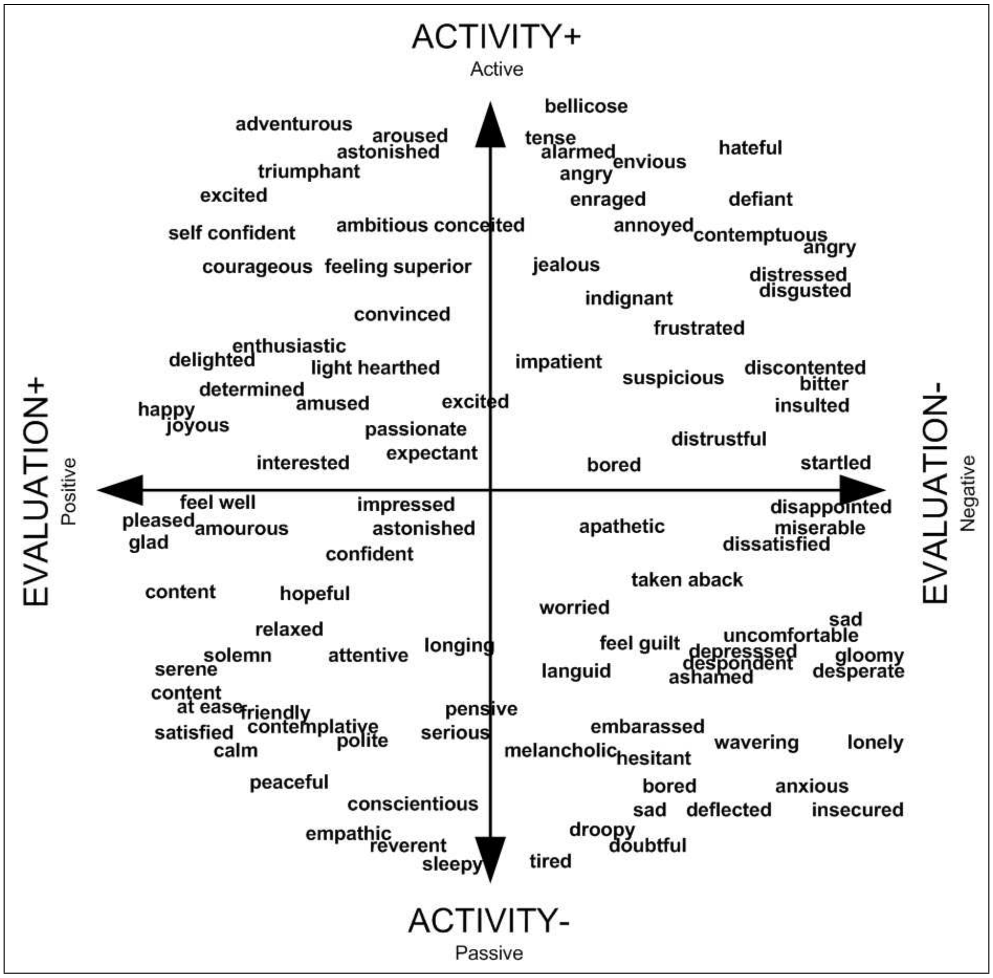

Figure 1: Typology of affects from Scherer et al.

3 in the training data.

Results

Our first hypothesis is that, if the four quadrants proposed by Scherer are a suitable arrangement for affective states in the EA space, a classifier should perform significantly better than chance (50%). Table 1 shows the results for the binary classification of the quadrants. By micro-averaging accuracy, we obtain at least 60% accuracy for the four binary classifications of the quadrants. Our second hypothesis predicts a correlation between accuracy and distance in the EA space. Table 2 summarizes the results for all classification tasks carried out: the *accuracy* column represents the average over all sets of training examples, in ascending order, while the *distance* is a geometric measure of the distance between two moods (or the center of gravity for each quadrant, assuming a uniform distribution of moods in the random sample). The *correlation* column is the cumulated Pearson correlation coefficient. For instance, from 57.93% accuracy up to 65.35% accuracy, *accuracy* and *distance* displays a correlation coefficient of 0.423.

BINARY CLASSIF.	Accuracy
Q1 vs Q2+Q3+Q4	67.40%
Q2 vs Q1+Q3+Q4	59.20%
Q3 vs Q1+Q2+Q4	56.50%
Q4 vs Q1+Q2+Q3	59.60%
Micro-accuracy	60.68%
Random	49.10%

Table 1: Distinguishing quadrants: Training 2000 and Testing 1000

Analysis of results

Figure 1: how likely is it to classify 1000 weblogs and obtain at least 60% accuracy only by chance? A simple calculation using a binomial distribution (with X=600 successes, n=1000 trials and probabilty(p) =0.5) yields to a probability of 0.0000000001[3]. Therefore, it is safe to conclude that the typology is very helpful in improving the performance of a

[3] $P(X>600, n=1000, p=0.5) = 1-P(X \leq 600) = 1-0.9999999999 = 0.0000000001$

BIN. CL.	$\overline{Acc.}$	Dist.	Corr.
Q3 vs Q1+2+4	57.93%	50	-
Act vs Pass	58.23%	60	1.000
Q2 vs Q3	58.65%	50	-0.096
Ann vs Frustr	58.80%	15	-0.666
Anx vs Calm	59.10%	90	0.204
Q3 vs Q4	59.83%	50	0.086
Calm vs Hop	61.10%	30	-0.252
Q2 vs Q1+3+4	61.28%	50	-0.192
Exc vs Anx	62.75%	100	0.334
Q4 vs Q1+2+3	63.35%	50	0.230
Q2 vs Q4	64.35%	60	0.228
Pos vs Neg	64.98%	60	0.227
Anx vs Q4	65.30%	95	0.387
Calm vs Exc	65.35%	75	0.423
Q1 vs Q2	66.53%	50	0.332
Ann vs Anx	68.30%	90	0.434
Calm vs Q1	69.20%	90	0.504
Exc vs Q2	69.35%	70	0.489
Ann vs Calm	70.05%	105	0.568
Ann vs Q3	70.30%	80	0.574
Q1 vs Q3	70.58%	60	0.521
Q1 vs Q2+3+4	70.98%	50	0.441
Q4 vs Q1	72.43%	50	0.360
Anx vs Sad	73.35%	15	0.162
Exc vs Ann	75.00%	45	0.095
Exc vs Amu	75.65%	15	-0.047

Table 2: Measuring correlation: Act = Active, Pass = Passive, Ann = Annoyed, Frustr = Frustrated, Anx = Anxious, Hop = Hopeful, Exc = Excited

classifier into four distincts groups of affective states. These results show that the abstraction offered by the four quadrants in the model seems correct. This is also supported by the observation that the classifier shows no improvements over the baseline if trained over a random selection of examples in the entire space.

Figure 2: interpreting the results with regards to our second hypothesis is more difficult. For most of the way while moving upward along the *accuracy* dimension, the geometric *distance* between the two classes correlates positively. However, the general trend turns out to be a slightly negative correlation. We believe there are mainly two factors that can contribute to the distortion of the nature of the correlation between *accuracy* and *distance*:

non-uniform distribution of moods We have assume that the random selection of moods within a quadrant was uniform, i.e. covering the whole of the space uniformely. This is perhaps too optimistic, because moods that appear more often in the corpus are likely to be over-represented and therefore move the center of gravity of the quadrant away from the center;

ambiguity of a mood location in the space There are two moods represented in the experiments which appear at more than one location in the typology: *excited* and *sad*. We have attempted to minimize the impact of such ambiguity in the results by taking a location in the middle, but this may have undesirabble side-effects.

The obvious solution to the first problem is to make sure that we populate the training data with moods uniformly distributed in the space. If further research reveals that there are further subcategory within a single mood, then a solution to the second issue would be to create a distinctive tag for each subcategory.

Future work

The next series of experiments should provide a more refined validation of the typology and some insights into the use of semantic orientation scores for the purpose of multi-mood classification of texts.

Whilst we have validated the proposed partitioning of the EA space for four regions as a whole, we now need to validate, for each quadrant, a more precise location for a number of individual, uncontroversial moods. This can be achieved by narrowing down the set of moods to those for which work in psychology exhibits a certain level of certainty with regards to their location, while at the same time are in agreement with our second hypothesis.

The long term aim of this work is to classify texts according to their mood, making it possible to classify a text according to more than two (positive vs negative or active vs passive) classes (angry, sad, happy, etc.). To achieve that, we will use directly the location of the affective states in the space in combination with automatic continuous “scores” (in the sense that they provide only a qualitative evaluation) of the E and A axis. Scores such as PMI-IR (Turney & Littman 2003) provide a real value for each axis. They can be converted to “measures” (in the sense that they offer a basis for comparison) between -1 and +1 by substituting the real PMI-IR scores for every semantically oriented phrase in a text with +1 or -1, and averaging over all phrases. By mapping the E and A measures to normalized values in the EA space, specific coordinates can be obtained and associated with surrounding affective states.

References

Joachims, T. 2001. *Learning to Classify Text Using Support Vector Machines*. Kluwer Academic Publishers.

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*.

Mullen, T., and Collier, N. 2004. Sentiment analysis using support vector machines with diverse information sources. In *Empirical Methods in NLP*.

Osgood, C. E.; Suci, G. J.; and Tannenbaum, P. H. 1957. *The Measurement of Meaning*. University of Illinois.

Scherer, K.; Dan, E.; and Flykt, A. 2006. What determines a feelings position in three-dimensional affect space? A case for appraisal. Cognition and Emotion. (in press).

Turney, P. D., and Littman, M. L. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Trans. Inf. Syst.* 21(4):315–346.