2004

Discerning Emotions in Texts

Victoria L. Rubin
Syracuse University, vlrubin@syr.edu

Jeffrey M. Stanton
Syracuse University, jmstanto@syr.edu

Elizabeth D. Liddy
Syracuse University, liddy@syr.edu

Discerning Emotions in Texts

Victoria L. Rubin*, Jeffrey M. Stanton, Elizabeth D. Liddy*

School of Information Studies
*Center for Natural Language Processing
Syracuse University
Syracuse, NY 13244-1190
{vlrubin, jmstanto, liddy}@syr.edu

Abstract

We present an empirically verified model of discernable emotions, Watson and Tellegen's Circumplex Theory of Affect from social and personality psychology, and suggest its usefulness in NLP as a potential model for an automation of an eight-fold categorization of emotions in written English texts. We developed a data collection tool based on the model, collected 287 responses from 110 non-expert informants based on 50 emotional excerpts (min=12, max=348, average=86 words), and analyzed the inter-coder agreement per category and per strength of ratings per sub-category. The respondents achieved an average 70.7% agreement in the most commonly identified emotion categories per text. The categories of high positive affect and pleasantness were most common in our data. Within these categories, the affective terms "enthusiastic", "active", "excited", "pleased", and "satisfied" had the most consistent ratings of strength of presence in the texts. The textual clues the respondents chose had comparable length and similar key words. Watson and Tellegen's model appears to be usable as a guide for development of an NLP algorithm for automated identification of emotion in English texts, and the non-expert informants (with college degree and higher) provided sufficient information for future creation of a gold standard of clues per category.

Introduction

In the Natural Language Processing (NLP) community subjectivity is defined as "aspects of language used to express opinions and evaluations" (Wiebe 1994, Wiebe 2000, Wiebe et al. 2001). There is increasing interest in, and need for text corpora annotated for subjectivity. Previous annotation schemes have been devised for a binary classification of objective sentences vs. subjective; and for positive and negative attitude types of subjective content (Wilson and Wiebe 2003). A finer grained distinction of emotions is needed for distinguishing evaluations in information extraction, summarization and question-answering applications. In this paper we describe a data collection facility for classification of evaluative subjective texts into a two dimensional map of emotion types with the assistance of non-expert raters.

This study combines theory and research methods from social and personality psychology with NLP analytical techniques as a possible approach for isolating, quantifying, and qualitatively describing emotions that readers identify in written texts. Psychology provides an empirically verified model of discernable emotions, the Watson and Tellegen's Circumplex Theory of Affect (1985).

A Circumplex Theory of Affect

Before one attempts to automatically discern emotions in written texts, it is important to ascertain whether people generally agree on the types of emotions identifiable in texts. Figure 1 depicts the essence of Watson and Tellegen's Circumplex Theory of Affect.

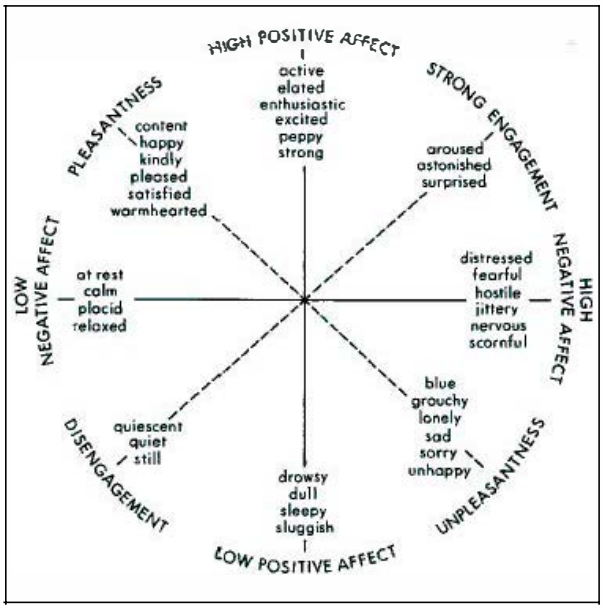

Figure 1. Watson and Tellegen's two-dimensional map. Reprinted as fair use excerpt from Watson and Tellegen (1985).

In their research on the structure of affect, Watson and Tellegen consistently encountered the same two major bipolar dimensions: positive affect and negative affect.

Positive affect reflects a combination of high energy and positive evaluation characterized in such emotions as elation. Negative affect comprises feelings of upset and distress (Watson and Tellegen 1985). Both positive and negative affect occur on bipolar continua, ranging from high to low. Note, however, that many affectively loaded words are not pure markers of either factor (see Figure 1).

Nonetheless, the choice of words to reflect each octant was based on a substantial body of research on self-report measures of experienced emotion (Tellegen et al. 1999). The pleasantness octant contains terms representing a mixture of high positive and low negative affect (e.g. pleased, satisfied). Unpleasantness includes combinations of high negative and low positive affect (e.g. blue, grouchy). Pleasantness-unpleasantness and strong-engagement-disengagement axes form an alternative orientation on the circumplex (imagine the figure turned clockwise by 45 degrees). Inclusion of two alternative rotations gives the map a circular appearance. This eight-octant categorization was the foundation of our study.

Research Questions

Using the theoretically specified structure for affect, we pose the following research questions:

1) Do people agree in discerning the types of emotions in texts, and if so, to what extent?
2) What are the most commonly identified emotion octants (for our sample data)?
3) Do people agree on ratings of affective terms represented in our model as sub-categories of the emotions' octants, and if so, to what extent?
4) If people agree on an emotion represented in a given excerpt, are they also consistent in selecting textual clues for the chosen emotion?

Data Collection

To answer these questions, we developed an on-line data collection facility for use in two phases of research: a pilot test and a main collection effort. Phase 1 was a pretest that assessed the usability of the data collection facility. Phase 2 included our full textual data set and a large pool of participants that were asked to discern emotion types, assess strength of emotions' presence in text, and select associated textual clues.

Textual Data. The database comprised 100 excerpts (min=12, max=348, average=86 words; 679 sentences) selected from several publicly available on-line sources. We used web-log and customer review genres as examples of texts that are typically written from the first person and are rich in a variety of emotional colorations. Writers' styles, grammar, and phrasing were preserved as much as possible, except for minor editing (e.g., name changes, pronoun substitutions, title paraphrases) to protect the writers' identities.

Length of Excerpts. Excerpts ranged from 2 to 21 sentences in their entirety as delimited by the writer, or by our segmentation of the texts closest to the 20-sentence limit (e.g., paragraph marks). Ongoing discussion in the NLP literature considers whether a sentence should be the unit of analysis for subjectivity marking. For instance, Wiebe et al. (2001) preferred sentence-level as a compromise between a fine grained expression-level analysis with higher inter-coder reliability, and a text-level analysis likely to have more noise. Our participants analyzed emotions on a segment level, and had freedom to specify the length of clues that influenced their ratings.

Data Collection Process. Each Phase 1 pre-tester was asked to look at minimum of 5 excerpts, and "think aloud" as they went through the tasks during a 30 - 45 minute session with one of the researchers.

For both phases, participants read an excerpt, identified emotions within the Watson and Tellegen's Circumplex, and provided their ratings of the perceived strength of presence the emotions on seven-point Likert scales. Next, participants cut-and-pasted the textual clues that influenced their ratings of the type and strength of emotion. Multiple emotion categories could be identified for the same excerpt. In the pre-test phase, the participants had the option of providing comments related to each excerpt.

Study Participants. In Phase 1, only 6 informants participated, 2 males and 4 females in educational occupations, 26 - 40 age group. For Phase 2, we invited four hundred informants from the StudyResponse online participant panel (http://www.Study Response.org) and 109 of them (46 males, 44 females, and 9 unidentified) participated in the study. Twenty seven participants were between 18 and 30 years of age, 33 between 31-40, 52 between 41-55, 8 were above 55 years old, and 9 did not identify their age. Due to the complexity of the tasks, all of the informants (but 3) that were invited, had at least a four year college education. Sixty had their master's or professional degrees, and 16 had their doctorate degrees.

Data Collection Tool. The online data collection tool consisted of the demographic page which recorded each participant's demographic data, and the main page with two simultaneously displayed panes (http://istprojects.syr.edu / ~StudyResponse / emotion / index.asp). The left-hand pane allowed participants to select one of the emotion octants related to the text shown in the right-hand pane. The right-hand pane also contained the Likert scale ratings for the strength of presence of octants' sub-categories, for instance, such affective terms like aroused, astonished, surprised for strong engagement octant (see Figure 1). It also had the clue collection box, and relevant user instructions. The interface provided an

option for identifying another emotion in the same excerpt or proceeding to a new text. The data collection database tracked elapsed time and recorded answers for each participant and for each analyzed text.

Data Analysis

Methodology. Phase 1 pre-test was used for an assessment of the online facility's functionality, the participants' comprehension of the task, the quality of the user interface, and the logical flow of data collection tasks. The pre-testers agreed that the instructions and tasks were clear enough, but made several usability suggestions.

In Phase 2, we addressed our research questions. We calculated informants' level of agreement on the most commonly selected emotion octant per each sample excerpt out of all submitted responses per each excerpt, and took an average of that agreement rate.

We found most commonly identified emotion octants for our sample data, based on overall counts of placements of excerpts into the most commonly agreed upon octants.

Next, we filtered out all the excerpts that had less than 3 person agreement on most commonly identified emotion octant, and analyzed the agreement within octants, based on affective terms and their ratings. Each emotion octant had 3 to 6 affective terms, or sub-categories, that the respondents rated with their strength between 1, emotion being strongly present, and 7 – absent. We calculated standard deviations of composite emotion ratings, averaged and compared the deviations of each affective term (or sub-category) within each emotion octant.

Finally, we looked at the most agreed upon excerpts and analyzed the consistency and lengths of clues selected by the informants.

Results and Discussion

In Phase 2, average agreement rate among the respondents on emotion octants categorization, based on most commonly selected emotion octant, was 70.7%, with 21.5% standard deviation. Fifteen texts had equal or lower than 50% agreement per text and should not be used as sample data for automation. Further investigation for reasons of such low inter-coder agreement should be conducted. The remaining 35 texts are useful for creation of an emotion-mining algorithm.

Of the 8 emotion octants, high positive affect and pleasantness were chosen 80 and 77 times. The octants of strong engagement and low negative affect were only chosen 32 and 22 times, respectively. The remaining octants – high negative affect, unpleasantness, disengagement, and low positive affect, were pointed to 17, 15, 11, and 6 times, respectively. This speaks to how common different categories of emotions were encountered in the data.

Table 1 depicts the amount of agreement expressed in standard deviation within pleasantness and high positive affect octants. Low standard deviations from a composite average rating suggest high agreement, and high standard deviation suggest higher interpersonal variability in either the interpretation of the sample texts, or in respondents' understanding of the affective terms. For instance, for high positive affect, the term enthusiastic was well agreed upon (*SD*=0.66 from composite rating), while "peppy" was interpreted with a great variability of *SD*=1.41.

Sub-Categories within High Positive Affect	Standard Deviations from Composite Likert Scale Ratings	Sub-Categories within Pleasantness Affect	Standard Deviations from Composite Likert Scale Ratings
active	**0.90**	content	1.06
elated	1.00	happy	1.26
enthusiastic	**0.66**	kindly	1.34
exited	**0.90**	**pleased**	**0.79**
peppy	1.41	**satisfied**	**0.83**
strong	1.43	warmhearted	1.58

Table 1. Average standard deviations of sub-category ratings in two most identified groups: high positive affect and pleasantness. The lower standard deviations are (in bold), the more agreement between respondents in ratings per sub-category there is.

In the final analysis, we looked at the sample text clues selected by the respondents in 28 texts where more than 3 people agreed upon the same emotion octant. For example, in one text 7 people agreed on the same octant. The text is rendered in full below:

> *She gives me hugs, nice hugs. The other day we were sitting on the bed waiting for something and she started crawling towards me, and she got most of the way and then just reached her arms out all the way, which just touched me, and threw herself forward onto me and hugged me, resting her head on me . Sometimes when I am doing things she will scuttle herself over to me and pull on my pants, or touch my leg. Hug me, Mummy. Sometimes when she feels like chatting, but doesn't know what to say, she just softly says Mummmeeee, mummmeee, mumeeee. I like that. <excerpt ID=1032>*

Throughout the dataset, the clue length varied from a few words to a sentence, and rarely the whole excerpt. Table 2 shows the length and choices of textual clues submitted by the 7 respondents. The clues show a lot of textual similarity. The most consistent clue word in the excerpt was the noun "hugs" sometimes accompanied by its adjectival modifier "nice". The combination was reported 6 out of 7 times. The verb "like" was reported 4 out 7 times. The consistency and persistence of this example indicates that for this particular text, non-trained

respondents were able to consistently agree on the most important clue for classifying the text into the pleasantness octant. What is particularly interesting in this case is that neither of the clues are exact same words as the sub-category headings in the pleasantness octant, and yet they are easily discernable by non-expert respondents.

Selected Clues	ItemID
nice hugs, touched me, softly, *like*	84
hugs, **nice hugs**	99
hugs, which just touched me, I *like* that	173
softly **nice hugs** I *like* that	206
She gives me **hugs**, **nice hugs.** I *like* that.	213
She gives me **hugs, nice hugs**	244
She gives me **hugs**, **nice hugs**	328

Table 2. Seven clues that helped 7 respondents to identify the sample text 1032 as belonging to the pleasantness octant.

In some excerpts, someone else's emotional experience (other than the excerpt writer's) was reported. The bold text in the excerpt below provides an example:

> *Yesterday was a very sombre affair indeed. I can't believe that it was six months to the day since those planes crashed into the World Trade Centre and the Pentagon.* ***Everyone seemed to have a solemn look on their faces and went about their daily business quietly -*** *it was very strange indeed. <excerpt ID=1012>*

Other researchers have noted similar issues of nested sources in subjective texts (Wilson and Wiebe, 2003). As the pre-tests showed, we effectively sidestep this problem by clearly indicating to the participants to identify and assess the emotions of the author of the excerpts, and not of a third party. Further, the designation of clue words or phrases by the participants disambiguated the source stimulus of the emotion ratings.

Conclusions and Future Research

The Watson and Tellegen's Circumplex Theory of Affect is useful as a guide for development of an NLP algorithm for an automated identification and an eight-fold categorization of emotion in texts, an Emotion-Miner. A sample of fifty texts was categorized by 110 respondents according to the Watson and Tellegen's model. The average inter-coder agreement on the most commonly categorized emotion octants per text was 70.7%. The texts with the inter-coder agreement ≤ 50% should not be used for an algorithm creation.

High positive affect and pleasantness octant were found to be most common in our sample data. Within those octants, enthusiastic, active, excited, pleased, and satisfied had the highest ratings agreement for the strength of presence of those particular affect sub-category terms.

We compiled a list of linguistic expressions identified as clues for each of the eight categories. In the future research, we intend to perform lexical, semantic, and syntactic analyses for patterns and regularities to answer the question of how each of the eight categories of affect is expressed linguistically. The results will become the basis of the Emotion-Miner algorithm.

The study benefits both the NLP and psychological research communities in terms of bridging previously unrelated research. It enhances understanding of the theory of perceived structure of emotions in written texts, and provides empirical data that link emotions to linguistic clues identified by a large number of non-expert participants.

Acknowledgements

We are grateful to our colleagues at Dr. Nakagawa's Language Informatics Laboratory, Information Technology Center at the University of Tokyo, and Dr. Noriko Kando and her doctoral students at the National Institute of Informatics in Tokyo, Japan for their valuable comments.

References

Tellegen, A., Watson, D., Clark, L.A., 1999. On the dimensional and Hierarchical Structure of Affect. Psychological Science, Vol. 10, No 4, 297-303.

Watson, D., Tellegen, A. 1985. Toward a consensual structure of mood. Psychological Bulletin, 98, 219-235.

Wiebe, J. M. 1994. Tracking point of view in narrative. Computational Linguistics 20 (2): 233-287.

Wiebe, J. 2000. Learning Subjective Adjectives from Corpora. Proc. 17th National Conference on Artificial Intelligence (AAAI-2000). Austin, Texas, July 2000.

Wiebe, J., Bruce, R., Bell, M., Martin, M., Wilson, T. 2001. A Corpus Study of Evaluative and Speculative Language. Proc. 2nd ACL SIGdial Workshop on Discourse and Dialogue. Aalborg, Denmark, September, 2001.

Wilson, T., Wiebe, J. 2003. Annotating Opinions in the World Press. 4th SIGdial Workshop on Discourse and Dialogue (SIGdial-03). Sapporo, Japan, July 2003.