The Conceptualisation of Emotion Qualia: Semantic Clustering of Emotional Tweets

E.Y. BANN

Advanced Emotion Intelligence Research
E-mail: eugene@aeir.co.uk
www.aeir.co.uk

J.J. BRYSON

Department of Computer Science, University of Bath,
Bath, BA2 7AY, United Kingdom
E-mail: J.J.Bryson@bath.ac.uk

A plethora of words are used to describe the spectrum of human emotions, but how many distinct emotions exist, and how do they interact? Over the past few decades, several theories of emotion have been proposed, each founded upon a set of basic emotions, and each supported by an extensive variety of research including studies in facial expression, ethology, neurology and physiology. Here we propose a theory that people transmit their understanding of emotions through the language they use that surrounds mentioned emotion keywords. Using a labelled corpus of over 21,000 tweets, six of the basic emotion sets proposed in existing literature were analysed using Latent Semantic Clustering (LSC) to propose the *distinctiveness* of the semantic meaning attached to the emotional label. We hypothesise that the more distinct language is used to express a certain emotion, then the more distinct the perception (including proprioception) of that emotion is, and thus more basic. This allows us to select the dimensions best representing the entire spectrum of emotion. We find that Ekman's set, arguably the most frequently used for classifying emotions, is the most semantically distinct. Next, taking all analysed (that is, previously proposed) emotion terms into account, we determine the optimal *semantically irreducible* basic emotion set using an iterative LSC algorithm. Our newly-derived set (Accepting, Ashamed, Contempt, Interested, Joyful, Pleased, Sleepy, Stressed) generates a 6.1% increase in distinctiveness over Ekman's set (Angry, Disgusted, Joyful, Sad, Scared).

Keywords: Basic Emotions, Latent Semantic Clustering, Lexical Analysis, Twitter.

1. Introduction

There are a great variety of words that describe the spectrum of human emotion. Many theories posit the existence of a set of 'basic emotions' that are hardwired into our brain as individual neurological circuits,[1–5] and that all other emotions are derived from these 'biological primitives' as either a combination or specific valence of these neural circuits.[6] Recently, however, the notion that emotion is a conceptualised act has been proposed,[7] and experimental results have been shown to support this hypothesis.[8] Emotion in this sense can be regarded in the same way as colour, insofar we categorise and communicate discrete colours within the confines of language, even though colour itself is in fact a spectrum of visible light.

The primary objective of this research is to evaluate existing basic emotion sets to discern which contain the most number of emotions expressed in the most distinct language, testing the hypothesis that the more distinct an emotion is (that is, unlike any other emotion), the more distinct the language is used to express the experience of that emotion. *Semantics* refers to the meaning of an expression; in particular, we consider co-occurring words to measure similarities of meaning. We attempt to show such semantic changes in emotion language from a corpus of explicitly expressed emotions extracted from the micro-blogging website Twitter, and evaluate six basic emotion sets on a scale of *semantic distinctiveness*, based on the hypothesis that the more distinct the language used to express a certain emotion, then conceptually (i.e. what we understand that emotion keyword to mean), the more psychologically irreducible that emotion is. The less semantically accurate a set of emotions is, the more similar these emotions are to each other, or in other words, if similar words are used when expressing two different emotions, then these emotions are, in theory, conceptually, and thus psychologically, similar. The secondary objective of this research is to identify a set of basic emotions by identifying the most semantically distinct emotion keywords relative to the underlying semantic features of each expression within the corpus.

2. The Psychology of Emotion

Emotion is that which leads the subject's condition to become so transformed that one's judgement is affected,[9] triggered by a subconscious appraisal process regarding an issue of personal value.[10] It is characterised by behavioral, expressive, cognitive, and physiological changes[11] and can be started and executed unconsciously.[12] The desire to experience or not

experience an emotion largely determines the contents and focus of consciousness throughout the life span.[13]

The above definition of emotion is not a conclusive definition of emotion, but amalgamates many of the important aspects from notable theorists' definitions. Attention is drawn to Aristotle's wording, stating that emotion "is *that* which...", implying that emotion is in fact a type of *quale*, that is, a subjective conscious experience that cannot be communicated, or apprehended by any other means other than direct experience.[14] Qualia refers to subjective 'raw' feelings, for example, the taste of red wine, or the experience of seeing the colour red. Emotion qualia thus refers to the raw feel of an emotion; the actual phenomenon of a particular emotion experienced may actually differ according to each person's perception of that emotion.

The dominant theory of emotion postulates the existence of a small set of hardwired, or 'basic', emotions, and consequently the majority of textual emotion recognition research has been based on such.

ANGER DISGUST FEAR JOY SADNESS SURPRISE

Ekman's basic emotion set[15] (shown above) is arguably the most frequently used within the field of computer science for emotion mining and classification. However, not only do the emotions comprising each basic emotion set vary amongst theorists, they do not always agree the definitions of emotion, thus adding to the confusion in delineating the set of basic emotions, or whether they exist at all. This can be viewed as a problem symptomatic of the vagueness of language, which suggests that there is a general problem about how to talk about the emotion qualia.[6] There are two viewpoints concerning the advocation of basic emotion sets: they are either based on biologically primitive or psychologically irreducible emotions (see Bann[16] for a detailed discussion).

Barrett's work[7] studied the act of conceptualising core affect, in other words, why people attach emotion labels to the experience of emotion qualia. She proposed the hypothesis that emotion is a psychological event constructed from the basic elements core affect and conceptual knowledge. In a study focusing on the conceptualisation of *fear*,[8] it was found that neither the presence of accessible emotion concept knowledge nor core affect alone was sufficient to produce the (world-focused) experience of fear. As emotions are constructed from conceptual knowledge about the world, we can see that emotions themselves are concepts that human beings begin learning in infancy and continuously extend and revise throughout life.[8]

This repeated experience of labelling a combination of core affect and the context in which it occurs as an emotion provides "training" in how to recognise and respond to that emotion; in this sense, Barrett described emotions as "simulations". This "skill" of conceptualising core affect as an emotion might be a core aspect of emotional intelligence — in much the same way as conceptual thinking is core to cognitive intelligence — defining how humans deal with their internal state, but more importantly, defining the emotion labels used as a combination of specific experiences. Each person's conceptualisation of their emotion spectrum is thus unique; it is this conceptualisation that we attempt to aggregate and analyse in this research.

Emotions can be expressed in a variety of ways including facial expressions, body language, tone of voice, and the language used in speech and text. This research focuses on the most explicit of these — the language used in communication — with the proposition that how humans communicate to one another can reveal individual conceptualisations of specific emotions, given that the specific emotion keyword is used within the communication. Defining *basic emotions* as emotions that are conceptually distinct from any other emotion, we explore the hypothesis that the language used in communicating basic emotions should be significantly different for each one, as each basic emotion should describe a significantly distinct concept.

3. Semantic Analysis

Over the past few decades there has been significant evidence that people's psychological aspects can be predicted through analysis of language style. One notable example is Rosenberg's work on verbal behavior and schizophrenia.[17] He found that, while the speech of those diagnosed with schizophrenia did not differ from unaffected people on the structural level, it did differ on the semantic level, i.e. with regard to the thematic concerns that were being addressed. It is this deviation from expected thematic concerns, which are linked to general and sex-specific social role expectations, that is associated with the diagnosis of schizophrenia.[17] Analysis of language semantics has been used extensively in research, including discovering individual differences in personality,[18] lie detection[19] and discovering individual differences in beliefs.[20] With respect to emotion analysis, French found that co-occurrence techniques such as Latent Semantic Analysis does not detect personality from short text samples,[21] but do reveal that texts expressing particular emotions have a greater semantic similarity to corresponding exemplar words.[22]

By analysing the semantics, specifically, co-occurrence statistics, of the language expressing individual emotion keywords, we can discern those emotions that are similar and those that are distinct. We postulate that similar emotions are represented by similar semantics, and propose to cluster emotional documents based on the underlying meanings of each document.

3.1. *Latent Semantic Analysis*

Latent Semantic Analysis (LSA)[23] is a variant of the vector space model that aims to create a semantic space by means of dimensionality reduction techniques and has been widely used in a variety of domains, from document indexing to essay grading. It has also been used in emotion classification of news headlines, performing better than Naïve Bayes in the case of recall but not as good as WordNet in terms of precision.[24] Given a raw co-occurrence matrix $\mathbf{M}$ using the entire vocabulary as `B`, this is transformed by `A` (the documented function for LSA is log-entropy normalisation), and `M` is applied to reduce dimensionality.

There are several techniques for `M` that reduce the dimensionality of words constituting the semantic space, the original method documented for LSA being Partial Singular Value Decomposition (PSVD). PSVD uses Singular Value Decomposition (SVD) to decompose the data matrix $\mathbf{M}$ into the product of three matrices:

$$\mathbf{M} = \mathbf{T}\Sigma\mathbf{D}^T \tag{1}$$

where $\mathbf{T}$ is the term matrix, $\mathbf{D}$ is the document matrix and Σ is a diagonal matrix with singular values sorted in decreasing order that act as scaling factors that identify the varience in each dimension. LSA uses a truncated SVD, keeping only the k largest singular values in Σ and their associated vectors:

$$\mathbf{M} \approx \mathbf{M}_k = \mathbf{T}_k\Sigma_k\mathbf{D}_k^T \tag{2}$$

This reduced-dimension SVD, or PSVD, $\mathbf{M}_k$, is the best approximation to $\mathbf{M}$ with k parameters, and is what LSA uses for its semantic space. The rows in $\mathbf{D}_k$ are the document vectors and the rows in $\mathbf{T}_k$ are the term vectors in LSA space.

4. The Emotional Twitter Corpus

Twitter is a public micro-blogging system that allows users to share short messages of up to 140 characters and, as these are publicly available, pro-

vides us with an ethical way of collecting a diverse range of public expressions. Coupled with the fact that a good proportion of tweets project the user's emotion — indeed, French found that some emotions, particularly those with strongly marked valence, can be accurately expressed and perceived in short blog excerpts[25] — we are able to assume that Twitter is a valid sample of human emotive expression and thus a suitable corpus for this project. There is somewhat of an explicit impulse to communicate emotions on Twitter and although the underlying cause is not always explicitly mentioned, it is this factor that we attempt to capture. Our experiences tested a collection of six basic emotion theories as described in Table 1 (see Bann[16] for details of our selection method).

Table 1 Basic Emotion sets from the most notable Basic Emotion theories that were analysed.

Basic Emotion Theory	**Identified Basic Emotions**
Izard	Anger, Contempt, Disgust, Distress, Fear, Guilt, Interest, Joy, Shame, Surprise
Russell's Categories	Angry, Depressed, Distressed, Excited, Miserable, Pleased, Relaxed, Sleepy
Plutchik	Acceptance, Anger, Anticipation, Disgust, Joy, Fear, Sadness, Surprise
Ekman	Anger, Disgust, Fear, Joy, Sadness, Surprise
Tomkins	Anger, Interest, Contempt, Disgust, Distress, Fear, Joy, Shame, Surprise
Johnson-Laird	Anger, Disgust, Anxiety, Happiness, Sadness

4.1. *Emotion Keywords*

The extraction mechanism and the selection of keywords to be mined from Twitter would form the structure of our eventual emotion corpus. Taking the union of all the emotion sets identified for analysis, we obtained a set of 21 unique emotion keywords, which, theoretically, constitutes the most distinct emotions. We extract unigrams created using the first person grammatical inflection of each keyword, similar to Russell,[26] as most tweets will contain this type of inflection: *"I am very **excited** today"* as opposed to *"I am feeling **excitement** today"*. This was chosen as opposed to mining for bigrams, for example *"feeling excited"*, *"feel excited"* and *"felt excited"*, as this resulted in far fewer tweets being returned due to Twitter's indexing focusing on single keywords. Moreover, tweets containing quantifiers would

have been ignored if we chose to extract bigrams, for example *"feeling **very** excited"*.

Contrary to Bollen's work,[27] we did not require tweets to contain the words *'feel'*, *'I'm'*, *'Im'*, *'am'*, *'being'*, and *'be'*, as an explicit mention of an emotion keyword would be sufficient to describe an experience of that emotion, reinforced by the fact that we will only be mining for the first person grammatical inflection of each keyword. We filtered out re-tweets — minimising duplicates — and negative tweets, because, for example, *'happy'* $\neq$ *'not happy'*; nor can we assume that *'not sad'* = *'happy'*. Tweets containing popular phrases which include *"Happy Birthday"* and *"Angry Birds"* were also filtered out. Initially, @ tags were not filtered, but we quickly realised that these tweets refer to messages either closely relating to other people or as part of a thread of messages; thus we filtered them out as the emotion expressed within such tweets did not describe an atomic emotional experience.We did not Porter stem collected words as Kim[28] notes that this might hide important semantic differences, for example, conceptual differences between *loved* and *loving*. To optimally harvest emotions, we substituted *fear* with *scared* as it was proven to be the most popular keyword out of *scared*, *frightened* and *afraid*. We also substituted *distressed* with *stressed*, due to an extremely low stream rate for this keyword (see Bann[16] for a detailed description of keywords).

4.2. *Emotion Streaming*

A PHP script was created that used the Gardenhose Level Twitter Streaming API — a streaming sample of about 10% of all public status updates on Twitter — that allows tracking of up to 400 keywords. We collected tweets that contained each of the selected emotion keywords, storing those which do not include a filtered phrase into a MySQL database. We programmed the PHP script to be cyclical in the sense that it streamed individual tweets, but changed emotion keywords every 5 minutes in order to collect the whole range of emotions. Ten days of data collection resulted in a labelled Temporal Emotion Database containing six emotion theories totaling to 21 unique emotion keywords each with at least 1100 documents to base our analysis on. It should be noted that by using WordNet[29] we could have expanded our initial list of 21 keywords by taking synonyms of each keyword and testing the stream rate for each emotion, selecting the most popular keyword; however in order to fairly test each theory, we opted against this as the selected emotion keywords had been carefully chosen by each theorist.

5. Semantic Emotion Analysis

Having created an emotional Twitter corpus, we analysed this data in order to evaluate the semantic distinctiveness of existing basic emotion sets, developing an iterative latent semantic clustering algorithm to discern the optimal semantically irreducible basic emotion set from all 21 emotions collected. Latent Semantic Clustering (LSC) is a simple modification of the LSA algorithm which we base our DELSAR algorithm on. Given a labelled corpus C with label set K, it calculates, using LSA, the semantic accuracy of each $label \in K$, thus providing an analysis of how distinct the labelling of C and the selection of K is. All analysis was performed on an Intel Core 2 U7700 CPU 2x1.33GHz with 2GB RAM using the GENSIM framework for Python[30] to create LSA spaces. Unless specified, we tested dimensions of the LSA space in increments of 10 and selected the dimensionality that performed optimally for each task, similar to Recchia.[31] For all tasks, we use Log-Entropy normalisation as our Association Function, found to generate optimal results[32] and recommended for LSA.[23]

5.1. *DELSAR*

Document-Emotion Latent Semantic Algorithmic Reducer (DELSAR) takes an emotion set and clusters each document's emotion to the emotion of its closest document vector (excluding itself), calculating a clustering accuracy for each emotion. The closest document vector is calculated as the maximum cosine value of the angle between the current document and each other document in the subcorpus. The emotion keyword in each document is removed before the closest document vector is calculated, so we focus purely on the words surrounding the emotion keyword for each document. DELSAR operates in the LSA space created from the subcorpus of all documents matching all emotion keywords in the set being analysed, in which there are $(doc\_limit \times number\_of\_emotions)$ documents.

If a document expressing a certain emotion, e, is not clustered with a document of the same emotion, then the words surrounding e is more similar to the words surrounding another emotion. Thus the clustering accuracy of an emotion set corresponds to how distinct that emotion set is; the more semantically accurate an emotion set is, the more distinct the language surrounding each emotion within the set is.

The reduction aspect of DELSAR initially starts with the set of all 21 emotions. After calculating the clustering accuracies for each emotion, it removes the least accurate emotion from the set and iterates until there

are n emotions remaining in the initial set, resulting in the optimal semantically distinct basic emotion set. The DELSAR algorithm is described in Algorithm 5.1.

Algorithm 5.1 DELSAR

Require: Final keyword set size *reduceTo*, Corpus **C** and Keyword Set **K**, where $\forall document \in \mathbf{C}\ \exists document \rightarrow emotion \in \mathbf{K}$
calculate cosine document similarity matrix of LSC(**C**, **K**)
for each *document* $\in$ **C do**
  delete *emotion* in *document*
  Find closest document vector *nearest* where $nearest \neq document$
  if
  $nearest(\mathbf{K}) == document(\mathbf{K})$ **then**
    document is a hit
  else
    document is a miss
  end if
end for each
for each *emotion* $\in$ **K do**
  calculate accuracy of *emotion* using (total *document* hits where *emotion* in *document*/total *document* where *emotion* in *document*)
end for each
if
length(**K**) $>$ *reduceTo* **then**
  delete least accurate *word* in **K**
  DELSAR(reduced **K**)
else
  return **K**
end if

We performed DELSAR1000 on the corpus and various subcorpora and report the results in Table 2. Note that DELSAR creates an LSA space of all documents within each emotion set; for each basic emotion set an LSA space of $(1000 \times number\_of\_emotions)$ documents is created. Evaluating all sets, our results show the accuracy of clustering each document to its nearest document, whether it is the same or another emotion. Of all the theories analysed, Ekman's set proved to be the most semantically distinct, with a 2.9% increase in accuracy compared to the average of the remaining sets. Russell's categories performed worst, which is surprising seeing as these emotions were taken as the basis for representing the entire emotion spectrum as a whole.

We performed DELSAR on the set of all 21 emotions, reducing the set

Table 2 DELSAR clustering accuracy of each basic emotion set using a corpus comprised of 1000 documents for each emotion within each set. Standard Deviation of all models is $\sigma = 0.027$.

	Model							
Dimension	**Izard 40**	**Russell 30**	**Plutchik 30**	**Ekman 30**	**Tomkins 30**	**Oatley 30**	**All 60**	**DELSAR 40**
accepting			0.583				0.452	0.553
angry	0.390	0.409	0.400	0.429	0.391	0.468	0.248	
anticipating			0.455				0.312	
anxious						0.535	0.272	
ashamed	0.452				0.467		0.366	0.534
contempt	0.550				0.575		0.356	0.574
depressed		0.292					0.193	
disgusted	0.364		0.417	0.484	0.422	0.527	0.251	
excited		0.407					0.227	
guilty	0.426						0.339	
happy						0.411	0.255	
interested	0.561				0.560		0.460	0.603
joyful	0.482		0.518	0.565	0.507		0.397	0.519
miserable		0.413					0.272	
pleased		0.548					0.359	0.506
relaxed		0.383					0.245	
sad			0.388	0.442		0.424	0.259	
scared	0.377		0.456	0.498	0.396		0.249	
sleepy		0.445					0.332	0.591
stressed	0.454	0.376			0.481		0.295	0.502
surprised	0.416		0.491	0.505	0.414		0.295	
MEAN	0.447	0.409	0.464	**0.487**	0.468	0.473	0.306	**0.548**
STDEV	0.068	0.072	0.065	0.049	0.069	0.057	0.072	0.039

to the eight most semantically distinct dimensions of emotion, these being:

ACCEPTING ASHAMED CONTEMPT INTERESTED JOYFUL PLEASED SLEEPY STRESSED

This set achieved a significant increase in terms of accuracy over Ekman's set of 6.1%; we could say that these emotions best represent the emotion spectrum in its entirety, or in other words, the remaining emotions could be expressed as a combination or a particular degree of intensity of these emotions.

In addition to performing DELSAR1000, we tested four subcorpora of varying document sizes to observe any temporal effects and found negligible temporal variance within our results.[16]

5.2. *ELSA*

While DELSAR is highly effective, its analysis is relative to a subcorpus of documents that express all the emotions contained within a particular basic emotion set — whilst it is a good measure of showing how distinct a particular emotion set is overall, it does not allow for each emotion to be mutually independent. This is important to take into consideration as it allows us to compare emotions without the constraint of it being in a set with other emotions — for example an emotion within a set may be considered distinct only because other emotions within the set are not. Emotional Latent Semantic Analysis (ELSA) is a modified version of DELSAR, in which emotions are treated separately from one another. ELSA takes the set of all

21 emotions and, *for each emotion*, creates an LSA space using documents matching only that particular emotion, in which there are (*doc_limit*) documents. For each ELSA space, the cosine value for the closest document vector to each document is determined, and an average of these is calculated. The higher this average value is for a specific emotion, the more similar the documents are for that emotion, in other words, the emotion cluster is tightly packed. Lower values mean less similar words being used in the expression of the same emotion — the emotion cluster is more dispersed — signifying a decrease in distinctiveness. The difference between ELSA and DELSAR, is that the latter evaluates whether a particular emotion *set* is representative of the entire emotion spectrum, as opposed to seeing which *emotions* are distinct.

Evaluating each basic emotion set according to ELSA is simply a matter of averaging the corresponding values of the constituent emotions, and discerning the most semantically distinct emotions requires selecting the emotions with maximum average values. The ELSA algorithm is described in Algorithm 5.2.

Algorithm 5.2 ELSA

Require: Corpus **C** and Keyword Set **K**, where $\forall document \in \mathbf{C}\ \exists document \rightarrow emotion \in \mathbf{K}$

```
for each emotion ∈ K do
  for each document ∈ C do
    if
    document(K) == emotion then
      delete emotion in document
      calculate cosine document similarity matrix of LSA(document, C)
      Find closest document vector nearest where nearest ≠ document
    end if
  end for each
  return average(nearest)
end for each
```

We performed ELSA in a similar fashion to DELSAR — testing the same copora — and report the results in Table 3. Out of all the basic emotion sets analysed, Tomkin's set proved to contain the most semantically concentrated emotions, although it must be pointed out that Tomkin's set is identical to Ekman's set without the emotion *sad* and four other emotions added; by swapping *disgusted* for *contempt*, Ekman's set would have been optimal at 0.747.

Table 3 ELSA average cosine values using dimensions 10, 20, 30, 40, 50, 60, 70, 80, 90 and 100. Each emotion uses a corpus of 1000 documents. Standard Deviation of all models is $\sigma = 0.010$.

	Model							
	Izard	**Russell**	**Plutchik**	**Ekman**	**Tomkins**	**Oatley**	**All**	**ELSA**
accepting			0.781				0.781	0.781
angry	0.727	0.727	0.727	0.727	0.727	0.727	0.727	
anticipating			0.717				0.717	
anxious						0.744	0.744	0.744
ashamed	0.743				0.743		0.743	0.743
contempt	0.838				0.838		0.838	0.838
depressed		0.695					0.695	
disgusted	0.708		0.708	0.708	0.708	0.708	0.708	
excited		0.708					0.708	
guilty	0.713						0.713	
happy						0.694	0.694	
interested	0.724				0.724		0.724	
joyful	0.761		0.761	0.761	0.761		0.761	0.761
miserable		0.744					0.744	0.744
pleased		0.742					0.742	0.742
relaxed		0.707					0.707	
sad			0.713	0.713		0.713	0.713	
scared	0.719		0.719	0.719	0.719		0.719	
sleepy		0.704					0.704	
stressed	0.736	0.736			0.736		0.736	0.736
surprised	0.723		0.723	0.723	0.723		0.723	
MEAN	0.739	0.720	0.731	0.725	**0.742**	0.717	0.731	**0.761**
STDEV	0.038	0.019	0.026	0.019	0.039	0.019	0.033	0.034

We obtained a slightly different optimal set consisting of the eight most semantically distinct emotions compared to DELSAR, taking away *interested* and *sleepy* and adding *anxious* and *miserable*:

ACCEPTING ANXIOUS ASHAMED CONTEMPT JOYFUL MISERABLE PLEASED STRESSED

This set achieved a 1.9% increase in accuracy compared to Tomkin's set. We could say that this basic emotion set contains those emotions which are the most *atomic* in the sense that the words surrounding these emotion keywords are semantically concentrated; people using these emotions are more likely to be actually referring to these emotions due to the similarity of language across all documents. Take *happy* as an example, which is the least atomic emotion: being the least semantically concentrated means that the language that people use when using the word *happy* varies the most, either due to describing a great variety of things, being used in a great variety of contexts, or varying perceptions of what the emotion *happy* actually means.

6. Conclusion

A vast majority of computer scientists tend to use Ekman's basic emotion set for emotion categorisation, and it appears that, semantically, it is the most distinct set, with a 2.9% increase in accuracy compared to the average of the remaining sets. Using an iterative algorithm based on LSC, we have discerned a set of eight (rather than Ekman's six) basic emotion keywords that have been calculated to be the most semantically distinct. This set

performed better in all semantic tests than all of the basic emotion models analysed, with a 6.1% increase in accuracy over Ekman's basic emotion set, providing evidence that by carefully selecting emotion keywords, more of the emotion spectrum can be accounted for. It must be noted, however, that the lack of varience of surrounding words of identified basic emotions may just depict a stricter consensus on the *definition* of the word, unrelated to any emotional phenomenological hierarchy.

Emotions must be seen as relative to a specific domain; it has been recently shown that facial expressions of emotion are not culturally universal.[33] Basic emotions are ultimately not universal and are correlated with underlying thematic concerns within the corpus under analysis.

Ranking emotions and basic emotion sets using our algorithms according to a metric of semantic distinctiveness allows us to compare the similarity of compound emotions, analyse the composite properties of emotions and highlight how specific emotions interact with each other, with applications ranging from clinical assessments to emotion engineering.

Emotion may contribute to evolution on a much grander scale than previously thought. Indeed, Izard[13] suggests that the main component in evolution could be Emotion Schemas, that is, evolution of actions through imitative learning of specific emotions. Memetic theory states that the ability to imitate is the only requirement for language to occur in evolution, and it has been shown in several studies that syntax and semantics emerge spontaneously (for a discussion, see Blackmore[34]). Thus, by analysing language we should be able to reverse-engineer the imitative mechanisms of humans. Mapping such processes could shed light on an updated and, combined with genetic algorithms, a more complete model of human evolution.

7. Acknowledgements

We would like to thank Paul Rauwolf and Yifei Wang for their helpful comments.

References

1. J. Watson, *Behaviourism* (University of Chicago Press, 1930).
2. C. E. Izard, *The Face of Emotion* (Appleton-Century-Crofts, New York, 1971).
3. R. Plutchik, *A general psychoevolutionary theory of emotion*, in *Emotion: Theory, research, and experience: Vol. 1. Theories of emotion*, eds. R. Plutchik and H. Kellerman (Academic press, New York, 1980), New York, pp. 3–33.

4. J. Panksepp, *Behavioral and Brain Sciences* **5**, 407 (1982).
5. J. Gray and N. McNaughton, *Nebraska Symposium on Motivation* **43**, 61 (1996).
6. A. Ortony and T. J. Turner, *Psychological Review* **97**, 315 (1990).
7. L. F. Barrett, *Personality and Social Psychology Review* **10**, 20(February 2006).
8. K. A. Lindquist and L. F. Barrett, *Psychological Science* **19**, 898 (2008).
9. Aristotle, *Rhetoric* 350BC.
10. P. Ekman, *Emotions Revealed: Understanding Faces and Feelings* (Phoenix, 2004).
11. J. Panksepp, *The Caldron of Consciousness: Motivation, affect and self-organization - An anthology (Advances in Consciousness Research)* 2000.
12. A. Damasio, *The Feeling of What Happens: Body Emotion and the Making of Consciousness.* (Vintage, 1999).
13. C. E. Izard, *Annual review of psychology* **60**, 1 (2009).
14. D. Dennett, Quining qualia, in *Consciousness in Contemporary Science*, eds. A. J. Marcel and E. Bisiach (Oxford University Press, Oxford, 1988) pp. 42–77.
15. P. Ekman, W. V. Friesen and P. Ellsworth, *Emotion in the Human Face* (Oxford University Press, 1972).
16. E. Y. Bann, *Discovering Basic Emotion Sets via Semantic Clustering on a Twitter Corpus*, tech. rep., University of Bath (2012), www.aeir.co.uk/pub.
17. S. D. Rosenberg and G. J. Tucker, *Archives of General Psychiatry* **38**, 1331 (1979).
18. J. W. Pennebaker and L. A. King, *Journal of personality and social psychology* **77**, 1296(December 1999).
19. M. L. Newman, J. W. Pennebaker, D. S. Berry and J. M. Richards, *Personality and Social Psychology Bulletin* **29**, 665 (2003).
20. A. Bilovich and J. J. Bryson, Detecting the evolution of semantics and individual beliefs through statistical analysis of language use, in *Naturally-Inspired Artificial Intelligence - Papers from the AAAI Fall Symposium*, 2008.
21. F. R. Gill, A.J., Level of representation and semantic distance: Rating author personality from texts, in *Proc. Euro Cogsci*, 2007.
22. A. J. Gill, D. Gergle, R. M. French and J. Oberlander, Emotion rating from short blog texts, in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08 (ACM, New York, NY, USA, 2008).
23. T. K. Landauer and S. T. Dumais, *Psychological Review* (1997).
24. C. Strapparava and R. Mihalcea, Learning to identify emotions in text, in *Proceedings of the ACM Conference on Applied Computing*, March 2008.
25. A. J. Gill, D. Gergle, R. M. French and J. Oberlander, Emotion rating from short blog texts, in *Proceedings of the 26th Annual SIGCHI Conference on Human Factors in Computing Systems*, 2008.
26. J. A. Russell, *Journal of Personality and Social Psychology* **39**, 1161 (1980).
27. J. Bollen, A. Pepe and H. Mao, *CoRR* **abs/0911.1583** (2009), informal publication.
28. S. M. Kim, A. Valitutti and R. A. Calvo, Evaluation of unsupervised emo-

tion models to textual affect recognition, in *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, (Association for Computational Linguistics, Los Angeles, CA, June 2010).
29. C. Fellbaum (ed.), *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)* (The MIT Press, May 1998).
30. R. Řehůřek and P. Sojka, Software Framework for Topic Modelling with Large Corpora, in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, (ELRA, Valletta, Malta, May 2010).
31. G. Recchia and M. N. Jones, *Behavior Research Methods* **41**, p. 647 (2009).
32. P. Nakov, A. Popova and P. Mateev, Weight functions impact on lsa performance, in *EuroConference RANLP'2001 (Recent Advances in NLP*, 2001.
33. R. Jack, O. Garrod, H. Yu, R. Caldara and P. Schyns, Facial expressions of emotion are not culturally universal, in *Proceedings of the National Academy of Sciences of the United States of America*, May 2012. in press.
34. S. Blackmore, *Journal of Consciousness Studies* **10**, 19 (2003).