University of Edinburgh
School of Informatics

An exploration of human emotion perception from short texts

4th Year Project Report
Artificial Intelligence with Psychology

Jonathan Ian Ellis

March 1, 2005

Abstract: In order to make artificial agents more human it is important that the agents can detect and convey some level of emotional presence. But how well do humans themselves perceive emotion? Working independently, 56 judges each rated 10 short and 10 long text extracts. These texts were taken from the first 200 words of selected weblogs (online diaries) and were then divided into 50 and 150 word sections. The main finding was that it is easier to rate longer texts due to the increase in contextual information. The secondary finding is that the level of contextual similarity that emotions have differs from emotion to emotion and that Acceptance is much less distintive (and appears more neutral) than all of the other emotions. Subjects gave low intensity scores to the long texts which had been selected to convey Acceptance but much higher intensity for shorter Acceptance texts. Comparing this with statistics from the BNC, we suggest that in situations where there is less context, emotional stereotypy cannot be used as effectively as in longer texts where there is more context. Anger and Anticipation are also hard to distinguish from the rest of the emotions, but to a lesser extent, and so they are slightly easier to rate than Acceptance, but still less so than the other emotions.

Acknowledgements

Thank you to Jon Oberlander for assistance, guidance and for reminding me of “The Prisoner”. Thanks also to Scott Nowson for assistance and for letting me use his data. Lastly, thank you to all of my subjects (and especially the experts) without whom I could not have done this experiment.

Contents

1. Introduction

The goal of this project was to investigate human perception of emotion from text. This document is structured as follows: an introduction to emotions and emotion rating, an introduction to automatic emotion rating design and a review of its uses. After this I explain my hypotheses about human emotion perception, explain my materials and results, and finally draw conclusions about my experiment.

1.1 Understanding and conveying emotions

Emotions are difficult to study in part because the experience of an emotion is a very personal experience and reaction to situations varies massively from person to person [9]. Fear of a bear will be a different experience for a park ranger than for someone who has never seen a bear before. This suggests that individual differences may be overwhelming for emotions such as love or depression.

Our own experience of emotion influences our behaviour towards others experiencing those same emotions. When having emotional experiences recounted to us each individual understands the emotions being conveyed based on their own past experiences and upon their knowledge of the other person; this means that we all understand emotions in a different way.

The way that you portray an event when recounting your own emotional experiences will be affected by personality, history, mood at the time, and mood when recounting the situation. The likeliness to report (and the way in which any report on emotional states is made) will vary also with personality [27].

Mathews et al. [23] suggested that emotional states may represent targets for self regulation; i.e. as drivers of voluntary attempts at mood regulation. People differ in their meta-cognitions of the importance of changing moods and cognitive states, and in their strategy preferences. The main work that relates emotional states to performance and information processing centres on the subject's overall mood. As with traits, states have complex effects depending on a host of moderating factors [21]. Extraversion is typically linked to positive affect and Neuroticism to negative (Watson [38]). The relationship between language production and personality has been investigated by Pennebaker and King [27] and more recently by Oberlander and Gill [11]. Their main finding was that extroverts use different (wordier, more abstract) language and they also found pointers towards a variance in the way that punctuation and sentence structure are used. Findings that high extraverts use more words (although these words tend to be less lexically specific) suggests that there would be differences between reading the emotion in a piece of text written by a high and low Extravert [6].

Language usage varies from person to person, not only in the main words that they choose to use but also in the form by which they express it and the content which they choose to express. Each person will then perceive these cues to differing degrees and interpret them in differing ways [34].

Emotions can also be hard to study because it is difficult to recreate a real experience of those emotions in a lab e.g. subjects may not ever show true happiness or anger in the safe environment of the lab.

Whilst language semantics convey emotion – as well as vocabulary choice – little is known about how we actually recognise emotion in text when little or no context is available. With most text interaction between people there is some history and so each person has some expectation about their normal language use, their expected mood, and their likely reaction to certain situations. Sometimes there is no prior knowledge of a person before someone is exposed to their emotions (in any form) and it is here that they must rely on some kind of emotional stereotype.

1.2 Emotion stereotypes

In reading we make subconscious judgements on the emotions that the author wants their readers to feel – these are emotional expectations and, to some extent, they are stereotypes. As with all stereotyping people tend to have one point of view on certain things and if a situation arises repeatedly then people view the new occurrence of that situation using prior examples to form expectations. With reading (and especially when reading material containing personal content such as online diaries) readers cannot help but to draw on past experiences of similar situations that they themselves have been in or have some understanding of. The way in which we are influenced by stereotypes should lead to there being a stronger consensus in texts where there is more background and context; the longer a piece of text is the stronger the consensus on what the emotion contained in that text is. This comes from the idea that increased context gives a reader more information to relate to their own past experiences. Readers drawing on their own emotional past are then going to be thinking of a similar emotion, and emotion level, and will therefore be more accurate in judging what that emotion is. Because of this I am going to compare human rating of texts in long and short texts in order to compare human ability to rate these texts to see if there is any difference in the levels of accuracy that are achieved.

1.3 Emotion scales

In order to analyse how people differ in their ability to rate texts for emotion I need to be able to compare each subject's opinions on the emotions that they think are contained in texts. In order to do this I am going to use a model of emotion that will allow subjects to rate the emotion contained in texts. There are many different models of emotion and below is an evaluation of how and why I chose between them.

There are many different emotional states. There are moods [cheerful, gloomy, irritable, listless, depressed, etc], interpersonal stances [e.g. distant, cold, warm, supportive], attitudes [e.g. like, love, hate, value, desire], and affect dispositions [e.g. nervous, envious, reckless, morose, hostile]. All of these have been used in different emotion scales. When people are writing they are typically writing about moods, but these moods come from emotion [5]. There are multiple ways of demonstrating the emotion space and there are many different words that can be considered to be emotion or emotion related words [14]. There is no agreement on a basic set of emotions thus pragmatic choices must be made between the different emotion scales. There is inevitably loss of information; and worse still, different ways of making the collapse lead to substantially different results. This is well illustrated in the fact that Fear and Anger are at opposite extremes in Plutchik's emotion wheel, but close together in Whissell's activation-emotion space. Thus caution is needed when selecting what scale to use. Hernández et al. [13] used an emotional space that had 5 emotions (Happy, Anger, Sad, Fear, Surprise). Watson and Tellegen's Circumplex theory of affect [39] made use of two major bipolar dimensions of positive and negative affect. This has been used as a guide to the development of Natural Language Processing algorithms for text classification [35] [37]. Both positive and negative affect occur on bipolar continua (high to low).

Another approach to affect lies in the field of Personality Psychology. The five factor model of personality traits represents a continuum of contrasting qualities that are polar opposites (see Mathews et al. for a review [22]). Another major trait model of personality is Eysenck's three factor PEN model (Eysenck and Eysenck, [7]).

The Psychologist Plutchik [31] derived a model of emotion by getting subjects to make lists of words which represented specific emotional states. Many similar or related emotion words can be grouped together under the umbrella of one emotion and Plutchik tried to arrange all of the words into eight primary emotions (as polar opposites): Joy-Sadness; Acceptance-Disgust; Fear-Anger; Surprise-Anticipation. Plutchik used multi-dimensional scaling to classify emotions according to their similarities. Subjects were given pairs of emotion describing words and asked to rate their similarity. From these judgments, a map of emotions emerged and the emotions were arranged as: Primary, Secondary, and Tertiary emotions. The primary emotions shown in the outer circle (see Figure 1.1) are each composed of two distinct secondary emotions. For example, Surprise is composed of Curiosity and Alarm. Acceptance is composed of Curiosity and Love, etc. In a like manner, each secondary emotion is composed of its two tertiary emotions. For example, Misery is composed of Shame and Pessimism. Pride is composed of Dominance and Optimism, etc (Plutchik [30]). Although Plutchik arranged the eight main emotions as polar opposites there are a number of different scales that have been derived from Plutchik's emotion scale with the emotions arranged in different ways.

In approaching the question of emotion I chose to use an emotion space derived from the outer ring of the wheel (see Figure 1.1). This version of Plutchik's emotion wheel was previously used by Makarova and Petrushin [19] and it makes use of activation and evaluation as the axis of the wheel. Here the emotions are placed in such a way that any angle from the centre would represent an emotion and any distance from the centre would represent the strength of that emotion. This provided a continuous spectrum of emotion states. I chose to make use of the layout and 8 main emotions of this wheel but redrew the wheel so that the space was not a continuous spectrum. This was so that subjects could see that the emotion space was plotting key points of that continuous spectrum but it limits user response in order to give as distinct results as possible. Feldman et al. [8] made use of a similar emotion space (although with the evaluation scale renamed pleasantness) as did Cowie et al. [4]. The activation dimension measures how dynamic the emotional state is; the evaluation dimension is a global measure of the positive or negative feeling associated with the state. This emotion space could be charted with only a few emotions or with many emotions as all emotions can be classified using the activation and evaluation scales. Figure 1.1 shows Plutchik's model of emotion. The emotion wheel that I used was based on the emotions found in the outer wheel. Instead of having the emotions shown as polar opposites (as with Plutchik's model) I chose to use the approach from [19] and gave the subjects the same eight emotions but on an activation-evaluation space. This model can be seen in Section 3.2.3.

As well as choosing which set of emotions to have subjects consider, the scale of these emotions must also be selected. Whilst it would be simple if an emotion were a binary state, this is not considered to be the case. The identification of the centre as a natural origin lets the emotional strength be measured as a distance from the origin to a given point of activation-evaluation space. The notion of 'full blown' emotion would be a state where emotional strength has passed a certain limit [5].

After considering the different options for emotion scales the adapted version of the Pluchik model was selected because this same scale has been used in a number of experiments concerning emotional utterances [19].

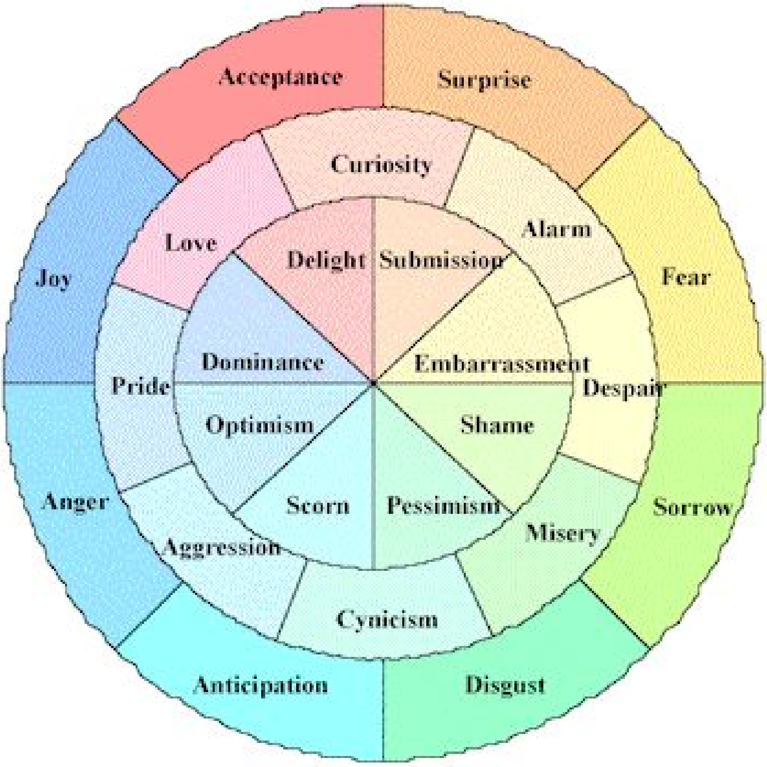

Figure 1.1: Plutchik's Model of Emotion. The outer wheel here shows the primary emotions (each made up of two secondary emotions [middle circle]; each made up of two tertiary emotions [centre]). The primary emotions in the outer wheel are the emotions used in my study.

1.4 Artificial emotion recognition

Emotion recognition software has recently started to be developed. There are many uses for this type of software. Informing someone that something is bad (such as a computer warning to a user) would certainly be responded to better if the language phrasing chosen was in keeping with the emotion supposedly being conveyed [26]. A talking robot would also sound more human if the words and the emotions that it was mimicking were matched. It would be beneficial to incorporate emotion into dialogue systems so that artificial communicative agents could conduct dialogues more naturally and be more engaging [1] [28]. These advances in technology are already becoming integrated into software but real advances need to be made before computers have any real model of our emotions.

A number of different methods and frameworks have been suggested in order to recognise emotions accurately. Lee et al. [17] combine two different forms of prior information to give better emotion recognition. Conati et al. [3] made use of a probabilistic framework using dynamic Bayesian Networks to do a similar task. These networks have drawbacks because they are very precise (non-heuristic) and thus the structure grows vastly over time. Work in textual analysis has resulted in a myriad of approaches to identifying emotions expressed by a sentence or discourse. Earlier approaches involve predetermined keyword search for words that convey strong emotion. Tagging depends on the keywords that a custom dictionary contains. This does have its drawbacks – especially with negation – and its accuracy is limited by the number of keywords that have been identified and captured in the dictionary.

Some automated learning techniques have been developed for detecting subjective text passages (Wiebe et al. [40]; Hatzivassilogou and Wiebe [12]; Yu and Hatzivassiloglou [41]). A range of techniques have been employed for this purpose, generally based on lexical clues associated with opinions. Bucci and Maskit [20] have been working to create a weighted dictionary for the computer modelling of the psycholinguistic variable Referential Activity (RA) in spoken and written language. The RA dimension concerns the degree to which language conveys non-verbal experiences in the listener or reader. High RA (vivid and evocative) and low RA (abstract, general, vague or dismissive) are not emotion ratings alone, but they do go some way to doing this. With the RA the lengths of the texts very much influenced the computability of the RA. Whilst RA is effective already in classifying some of the linguistic context of the texts it is not solely related to emotions. Boucouvals et al. [2] used a tagged dictionary along with grammatical analysis in order to determine which words held most weight in real-time internet communication. The main problem with this is that someone must create the tagged dictionary that covers the target domain. It is not possible to do this without at least some data manipulation by hand.

The use of common sense rules to tag affect was a successful technique from Liu et al. [18] which involved the manual encoding of everyday knowledge. Knowledge had to be encoded this way because whilst it is taken for granted by humans it is not normally captured by statistical approaches. A paper on flaming [16] outlines a complex set of heuristics for identifying flaming (insulting or offensive material) text from the web or emails in a similar way to the knowledge used by Liu et al. Some of these heuristics were then used by Eudora [32] in the development of an email program add-on. Interest in the expression of emotions in the online environment is becoming increasingly prevalent because of the interest in the changes in the ways that we work under certain emotional pressures. Eudora has created a feature that analyses the content of outgoing emails for flaming. This feature detects the mood of the sender in a number of ways including the sentence structure, comparison with a dictionary of words that have bad associations and also a dictionary of sentences that have specifically aggressive meaning. The email is analysed before it is sent and the user is informed with an ice cube or a chilli rating (1, 2, and 3; hot, hotter, and hottest respectively) as a way of recommending to the user that they may want to think again before sending the email.

Previous dialogue act tagging methods have only managed 71% accuracy in emotion tagging (Stolcke et al. [36]); this study found human annotation scoring to be 84% accurate and baseline 35%. One group [15] is currently experimenting with text revision for texts which are more positive and more negative than an existing text through the replacement of neutral words with synonyms of greater positive or negative weighting.

1.5 How a language processor works

A specific tool originally made to classify the personality of a writer of text (Personality language checker (Oberlander et al. [10])) can be adapted for emotion classification. The underlying mechanism for classification relies on McDonald's semantic space approach. Users can be compared with regard to the similarity of the context in which their writing appears. For instance, contexts which are similar to each other would get a low score and highly dissimilar contexts would get a high score. To date the language checker has only been used with texts in order to analyse the personality content but the dictionaries could easily be adapted to suit emotions. We might find the following with regard to Happy:

- Happy 0.03
- Sad 0.11
- Glorious 0.12
- Great 0.31
- Fortunate 0.36
- Perfect 0.40
- Squirrel 0.96
- ...

Ostensibly non-emotion words will still be given scores for each emotion. This means that any text analysed would have a rating for all emotions. The language checker used by Scott MacDonald works on the British National Corpus (which has approximately 100 million words). The personality checker already has an emotion factor in it and this could essentially be used as a quick comparison between the human and the computer elements of emotion detection in text. Rather than using or adapting the PLC, this study explores semantic space (computing a small set of inter-emotion comparisons to compare emotion terms with one another).

2. Hypotheses

The main hypothesis is that the more background information that the readers are presented with – the more context – the better the reader is able to understand what the emotional content is. With increased context they will be able to rely more on emotional stereotypes and these will be more accurate given the increased context. With fewer words the reader in placed in a situation where they are increasingly likely to have to guess whether a situation is in a good or bad context – and if there is not a clear enough context then the reader is more likely to be less accurate about what they think the emotional content is. The reader will be relying on emotional stereotypes that they have and they will judge the text based on single words with strong emotional content in order to label the emotion of the writer. However, given the decreased context, making a judgment based on a few (possibly conflicting) words will make the subject more likely to be incorrect. This is a very similar way to how the language processor works as, to human and computers, certain words will have stronger weighting, and thus the judgement will be based on these words. This experiment does not compare computer performance with human performance, instead it checks how humans fare with more or less information.

The main difference between computers and humans is that humans are able to get an impression of the emotions as a whole, regarding the whole text. Humans have the advantage of not being confused by non-emotion words (or words which are not necessarily strongly associated with any particular emotion). Subjects are expected to do worse at identifying the main emotion in the shorter texts because of a decrease in the number of emotion words and the stereotypes that they draw on based on the available words may not be the correct stereotypes because they may not be based on proper emotion words. Subjects are not, however, expected to fail at rating emotions completely. Instead they are expected simply to do worse with shorter texts. Current computer based emotion work suggests that human subjects would probably still do better than a computer on short texts, but that a computer would be expected to do much worse with long texts because it would analyse all of the words and this would add noise. This in turn would make the result more diluted because of the noise from the increased number of non–emotion words. In order to test human ability at text rating my experiment will have two lengths of text for each subject to read. I will be aiming to establish whether or not the subjects will have a more accurate and stronger opinion on the emotional content in the longer examples of the text.

By making use of statistics from the British National Corpus (BNC) with regard to the contextual similarity of emotion words it should be possible to calculate how distinct emotions are from each other. Emotions which are more distinct should be easier to identify for the subjects and so the scores which they have for these emotions should be reflected by the BNC results. In order to calculate distinctiveness I will compare the contextual similarity of each emotion with the others and then create a single distinctiveness score. I hypothesise that these scores will be ranked in reverse order from the rank order of emotional certainty as this would reflect less contextually similar emotions as being easier to classify and more contextually similar emotions as being harder to classify. In essence, it is claimed that emotional stereotypes rely on the content and context of a text. The greater the context given implies a greater possible understanding from that text. The greater the contextual dissimilarity the easier that emotion will be to identify. Any contextual similarity would indicate why some emotions are considered more similar than others, and hence, why people making judgements about them are more likely to confuse them than they are to confuse others.

3. Method

3.1 The Judges

The 58 judges were all undergraduate and postgraduate students, or recently graduates living in Edinburgh (28 males, 30 females, mean age = 21.4 years, S.D. = 1.56). All were naïve to the rating of emotion. None had previously taken part in any emotion rating experiments.

3.2 Materials

3.2.1 Weblogs

The target texts were taken from a weblog corpus collected previously (Nowson et al. [25]) for a project involving analysing texts for personality information using a machine learning language checker. This consisted of 71 subject's one month long weblogs. The texts had not been written under experimental conditions; instead they came straight from the internet. The weblogs ranged widely in length, from only a few short postings to near daily massive posts of up to a few thousand words. Permissions for further use of each weblog was granted by the authors before I used them. They were all extracted in an anonymous manner. "The internet is increasingly considered as a resource for linguistic study, and a number of studies have focused on the nature of various types of computer-mediated communication (CMC), such as asynchronous e-mail and synchronous chat. A key reason for studying these genres is to determine how they differ from non-computer-mediated analogues . . . To date, however, relatively little work has considered blogs though they are now being discussed as both a topic . . . and a tool for study" – Scott Nowson [25]. In order to select the extracts that I used I read all of the weblogs. This was an arduous task, but it meant that I could pick out any post that I believed to have emotional content. I decided that any post that had emotional content (or no emotion content at all for the Neutral control case) would have the first 200 words read by the experts and they would then pick the strongest examples which would then be used in my subject questionnaires.

An example of an extract from a weblog can be found in the Appendix.

3.2.2 Expert categorisation of the texts

First pass of the content derived 135 extracts (all 200 words long). These had next to be categorised in order to pick the extracts that were thought to have fairly strong and clear emotional content. To do this all 135 texts were read and rated by five experts. Given that the experiments would not fail even if the experts agreed on an emotion and then every single subject classified it as a different emotion an expert could simply be someone who had been exposed to the master set of data (all 135 extracts). In order to increase the level of prior experience that my experts had with the texts 3 of the experts had some prior knowledge of the texts. The other two experts were chosen from the subject pool to analyse the full 135 texts instead of rating just the 20 (150 and 50 word) texts. Whilst the first three experts have been involved at all stages the recruited subjects were simply recruited because they had the time available to read all 27000 words and rate them based on the eight main emotions (and Neutral). Hence, all of the experts had seen a large volume of emotional material and could all be classed as experts of emotion rating. In order to select 20 texts from 135 only texts with 100% expert agreement on emotion were chosen. From these two texts were selected for each of the emotions (and four Neutral texts) and the texts that were chosen were the ones with the highest mean emotion scores from the experts.

3.2.3 Questionnaires

Each subject was given one of four randomised questionnaires. There are 2 examples of each of the 8 target emotions (as classified by experts) as well as four Neutral examples per questionnaire. In the questionnaire these are split into ten long (150 word) and ten short (50 word) extracts. The first 200 words were taken from the start of each original weblog. This was then split into a long and a short section, either the long coming from the first or last 150 words. This cut was made regardless of sentence structure etc. This did mean that some extracts started and ended mid–sentence but it was a uniform cut in order to maintain consistency.

Differences in writer and writing style mean that writers vary hugely in the way that they report events in weblogs. Some writers will report the major thing that happened to them that day first while others may tell the whole story before reviewing how it affected them or before they really use emotional descriptive words. In order to choose which texts to use I had to use a uniform way to select them and the most obvious way was to take the first section of postings in order to a standard extracting method. If these extracts then had more or less of an emotional content it would not matter as the experts would choose the best from these anyway.

From this set 10 short and 10 long extracts were selected for each questionnaire. The texts were arranged in random order. The booklet was prefixed by an explanatory page informing judges of the format of the experiment.

Subjects were asked to give the text extracts a score (one to seven) on the emotion that they think is the main emotion in the extract. The questionnaire was lain out so that there was no priming toward any particular emotion and, further to that, the Neutral appeared more often than any single emotion so that the subjects cannot just presume that each emotion appears once or twice. The questionnaire was 150 words*10 + 50 words*10 and did not take long to complete; under 30 minutes for all subjects.

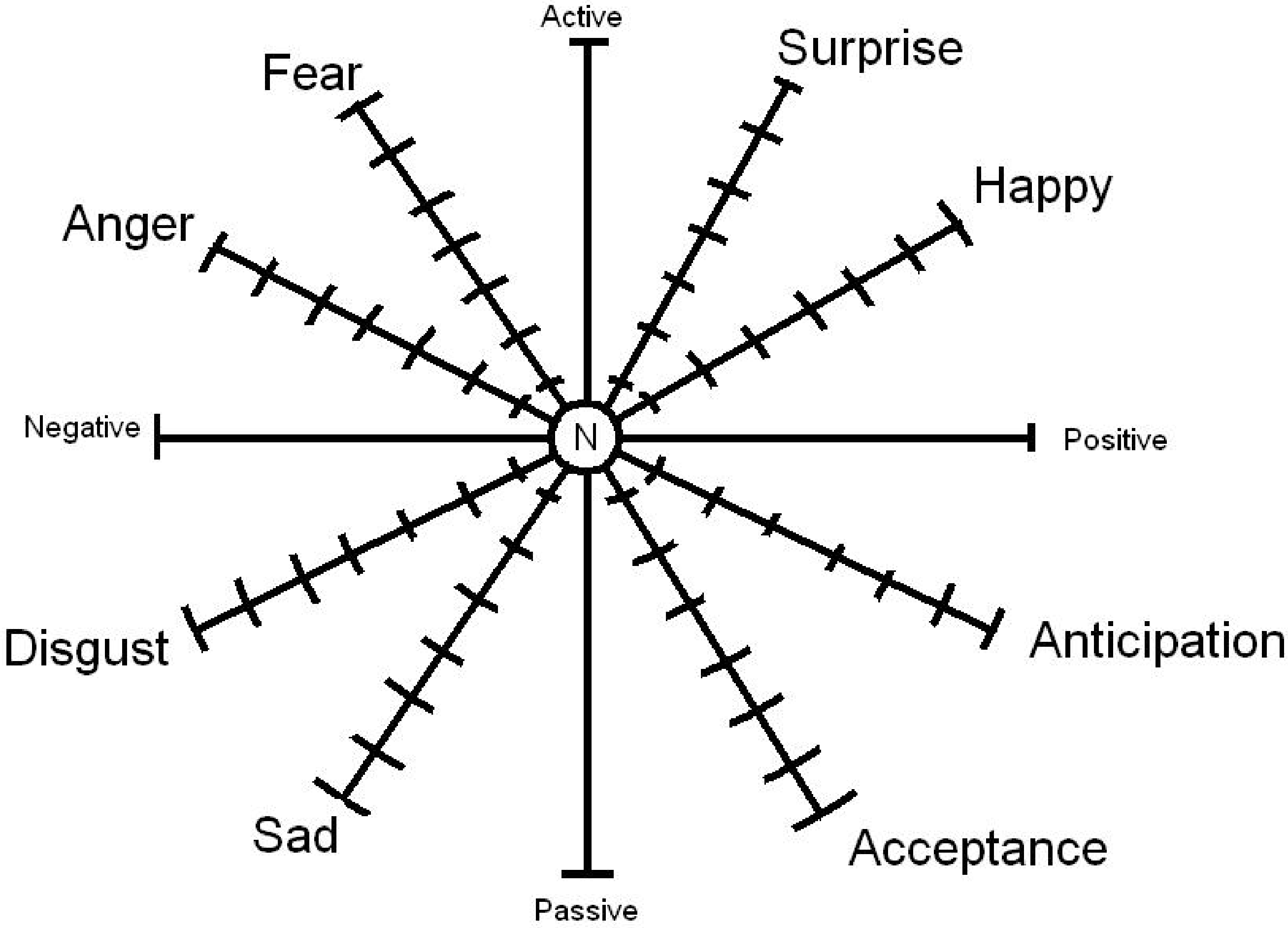

Figure 3.1: Emotion wheel used in my experiment

The emotion scale eventually chosen was the activation-evaluation wheel (see Figure 3.2.3) [5], which has been derived from the Plutchik's emotion wheel [29] [19]. The x-axis is the evaluation or valence axis with positive values on the right side, and the y-axis is the activity axis with high activity on the top. The strength of emotion corresponds to the distance from the center of the circle. The rating scale is displayed in this way in order to give the subjects a simplified view of the emotion space while using the same measures as in [5] which had a continual spectrum of emotional states. Having the emotions split into active and passive, and into positive and negative, should lead the subjects to think about the closeness of the emotions and the way in which the emotion in the text lies on the wheel. The main bonus of the emotion wheel is that people can see the interaction between the different emotions and are able to score the different emotions based on this.

4. Results

4.1 Summary of the findings

The main finding is that it is easier to rate longer texts due to the increase in contextual information. The secondary finding is that the contextual distinctiveness of Acceptance (and to a lesser extent, Anticipation) (from the BNC) is much lower than the other emotions.

4.2 Expert agreement

Expert rating						
Expert	A	B	C	D	E	Mean rating
Anticipation	4.5	4.5	2.5	5.5	6.0	4.6
Acceptance	5.5	4.0	5.0	3.0	6.5	4.8
Surprise	5.5	6.0	4.0	4.0	5.0	4.9
Disgust	4.5	5.0	5.0	5.5	6.0	5.2
Fear	6.0	5.0	4.0	5.5	6.0	5.3
Anger	5.0	6.0	6.0	3.5	6.5	5.4
Sad	7.0	5.0	5.0	6.0	5.5	5.7
Happy	6.5	5.0	5.5	5.5	6.5	5.8
Mean rating	5.6	5.1	4.6	4.8	6.0	5.2
S.D.	0.9	0.7	1.1	1.1	0.5	

Table 4.1: Average emotion score per expert (in ascending order mean rating)

This table shows the expert emotion scores of the selected 16 out of 20 texts grouped into emotions (the remaining 4 texts were categorised as Neutral). These scores are for the intensity of emotion [1 to 7] and are scores of the 200 word long texts and not any shorter texts. These texts were selected for the materials because there was 100% agreement from the experts on the base emotion within that text. Neutral texts do not appear here because they had no emotion content and were chosen on a different scale from the other emotions – instead subjects would say that they were not any of the other emotions and that they contained no real emotional content.

As can be seen in Table 4.1, the level of agreement from the experts varied across the emotions. Results are best for Happy and Sad and worst for Acceptance and Anticipation. Happy and Sad are emotions which are included on almost every emotion scale and which are considered to be polar opposites of each other. Intuitively, Happy and Sad are also emotions to which we are most often exposed. Acceptance and Anticipation, on the other hand, are less vivid but are still considered as key emotions. The standard deviation of the experts differs, with Expert E (S.D.= 0.5) scoring in a small range (between 5.5 and 6.5) and with Expert C (S.D.= 1.1) having a much wider range of scores (from 2.5 up to 6). This shows that even people who have rated 135 texts still have a fairly wide range of opinions in what intensity of emotion they think are being conveyed.

Neutral is set apart from the other emotions because it is not rated using the same kind of scale; it is simply a binary state (a text is or is not Neutral) whereas the other emotions are scaled. It is clear that something which is therefore to be considered as being neutral could more easily be confused with another emotion because subjects may find trace emotion and therefore label that text as having a small amount of an emotion rather than no emotion at all. For this reason, Neutral texts have not been investigated further here. Human emotion stereotypes seem to have a place for almost any word and given any "Neutral" text there will probably be words in it which are linked to at least some emotion for some people.

4.3 Subject–Expert agreement

Table 4.2 shows the number of times each emotion was chosen compared with the gold standard (see first column) emotion in the long texts. The last column shows the overall percentage of answers which agree with the gold standard; these are considered "correct". Across all emotions, the percentage correct was 69%. The experts were 100% in agreement of the emotions in the 200 word texts and so it is interesting that the emotion of Acceptance gained such a low level of agreement in the 150 word segments. Again, emotions which were rated correctly most often are the ones that seem to be more commonly occurring – especially Happy and Sad. Numbers in bold are the "winners" of each emotion; as can be seen in every case other than Neutral the emotion with the most subjects choosing it is the correct emotion. The neutral case is separate (as mentioned above).

Table 4.3 shows the number of times each emotion was chosen compared with the gold standard (see first column) emotion in the short texts. The last column again shows the overall percentage of answers which agree with the gold standard; these are considered "correct". In this table, the percentage correct (41%) is much lower than in long texts, suggesting that the 50 word texts are harder to classify.

					Subject	Rating				
Expert rating	Surprise	Happy	Anticipation	Acceptance	Sad	Disgust	Anger	Fear	N	% correct
N	0	0	4	**26**	2	2	6	4	12	21%
Acceptance	4	4	4	**20**	0	3	17	0	4	36%
Anticipation	0	8	**28**	20	0	0	0	0	0	50%
Anger	0	0	0	0	8	0	**31**	17	0	55%
Surprise	**40**	5	11	0	0	0	0	0	0	71%
Disgust	0	0	0	0	0	**44**	12	0	0	79%
Fear	0	0	0	0	0	4	6	**48**	0	83%
Happy	8	**48**	0	0	0	0	0	0	0	86%
Sad	0	0	0	0	**56**	0	0	0	0	100%
Average Correct										69%

Table 4.2: Long text emotion rating – number of people choosing that emotion (shown in ranking of percentage ranked “correct” against the experts)

However, subjects still selected the correct emotion overall in 6 of the emotions and so they were still very much getting the correct emotions. Other than with Surprise and Fear, subjects (as a group) were correct in classifying emotion in both the long and short texts. It can be clearly seen that even in the short cases, the emotion that a text is meant to be conveying is detectable.

Turning from emotion type to intensity, Table 4.4 shows a set of combined scores for each emotion which combines the number of people who gave this emotion label to the text and the mean score of all of the people who recorded this emotion. In effect, the greater the number here, the greater the confidence in this emotion.

N (number of people) * M (mean user rated score) = confidence

This score is a slightly abstract interpretation of the weighting that people have given this emotion compared with the other emotions (and compared to the gold standard). Neutral is not shown here for reasons discussed above. The percentages show that there is a distinct difference between the levels of emotion that people are rating in longer texts and the levels being recorded in shorter texts. The total intensity with which emotions were rated (shown by the second last column) for the long text is more than double than for the short texts. This suggests that subjects were much more confident in rating the longer texts. As all of the short texts came from the same extracts as the longer texts this seems to suggest that the shorter texts simply do not convey as much emotion as the longer texts. It can be seen that Acceptance jumps from being only 64% confidence in the long texts to being 97% in the short texts. Not only is there an increase in the people agreeing with the experts, but the certainty that they have is much greater. Anticipation and Anger also both climb in subject confidence of rating between long and short whilst all others fall, some by a large amount. Acceptance is much lower than most of the emotions in the long text which would suggest that it is harder to identify there. This difficulty in identifying Acceptance suggests that it is much less distinct from the other emotions. The scores suggest that the subjects were the least confident at rating Acceptance and Anger in the long texts but that in the short texts, subjects were more confident at rating these emotions. Fear stands out as having a very low confidence rating in the short texts. One of the hypotheses for investigation was that the longer texts would have more certain ratings from the subjects and this held true overall.

4.4 Emotion contextual similarity in the BNC

The difference in ratings of emotions may have come either from the selection of the texts or from the difference in difficulty in rating the different emotions. As these texts were selected as being the texts with the highest expert agreement an

					Subject	Rating				
Expert rating	Surprise	Happy	Anticipation	Acceptance	Sad	Disgust	Anger	Fear	N	Correct
N	0	**14**	2	**14**	8	4	2	6	6	11%
Fear	4	4	0	4	**16**	4	8	8	8	14%
Surprise	16	4	8	9	0	0	0	0	**19**	29%
Sad	0	0	0	8	**24**	12	4	8	0	43%
Anticipation	0	4	**32**	12	0	0	0	0	8	57%
Disgust	4	4	4	4	0	**32**	8	0	0	57%
Happy	8	**32**	0	0	0	8	5	0	3	57%
Anger	0	0	4	0	4	8	**36**	4	0	64%
Acceptance	0	4	0	**36**	8	0	0	0	8	64%
										41%

Table 4.3: Short text emotion rating – number of people choosing that emotion (shown in ranking of percentage ranked "correct" against the experts)

		Surprise	Happy	Anticipation	Acceptance	Sad	Disgust	Anger	Fear	N	Total #	% Correct
Long	Acceptance	16	64	32	1600	0	24	748	0	0	2484	64%
Long	Anger	0	0	0	0	448	0	5580	1360	0	7388	76%
Long	Surprise	6560	60	528	0	0	0	0	0	0	7148	92%
Long	Anticipation	0	352	6048	0	0	0	0	0	0	6400	95%
Long	Happy	320	12096	0	0	0	0	0	0	0	12416	97%
Long	Fear	0	0	0	0	0	64	84	11520	0	11668	99%
Long	Disgust	0	0	0	0	0	8272	48	0	0	8320	99%
Long	Sad	0	0	0	0	16800	0	0	0	0	16800	100%
											72624	90%
Short	Fear	48	32	0	48	960	16	224	64	0	1392	5%
Short	Surprise	576	0	288	144	0	0	0	0	0	1008	57%
Short	Sad	0	0	0	256	1728	384	64	128	0	2560	68%
Short	Anticipation	0	88	7308	1440	0	0	0	0	0	8836	83%
Short	Disgust	48	32	32	32	0	2176	160	0	0	2480	88%
Short	Anger	0	0	96	0	64	256	7488	48	0	7952	94%
Short	Happy	160	4736	0	0	0	64	20	0	0	4980	95%
Short	Acceptance	0	16	0	4320	128	0	0	0	0	4464	97%
											33672	73%

Table 4.4: Intensity – These are constructed scores multiplying the number of people who rated a text a particular emotion with the scores that they rated this emotion.

exploration was needed into why the differences that occurred. There may be an inherent difference in the way in which emotions are perceived in long and short texts and more investigation was needed into why this is. By giving the subjects an emotion wheel with 8 emotions to choose from, I may have influenced the subjects to choose from emotions that they perhaps did not think included the emotions they considered to be actually in the text. When debriefing subjects, most said that there were other emotions that they wished to use as labels that were not available on the emotion wheel.

In order to explore the perception issues with emotion and the emotion wheel I have looked at the relative semantic distance from each emotion to each other emotion. In order to do this I used an online search [24] to calculate the Contextual Similarity between each emotion word on the wheel and all other emotion words on the wheel. The default similarity measure is calculated as the cosine of the angle between word vectors normalized so that the similarity values are z-scores (allowing words of markedly different frequency (with consequent differences in vector sparseness) to be legitimately compared). This gives a bit vector for all contexts which then are added together and semantic similarity is computed from the cosine of the angle between them. These bit vectors are derived from the frequency distributions of the words occurring in the immediate context of a target word (emotion word), computed over the British National Corpus. Words that occur in the same sorts of contexts are thus contextually similar, and tend to be similar in meaning (in the case of emotions, they share emotional context and so are likely to convey the same or related types of emotion). The words were searched for in their lemmatised form. The numbers in the table show the similarities of all of the emotions against one another. These numbers summed and averaged show the comparative contextual similarity of each emotion to all other emotions. Note that the co–occurrence of A and B is not the same as B and A. This is because the scores are normalised and so the relation of A to B will be altered depending on what the contextual similarity scores are for A and B in order to make all of the contextual similarity scores comparable to one another.

The BNC is made from human texts, so finding contextual similarities between some emotions from the BNC would suggest that subjects will also see these contextual similarities in the texts via the emotions which they are rating. Hence subjects should find that some emotions are more (and some are less) distinctive from each other, and from Neutral.

Table 4.5 shows a comparison of contextual similarity of emotion words from the BNC. The contextual similarity scores for target versus comparison and for comparison versus target are summed and then divided by seven. It is divided by seven because there are seven emotions other than the one being rated and we gain an average for that emotion against all of the other emotions. This shows that the scaling varies between emotions and therefore some emotions should be

	Surprise	Happy	Anticipation	Acceptance	Sad	Disgust	Anger	Fear	Sum
Surprise	-	2.8437	1.1180	1.2561	4.0677	3.5578	3.2512	3.3390	19.4335
Happy	3.0466	-	1.2851	0.7791	5.7500	3.0143	1.9470	2.4738	15.2493
Anticipation	1.1375	1.40	-	3.49	0.19	0.54	0.14	2.46	8.23
Acceptance	1.0102	0.3378	3.2008	-	0.3189	0.0464	0.3152	0.8013	5.0204
Sad	4.6634	6.3206	1.4144	0.2403	-	4.9834	4.6781	4.1581	21.7949
Disgust	3.6243	2.8790	1.4144	0.3037	4.4215	-	6.1865	4.3720	19.5771
Anger	3.5304	1.9023	3.2008	0.5420	4.4514	6.6567	-	6.1539	22.9071
Fear	3.2202	2.1910	2.4408	1.0464	3.4545	4.1339	5.4227	-	18.6893
C vs.T Summed	20.23	15.03	12.96	6.40	18.59	19.38	18.69	20.42	
T vs.C Summed (Column10)	19.43	15.25	8.2	5.02	21.80	19.58	22.91	18.69	
Average	19.8	15.1	10.6	5.7	20.2	19.5	20.8	19.6	
Divided by 7	2.83	2.16	1.51	0.82	2.88	2.78	2.97	2.79	

Table 4.5: This table shows the comparisons (C = comparison; T = Target) between each of the emotions with regard to contextual similarity. Target emotions are shown along the top row and Comparison emotions are shown down the first column. The larger a number is, the less similar the context of that emotion from the context of all of the other emotions; the smaller the number is, the more contextually similar an emotion is.

easier to determine the strength of because they are more distinctive from all of the other emotions.

As can be seen, six of the emotions are approximately level but Anticipation and Acceptance are of a much lower level of contextual distinctiveness (even normalised) and so this means that they are not comparable in terms of rating, because they can be perceived to be closer to neutrality than the other emotions. The emotions that are in this experiment therefore differ in the way in which they are perceived to lie on the emotion wheel. If I were to redraw the emotion wheel based on this information then the relative lengths (and the relative distances from zero) of emotions would vary by emotion. Here the crossover point would still be considered to be neutral, and the relative distances would be adjusted so that emotions with high contextual similarity and low distinctiveness (Acceptance and Anticipation) would be considerably closer to the zero point than the emotions that are highly dissimilar. Another alternative would be to retain the same emotion wheel and then use the similarity scores as weightings in order to have a normalised emotion space reflecting how easy or hard it is to score emotions.

Further analysis could be performed on this information to compare the different emotions in order to see where they should lie compared with one another but, for now, it is useful simply to prove that there may be some variance in the strengths of the different emotions and that, based on the figures from the BNC, acceptance and anticipation should be being seen to be much less distinctive than the other emotions.

4.5 Comparison of text rating with the BNC

Compare the order of the intensity from experts (table 4.6 – ascending) with the order which the emotions can be ranked based on contextual similarity (table 4.7) and a distinct similarity can be seen.

As can be seen in tables 4.6 and 4.7 the only wildcard here is Happy, as it was easiest to classify but was not very emotionally distinct. Possible reasons for this are proposed in the Discussion.

There is a general agreement between the ranking of agreement in subjects for the long texts and of the contextual similarity of emotion words from the BNC. Comparing the shorter texts to the BNC score rankings found a general lack of agreement. The contextual dissimilarity of emotions means that they fall more easily into an emotional stereotype. Being an emotional stereotype means that an emotion is easier to classify and hence the more context that a subject is presented with the more the emotionally stereotypical emotions will be identified. Hence

Gold Standard	Mean score
Anticipation	4.6
Acceptance	4.8
Surprise	4.9
Disgust	5.2
Fear	5.3
Anger	5.4
Sad	5.7
Happy	5.8

Table 4.6: Expert intensity of score for each emotion

BNC emotion word contextual similarity	
Acceptance	0.82
Anticipate	1.51
Happy	2.16
Disgust	2.78
Fear	2.79
Surprise	2.83
Sad	2.88
Anger	2.97

Table 4.7: This is the final row of table 4.5 sorted by contextual similarity score. This is a measure of distinctness of emotions (the larger the number, the more distinct the emotion is from the others).

	Long Ranked	L%	Short Ranked	S%	% Difference
Acceptance	8	36	1	64	28
Anger	6	55	2	64	9
Anticipation	7	50	5	57	7
Disgust	4	79	4	57	(22)
Happy	2	86	3	57	(29)
Surprise	5	71	7	29	(42)
Sad	1	100	6	43	(57)
Fear	3	83	8	14	(69)

Table 4.8: This table shows the long and short texts ranked by percentage of “correct” responses and sorted by percentage change from long to short (brackets show negative numbers). As can be seen Acceptance had a huge increase in correct responses between long and short whereas five of the emotions had a big drop in correct subjects, and Anger and Anticipation changed only a little.

this archetypicality means that the intensity of an emotion will be more easily identified and recorded as being of a higher level. The emotions which are less distinct are less stereotypical. This makes them harder to separate from all of the other emotions (and from Neutral).

5. Discussion

The expectation was that subjects would not find the emotion content of shorter texts as easy to rate as in the longer texts. This was generally found to hold true, but the contextual similarity of emotions had an impact on what emotions subjects selected in both the long and short texts. Some emotions appear to be more 'distinctive' than others; these also appear to be easier to rate. When information in a text is scarce, subjects seem to gravitate towards the less distinctive. Perceived intensity also varies between emotions, and this makes it harder to calibrate one against another. When reading a text, the emotions that the text tries to convey and the emotion that it makes you feel can be very different. Asking someone to rate the emotion that the writer felt when they wrote it turns out to be more complex than I had realised because it requires complex reasoning about the intentions of the writer. There could be many reasons why there is a lack of agreement as to emotion type but I think that the main one is most likely to be the fact that the longer texts simply have more context and are therefore easier to rate for humans. The shorter texts are not as reliable as sources of evidence, they have less subject agreement and are therefore showing that the lack of emotional information is in fact hindering the analysis process. Shorter texts are less intense than longer texts and this encouraged people to choose emotions which lie closer to the neutral point (because they are less distinct from Neutral than the other emotions).

Subjects seemed to find it particularly hard to classify Acceptance in longer texts. The reason for this may be that Aceptance may be an emotion that, when being used in a weblog, is used to talk about emotional things without conveying much emotion. This would be another reason why Aceptance falls closer to the neutral point of the wheel than a more direct emotion like Happiness e.g. writers talking about a happy event inherently use emotion and in particular, the emotion of Happiness a lot.

With Neutral texts it seems that almost all were rated as having some (non-neutral) emotional content by subjects. I think that this is because it is almost impossible to select or write an entirely neutral passage of text and because there will always be words which can be considered by some people to show a trace of a particular emotion. The subjects were given instructions which asked them to identify the main emotion in the text and so the lack of Neutral responses to Neutral texts may have been caused by the effect of the task instructions.

Happiness was predicted from the BNC contextual comparison not to be the most distinctive (and therefore easy to rate) emotion but subjects actually found it quite easy to rate. The main reason for this is probably that humans simply have a good enough idea of what things make people happy (without actually using the word happy or any other emotion words relating to happy). It is probably enough to say that happiness is such a deep rooted human emotion that it is easy to classify whether or not certain contextual triggers for it are present. There are many different ways of expressing the emotion of happiness (and many different words to use). This means that contextual similarity does not always accurately convey that an emotion will be easy to distinguish from other emotions. Nonetheless the BNC results have reflected the human results fairly well. The Happiness anomaly suggests that deeper thought about 'distinctiveness' may be needed.

The book "Exercises in style" [33] is made up of one short two paragraph story repeated in around a hundred different writing styles. Not only does the book show that there are many different ways of conveying almost the same basic information but in (almost) every case a casual reader would be able to pick out distinct differences between the different versions of the story. Although the different chapters are conveyed in different writing styles and were all written by the same author it is not difficult to go along with the pretence that these styles represent writing styles of different people. Writers of weblogs did not differ to the extent of writing in sonnet form but there was a big range in writing style. Further work could consider how conventional notions of style interact with perception of emotion.

One final consideration about human emotion scoring is that people may simply all form their own slightly different models of emotion. Hence, the emotion space that I used for this experiment may simply not be consistent with the model that some people have. Past emotion work has tried to come up with "the ultimate emotion space" but has simply produced many different emotion spaces. All of these emotion spaces are valid in particular contexts of use.

Human emotion stereotypes seem similar to emotion dictionaries. Humans seem to relate many different words to emotions. Some words aren't directly related to an emotion but they still have a very distant psychological connection to that emotion. Every person will differ in their judgments, (and differ more the further away a word is from a target emotion) but generally common emotion words will be shared by almost everyone within a community. I think that given a short text humans could pick out almost any single word and (even if the emotion rating of that word is actually very small) say that that word somehow related to an emotion. Because of this, below a certain volume of text, the subject will really only be able to rate texts which have particularly strong emotion words in them. Subjects seemed to prefer to decide that a text belongs to an emotion rather than say that the text is Neutral even when Neutral was a clearly available option. Perhaps in the real world purely text driven emotional understanding isn't always as important as we think. People seem to be able to get through life using email, text messages, and instant messaging (which all often have a word count lower than 50 words) without too many problems. To put the point more strongly, maybe deep emotional understanding is not always essential and instead, as humans, we are just good at adapting to any situations that arise based on any new emotional signals that are received.

6. Conclusions

When conducting an experiment involving emotion rating, experimenters should note that the behaviour of raters varies between long and short texts. This may affect some experimental designs. Subjects appear to rate texts based on emotion with higher accuracy in longer texts. From a text of 150 words, the 58 subjects consistently agreed with a gold standard rating of the text's author's emotion for five of the emotions, but not for Anger, Anticipation or Acceptance. However, with shorter 50 word texts, subjects were only able to rate Anger and Acceptance effectively. Acceptance stands apart from all of the other emotions as being rated the main emotion in short texts much more than in long texts. The reason for this may be that Acceptance is a particularly neutral emotion and it has proved much harder to stereotype than any other emotion because it shares little contextual similarity with any other emotion. In a situation where subjects were unsure of the main emotion they chose a weaker emotion in its place; Acceptance.

Bibliography

[1] J. Bates. The role of emotion in believable agents. In *Communications of the ACM 37(7)*, 1997. p122–125.

[2] AC. Boucouvals and X. Zhe. Text-to-emotion engine for real time internet communication. In *International Symposium on CSNDSP*, 2002.

[3] C. Conati and X. Zhou. A probabilistic framework for recognizing and affecting emotions. In *Proceedings of the AAAI 2004 Spring Symposium on Architectures for Modeling Emotions*, Stanford University, CA, 2004.

[4] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schröde. Feeltrace: An instrument for recording perceived emotion in real time. In *Proc, ISCA Workshop on Speech and Emotion*, Newcastle, NI, 2000. pp.19–24.

[5] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and JG. Taylor. Emotion recognition in Human-Computer interaction. In *IEEE Sig. Proc. Mag.* Vol. 18(1), Jan 2001. pp.32-80.

[6] J. Dewaele and A. Furnham. Personality and speech production: A pilot study of second language learners. In *Personality and Individual Differences, 28, 355-365*, 2000.

[7] HJ. Eysenck and SB. Eysenck. Manual of the eysenck personality scales. Sydney, 1991. Hodder and Stoughton.

[8] L. Feldman Barrett and JA. Russell. Independence and bipolarity in the structure of affect. In *Journal of Personality and Social Psychology, 74, 967-984*, 1998.

[9] JM. Fellous and MA. Arbib. Emotions: From brain to robot. In *Trends in Cognitive Science, 8(12):pp.554-561*, 2004.

[10] A. Gill and J. Oberlander. Perception of e-mail personality at zero-acquaintance: Extraversion takes care of itself, but neuroticism is more of a worry. In *Proceedings of the 25th Annual Conference of the Cognitive Science Society*, Boston, 2003. pp456–461. Boston, July 31-August 2, 2003.

[11] A. Gill, J. Oberlander, and E. Austin. Rating e-mail personality at zero acquaintance. Submitted for journal publication.

[12] V. Hatzivassilogou and J. Wiebe. Effects of adjective orientation and gradability on sentence subjectivity. In *Proceedings of the conference on Computational Linguistics*. Coling, 2000.

[13] DJ. Hernández, O. Déniz, J. Lorenzo, and M. Hernández. BDIE: a BDI like architecture with emotional capabilities. In *Architectures for modelling emotion: Cross-Disciplinary foundations.* From AAAI Spring Symposium, 2004.

[14] http://www.humanityquest.com/. *Home Page For 500+ Human Value of emotion(shows a list of emotion related words).* Berkely, Retrieved:18 Feb 2005, Last updated: 20 Feb 2005.

[15] DZ. Inkpen, O. Feiguina, and G. Hirst. Generating more-positive or more-negative text. In *Computing Attitude and Affect in Text*, Dordrecht, The Netherlands, 2005. Springer.

[16] D. Kaufer. Flaming: A white paper. Carnegie Mellon, 2000.

[17] CM. Lee, S. Narayanan, and R. Pieraccini. Combining acoustic and language information for emotion recognition. In *Proc. of ICSLP*, Denver, CO, 2002.

[18] H. Liu, H. Lieberman, and T. Selker. A model of textual affect sensing using real-world knowledge. In *Proceedings of the seventh international conference of intelligent user interfaces*, 2003.

[19] V. Makarova and VA. Petrushin. Ruslana: A database of russian emotional utterances. In *International Conference on Spoken Language Processing.* ICSLP, 2002.

[20] B. Maskit, W. Bucci, and AJ. Roussos. Capturing the flow of verbal interaction: The discourse analysis attributes program. In prep.

[21] G. Mathews. Designing cognitive architectures and beyond. In *Architectures for modelling emotion: Cross-Disciplinary foundations.* From AAAI Spring Symposium, 2004.

[22] G. Mathews and I. Deary. *Personality Traits.* Cambridge University Pres, Cambridge, 1998.

[23] G. Mathews, M. Zeidner, and R. Roberts. *Emotional intelligence: science and myth.* Cambridge mass: MIT Press, 2002.

[24] S. McDonald. http://www.hcrc.ed.ac.uk/~scottm/semantic_space_model.html. In *Semantic Space Model Demo.* Edinburgh University HCRC, 14/02/2005.

[25] S. Nowson, J. Oberlander, and A. Gill. Weblogs, genres and individual differences. 2005. Submitted.

[26] A. Ortony. On making believable emotional agents believable. In *Emotions in humans and artefacts (Trappl, R., Petta, P., Payr, S., eds), p 189*, Cambridge, MA, 2002. MIT Press.

[27] JW. Pennebaker and L. King. Linguistic styles: Language use as an individual difference. In *Journal of Personality and Social Psychology, 77, 1296–1312*, 1999.

[28] P. Piwek. A flexible pragmatics-driven language generator for animated agents. In *Proceedings the Research Note Sessions of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL'03)*, 2003. pp.151–154.

[29] R. Plutchik. *Emotion: A Psychoevolutionary Synthesis.* Harper and Row, New York, 1980.

[30] R. Plutchik. Emotions and their vicissitudes: Emotions and psychopathology. In *M. Lewis and J. M. Haviland (eds.), Handbook of Emotions*, New York, 1993. Guilford. pp.53–65.

[31] R. Pultchik. *The psychology and biology of emotion.* HarperCollins, New York, NY, 1994.

[32] Qualcomm. http://www.eudora.com/email/features/moodwatch.html. In *Eudora MoodWatch*, 14/02/2005.

[33] R. Queneau. *Exercises in style.* Translated by John Calder.

[34] ET. Rolls. *The brain and emotion.* Oxford University Press, Oxford, 1998.

[35] VL. Rubin, JM. Stanton, and ED. Liddy. Discerning emotions in texts. In *The AAAI Symposium on Exploring Attitude and Affect in Text*, Stanford, CA., 2004. AAAI-EAAT.

[36] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, and R. Martin. Dialogue act modelling for automatic tagging and recognition of conversational speech. Computational linguistics, 2000.

[37] A. Tellegen, D. Watson, and LA. Clark. On the dimensional and hierarchical structure of affect. In *Psychological science, Vol. 10, No 4*, 1999. pp.297–303.

[38] D Watson. *Mood and temperament.* Guilford Press, New York, 2000.

[39] D. Watson and A. Tellegen. Toward a consensual structure of mood. In *Psychological Bulletin, 98*, 1985. pp.219–235.

[40] J. Wiebe, R. Bruce, and T. O'Hara. Development and use of a gold standard set for subjective classifications. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics.* (ACL-99), 1999. pp.246–253.

[41] H. Yu and V. Hatzivassiloglou. Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences.

In *Proceedings of the conference on Empirical Methods in Natural Language Processing*. EM NLP, 2003.

Appendix A. Weblog Extract

The following is an extract from the first 200 words of one of the anoymised weblogs. The markers indicate where the text would have been cut (either to have 150 words at the start and 50 at the end or 50:150).

Last night I had a horrible nightmare about being at Julia's house. I can't remember the specifics of it, but I was reaching for the door handle, and a spider jumped on my hand. I freaked out, but I couldn't move any of my limbs. I tried to smack the

The cut could be made here

thing off my hand, but my stupid arm just wouldn't move. I was terrified, horrified, my God it was awful. I woke up unable to move, from sheer terror. It took me forever to separate the reality from the dream enough that I could get up and out of the room. And then, a few minutes ago, I walked into my bedroom and saw a spot on my bed. I thought, hmm, must be a dust ball. But then it moved. I jumped. I looked around for something to kill it. I couldn't reach my shoe, it was on the

The cut could be made here

other side of the spider. My eyes shifted wildly around the room, looking for anything to club it with. I grabbed my machete; it's in a case, it was sort of flat. But then I realized that if I hit it and didn't kill it, or worse, if I just