Gill, A.J., Gergle, D., French, R.M., and Oberlander, J. (2008). Emotion Rating from Short Blog Texts. *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 2008)*.

Emotion Rating from Short Blog Texts

Alastair J. Gill*, Darren Gergle*, Robert M. French†, Jon Oberlander‡

*Center for Technology
and Social Behavior
Northwestern University,
2240 Campus Drive,
Evanston, IL 60208, USA
{alastair|dgergle}@northwestern.edu

†LEAD-CNRS UMR 5022,
University of Burgundy
Pôle AAFE, BP 26513,
Dijon 21065, France
robert.french@u-bourgogne.fr

‡School of Informatics,
University of Edinburgh
2 Buccleuch Place,
Edinburgh, EH8 9LW, UK
J.Oberlander@ed.ac.uk

ABSTRACT

Being able to automatically perceive a variety of emotions from text alone has potentially important applications in CMC and HCI that range from identifying mood from online posts to enabling dynamically adaptive interfaces. However, such ability has not been proven in human raters or computational systems. Here we examine the ability of naive raters of emotion to detect one of eight emotional categories from 50 and 200 word samples of real blog text. Using expert raters as a 'gold standard', naive-expert rater agreement increased with longer texts, and was high for ratings of joy, disgust, anger and anticipation, but low for acceptance and 'neutral' texts. We discuss these findings in light of theories of CMC and potential applications in HCI.

Author Keywords

Computer-mediated communication, emotion, affect, language.

ACM Classification Keywords

H.5.m [**Information Interfaces and Presentation (HCI)**]: Miscellaneous.

J4 [**Social and behavioral systems**]: Psychology.

INTRODUCTION

Face-to-face or on the phone, people can often guess a speaker's emotion 'just from their tone of voice': that is, without being able to identify the words being used, let alone their specific meanings (for an overview, see [4]). But would we ever want to rely on words alone – without using information from the speech signal? In some cases we have to: computer-mediated communication (CMC), email, text-chat and websites all offer reduced media richness. Knowing the emotional tone of comments circulating about one's company can be useful business intelligence. Blogs (or personal weblogs) around the world can discuss a company's performance one day, and perhaps influence its share price the next. Some tools exist which look at usage of mood terms in blog posts, analyzing large amounts of text to capture national responses to news or sporting events [2]. These tools measure emotion at a very coarse (e.g., national) level, however, often greater specificity is needed: Smaller text segments reflecting particular opinions may need to be extracted and classified for opinion or emotion. Indeed, detecting emotion from short sections of text may facilitate the development of technologies to automatically detect emotion in email clients or in a friend's recent blog posts. Eventually, user interfaces which can automatically detect and adapt to user emotion may be possible.

Additionally, there is an empirical question regarding the text-based communication of emotion, with different theories proposing varying degrees to which it is possible to understand social information, such as emotion in a computer-mediated environment. One extreme perspective put forward in Social Presence Theory [13] is that less rich environments, such as text-based CMC environments, inhibit communicating emotional expression. While in much richer environments (face-to-face) in which intonation and non-verbal cues are available, interlocutors are able to communicate a full range of emotional and interpersonal information due to greater social presence.

Alternatively, another theory (Social Information Processing, SIP; [14]) proposes that interpersonal cues, such as emotional information, are present in computer-mediated environments, but it just takes longer to derive the same information. Therefore, in a CMC environment with potentially unlimited time, interlocutors would be expected to derive the same perceptions as is possible in face-to-face communication, either by placing greater emphasis on existing cues (linguistic features), or by developing new strategies such as emoticons.

Following a recent increase of interest in the area of emotions, opinions, and their classification [6,15], one recent study has made significant first steps in addressing these questions: Hancock, Landrigan, & Silver [8] asked

CHI 2008, April 5–10, 2008, Florence, Italy.

participants in a text chat environment to express either positive (happy) or negative (unhappy) emotions to their naive conversational partner without explicitly describing their (projected) emotional state. Naive judges (the text-chat partners) could accurately perceive their interlocutor's emotion, and were less likely to enjoy or want to meet the authors of negative messages relative to positive ones. Additionally, a linguistic analysis of the transcripts found that authors portraying positive emotion used a greater number of exclamation marks, and used more words overall, whereas authors' texts portraying negative emotion used an increased number of affective words, words expressing negative feeling, and negations. Punctuation features matched the self-reported strategies used by the portrayers of emotion, with this regarded as evidence for the Social Information Processing hypothesis [8].

In this paper, we further explore the text-based communication of emotion in CMC, and build upon Hancock et al. in three main ways: (1) we expand their classification of emotion from positive and negative into the eight main categories as proposed by the literature [5,11]; (2) rather than focus on extended interactions, we examine whether emotion can be accurately classified on the basis of asynchronous short blog text extracts of 50 and 200 words, derived from (3) real emotional blogs (not actors). Previous work has shown that perception of personality is possible using 'thin slices' of email texts [1,7].

METHOD

Participants

The 65 judges of emotion were students at a Scottish university (23 males, 42 females, mean age = 22.24 years). Debriefing revealed all were frequent email users (mean score 6.30 on a 0-7 Likert scale); conversely, very few used blogs frequently (mean score 1.38), with 33 participants never using blogs. All were naive raters of emotion. An additional two participants are excluded from this analysis because they provided multiple or unclear responses.

Materials

The target blog texts used for emotion rating were taken from a previously collected corpus in which authors contributed their writings from one pre-specified month [10]. Blog lengths ranged from a few short postings to near daily posts of a few thousand words. Authors granted permission for further use of each blog before collection.

From our blog corpus, an 'expert' research assistant (not involved in subsequent rating) selected the first 200 words of each post if they contained some emotional content or were 'neutral', that is, apparently contained no emotion. This yielded 135 text extracts totaling 27,000 words. We do not have author ratings of emotions for the blog texts. Therefore for each extract, expert raters rated these texts as expressing one of eight emotions (anticipation, acceptance, sadness, disgust, anger, fear, surprise, joy) or neutral. Five expert raters who had extensive exposure to the texts were used: three had been closely involved with collecting and analyzing the original blog corpus; two experts were recruited from the university experimental participant pool, and familiarized themselves with the blog texts before their expert rating task. After all experts had assigned an emotional category rating to each of the 135 texts, 20 were selected as expressing strong and clear emotional content. This was based on all expert raters agreeing on the emotion assigned, and having the strongest emotion rating (two texts for each emotion, and four for 'neutral').

For each of these 20 texts we use two versions in the subsequent analysis: For the long version, we retain all 200 words; for the short version we extract the middle 50 words of the 200 word text. Note that in doing so, we ignore the sentence boundaries and so the start and finish of the sections may occur mid-sentence. This avoids bias resulting from the experimenters selecting texts which they believe contain features important for emotion expression. Comparison of long and short text versions revealed consistency of language and topic.

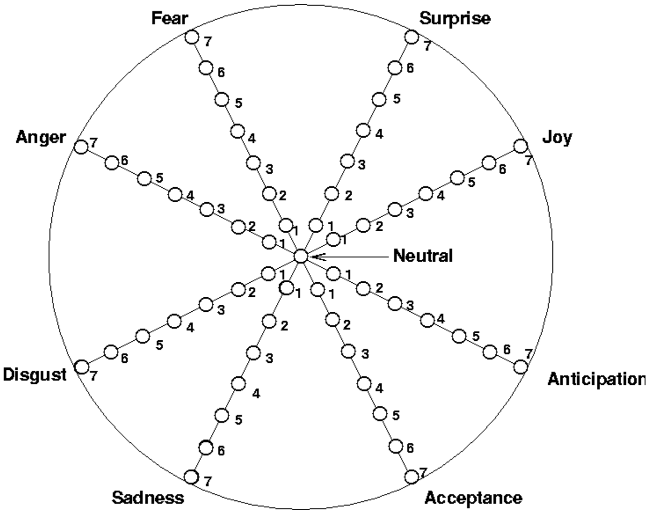

Figure 1: Emotion wheel used for text rating.

Procedure

All 20 texts were presented to the naive raters of emotion in random order. Participants were randomly assigned to two groups in a counter-balanced design: One group saw 10 long versions, then 10 short versions of the texts; The second group saw 10 short, then 10 long text versions.

All participants used the activation-evaluation wheel shown in Figure 1 [5,11]. Imagining x- and y-axes: Evaluation (valence) is on the x-axis, with positive values on the right, and activity on the y-axis, with high activity at the top. The strength of emotion corresponds to the distance from the center of the circle (between 1 and 7), with the center of the circle used to score 0 or 'neutral' emotion. This model is considered well-suited to computational work [3], has previously been used for rating emotion in speech [9], and allows comparison with findings for valence [8]. Alternative approaches to emotion are described in [4,12].

In the rating instructions, the judges were asked to rate 'how they perceive the author's emotions' but 'not to spend too long thinking about their answer, as we are particularly interested in [their] initial response'. All ratings took less than 30 minutes, and were combined with another (non emotion) text rating task not reported here.

Analysis

Nominal logistic regression was run on the emotion judgment data. We ignored the strength of emotion rating (1-7), simply coding expert-naive rater agreement as a binary value (agreement=1; disagreement=0; we leave analysis of emotional intensity to future work), and entered as a dependent variable. Text emotion (surprise, joy anticipation, acceptance, sadness, disgust, anger, fear, or neutral), and Text length (long, short), were entered into the equation as categorical variables; and an Expert text emotion × Text length interaction variable was included. A participant variable was included to account for individual judge biases. We avoid drawing conclusions from the ratings of 'neutral' texts, given the lower probability of assignment due to the emotion wheel design.

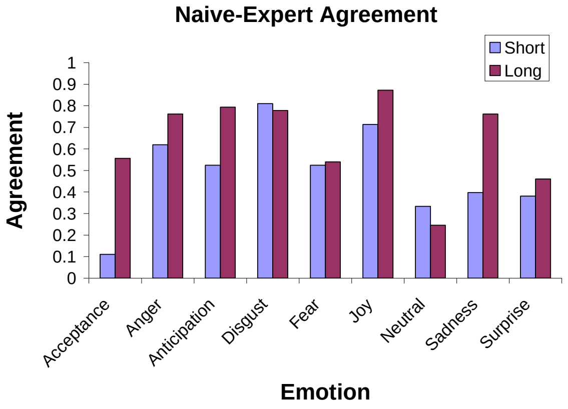

Figure 2: Agreement by Emotion

RESULTS

Figure 2 illustrates expert-naive judge agreement as a percentage (for comparison, chance agreement for the categories is approximately 0.11 or 11%). We note that there were significant interactions between Text length and Naive-Expert judgments for Acceptance (χ^2=13.6, p<.001[1]), Sadness (χ^2=5.85, p=.015, and to a lesser extent, Fear (χ^2 4.11, p=.043), where the agreement was higher for longer texts. For Disgust (χ^2=5.50, p=.019) we note the opposite interaction, with Naive-Expert judge agreement decreased for longer texts (also the case for Neutral, χ^2=18.45, p<.001).

Overall, there is a significant main effect of Text length, with longer texts leading greater Naive-Expert judge agreement (χ^2=32.5, p<.001), even though in both long and short samples there appears to be no difference in language or topic. Turning now to the effect of text emotion on Naive-Expert judge agreement, we find main effects indicating significant Naive-Expert judge agreement for Joy (χ^2=29.1, p<.001), Disgust (χ^2=28.6, p<.001), Anger (χ^2=8.77, p=.003), and Anticipation (χ^2=5.48, p=.019). This reveals that the judges were able to accurately rate these emotions in the text regardless of length. For Acceptance a main effect indicates significantly lower agreement between Naive and Expert judges (χ^2=33.7, p<.001; with this also the case for Neutral texts, χ^2=79.3, p<.001).

DISCUSSION

The results show greatest naive-expert judge agreement for the ratings of texts expressing joy, disgust, anger and anticipation. Additionally, we note that overall greater text length increases naive-expert agreement, however examination of the interactions indicates that this is mainly for the texts with low agreement (sadness, fear or acceptance). In the case of disgust, for which there are already high levels of agreement, the extra availability of textual information in the longer text slightly hurts naive-expert agreement.

We note that the greatest naive-expert rater agreement is related to strongly positive and negative emotions (anger, disgust, joy, anticipation): Apparently naive judges were better able to rate texts with strongly marked valence (consistent with [8]). Conversely, texts characterized more by their activity appeared to be assessed around chance levels, and in the case of acceptance showed disagreement.

Discussing our findings in the context of CMC theories indicates that *some* emotion can be accurately expressed and perceived in short blog excerpts. This contradicts predictions by the Social Presence Theory regarding less-rich media such as asynchronous text-based CMC. However, what sense can be made of the behavior of individual emotions in CMC? For the emotions which strongly express valence, these appear to be clearly discernable through thin slices of textual CMC, regardless of length. In the case of perceiving emotions primarily related to activity, here the naive judges seem to have more difficulty. The improvement in performance resulting from increase in text length appears to offer some support for Social Information Processing theory, however exposure to a much greater length of text may be required for significant agreement with the expert judges. Additionally, since we are not able to contrast emotion perception performance for blogs with either synchronous CMC or other media, we do not make stronger theoretical claims.

Limitations

Emotion, like many spontaneously occurring behaviors is difficult to manipulate and measure experimentally without disrupting its expression. One strength of this study is that it uses naturally occurring personal blogs expressing genuine emotion. However, relying upon expert raters to provide the

[1] Throughout this section we report the parameter estimates and their corresponding one degree of freedom Wald Chi-square tests for N=1260.

‘gold standard’ for text emotions has limitations: Our ‘experts’ were very familiar with personal blogs, but they were not psychologically trained for emotion rating. Additionally, the experts may have only selected blog texts which express emotion very saliently, although that may not be the case given the lack of agreement in some cases between the expert and naive raters. Future studies would ideally draw upon self reports or even physiological measures of emotion from the authors during writing, and also contrast this with other forms of communication. Finally, in this paper, we have not examined the differences in the strength of ratings provided by different judges, nor have we examined other background information collected for the judges as part of this experiment (e.g. personality). We leave this to future work.

Contributions

This study builds upon previous work to study the way in which emotion is expressed and assessed in CMC. We note that previous work in this area has been limited to positive and negative emotions (happy vs. sad), that the naive judges of emotion had a 30 minute interaction upon which to base their judgments, and finally that emotions were acted out through a confederate. In the current study, we show that naive raters with little experience of using blogs are able to: (1) identify four emotions (joy, disgust, anger and anticipation) with relatively high agreement with expert judges from naturally occurring data; (2) perform these accurate ratings based on short, asynchronous blog texts, which are (3) genuine emotions collected from real authors.

Further, these findings suggest that some emotions are expressed and perceived through asynchronous text-only environments, apparently contradicting Social Presence Theory which would expect such emotional expressions to be inhibited. Rather, the fact that emotion rating agreement improves with text length is in line with the Social Information Processing theory. However, we are reserved in generalizing this claim since our study does not contrast blog performance with other media.

Finally, although we do not apply machine learning classification to the blog emotions, we note that since it is possible for naive human judges, this sets an interesting challenge for future computational work (cf. [15]). Potential applications may include emotion monitoring of blog posts, or dynamic interfaces which adapt to user state based on linguistic features of the texts.

ACKNOWLEDGMENTS

We thank Scott Nowson for the use of his blog corpus, Jonanthan Ellis for preparing rating materials, and our reviewers for their helpful comments. This work was supported in part by Région de Bourgogne FABER Post-Doctoral Fellowship (05512AA06S2469) (first author), NSF IIS-0705901 (second author), and European Commission Sixth Framework grants NEST 516542 and NEST 029088 (third author).

REFERENCES

1. Ambady, N., LaPlante, D., & Johnson, E. (2001). Thin-slices judgments as a measure of interpersonal sensitivity. In J. Hall & F. J. Bernieri (Eds.), *Interpersonal sensitivity: Theory and measurement*. Mahwah, NJ: Erlbaum.
2. Balog, K., Mishne, G., & Rijke, M. (2006) Why Are They Excited? Identifying and Explaining Spikes in Blog Mood Levels. In *Proceedings of the European Chapter of the Association of Computational Linguistics (EACL 2006).*
3. Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., & Taylor, J. (2001). Emotion Recognition in Human–Computer Interaction. *IEEE Signal Processing Magazine 18 (1),* 32–80.
4. Ekman, P. (1982). *Emotion in the human face* (2nd ed.). New York: Cambridge University Press.
5. Feldman Barrett, L., & Russell, J.A. (1998). Independence and bipolarity in the structure of affect. In *J Personality and Social Psychology, 74,* 967-984.
6. Fussell, S. R. (2002). The verbal communication of emotion: Interdisciplinary perspectives: Introduction and overview. In S. R. Fussell, (Ed.) *The verbal communication of emotion: Interdisciplinary perspectives.* Mahwah, NJ: Erlbaum .
7. Gill, A.J., Oberlander, J., & Austin, E. (2006). The perception of e-mail personality at zero-acquaintance. *Personality and Individual Differences, 40*, 497-507.
8. Hancock, J.T., Landrigan, C., & Silver, C. (2007). Expressing emotion in text-based communication. *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 2007),* 929-932.
9. Makarova, V., Petrushin V. A. (2002). RUSLANA: A Database of Russian Emotional Utterances. *Proc. Int. Conf. Spoken Language Processing (ICSLP 2002),* pp. 2041-2044.
10. Nowson, S., Oberlander, J., & Gill, A.J. (2005). Weblogs, genres and individual differences. In *Proceedings of the 27th Annual Conference of the Cognitive Science Society*, pp. 1666-1671.
11. Plutchik, R. (1994). *The psychology and biology of emotion*. New York: HarperCollins.
12. Scherer, K. R. (2005). What are emotions? And how can they be measured? *Social Science Information, 44(4),* 693-727.
13. Short, J., Williams, E., & Christie, B. (1976*). The social psychology of telecommunications.* New York: Wiley.
14. Walther, J. (1992). Interpersonal effects in computer-mediated interaction: A relational perspective. *Communication Research, 19*, 52-90.
15. Wiebe, J., Wilson, T., & Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation, 39,* 65-210.