Journal of Classification (2019) 36:619–648
https://doi.org/10.1007/s00357-019-9307-0

Comparing the Utility of Different Classification Schemes for Emotive Language Analysis

Lowri Williams[1] · Michael Arribas-Ayllon[2] · Andreas Artemiou[3] · Irena Spasić[1]

Published online: 10 May 2019

Abstract
In this paper we investigated the utility of different classification schemes for emotive language analysis with the aim of providing experimental justification for the choice of scheme for classifying emotions in free text. We compared six schemes: (1) Ekman's six basic emotions, (2) Plutchik's wheel of emotion, (3) Watson and Tellegen's Circumplex theory of affect, (4) the Emotion Annotation Representation Language (EARL), (5) WordNet–Affect, and (6) free text. To measure their utility, we investigated their ease of use by human annotators as well as the performance of supervised machine learning. We assembled a corpus of 500 emotionally charged text documents. The corpus was annotated manually using an online crowdsourcing platform with five independent annotators per document. Assuming that classification schemes with a better balance between completeness and complexity are easier to interpret and use, we expect such schemes to be associated with higher inter–annotator agreement. We used Krippendorff's alpha coefficient to measure inter–annotator agreement according to which the six classification schemes were ranked as follows: (1) six basic emotions (α = 0.483), (2) wheel of emotion (α = 0.410), (3) Circumplex (α = 0.312), EARL (α = 0.286), (5) free text (α = 0.205), and (6) WordNet–Affect (α = 0.202). However, correspondence analysis of annotations across the schemes highlighted that basic emotions are oversimplified representations of complex phenomena and as such likely to lead to invalid interpretations, which are not necessarily reflected by high inter-annotator

✉ Irena Spasić
SpasicI@cardiff.ac.uk

Lowri Williams
WilliamsL10@cardiff.ac.uk

Michael Arribas-Ayllon
Arribas-AyllonM@cardiff.ac.uk

Andreas Artemiou
ArtemiouA@cardiff.ac.uk

[1] School of Computer Science & Informatics, Cardiff University, Cardiff, UK

[2] School of Social Sciences, Cardiff University, Cardiff, UK

[3] School of Mathematics, Cardiff University, Cardiff, UK

agreement. To complement the result of the quantitative analysis, we used semi–structured interviews to gain a qualitative insight into how annotators interacted with and interpreted the chosen schemes. The size of the classification scheme was highlighted as a significant factor affecting annotation. In particular, the scheme of six basic emotions was perceived as having insufficient coverage of the emotion space forcing annotators to often resort to inferior alternatives, e.g. using happiness as a surrogate for love. On the opposite end of the spectrum, large schemes such as WordNet–Affect were linked to choice fatigue, which incurred significant cognitive effort in choosing the best annotation. In the second part of the study, we used the annotated corpus to create six training datasets, one for each scheme. The training data were used in cross–validation experiments to evaluate classification performance in relation to different schemes. According to the F-measure, the classification schemes were ranked as follows: (1) six basic emotions (F = 0.410), (2) Circumplex (F = 0.341), (3) wheel of emotion (F = 0.293), (4) EARL (F = 0.254), (5) free text (F = 0.159) and (6) WordNet–Affect (F = 0.158). Not surprisingly, the smallest scheme was ranked the highest in both criteria. Therefore, out of the six schemes studied here, six basic emotions are best suited for emotive language analysis. However, both quantitative and qualitative analysis highlighted its major shortcoming – oversimplification of positive emotions, which are all conflated into happiness. Further investigation is needed into ways of better balancing positive and negative emotions.

Keywords Annotation · Crowdsourcing · Text classification · Sentiment analysis · Supervised machine learning

1 Introduction

Traditionally, in domains such as market research, user subjectivity has been accessed using qualitative techniques such as surveys, interviews and focus groups. The proliferation of user-generated content on the Web 2.0 provides new opportunities for capturing people's appraisals, feelings and opinions. However, the sheer scale of text data generated on the Web poses obvious practical challenges of classifying user subjectivity using traditional qualitative techniques. Text mining has emerged as a potential solution for overcoming information overload associated with reading vast amounts of text from diverse sources. Recently, sentiment analysis has emerged as an approach that aims to automatically extract and classify sentiment (the subjective component of an opinion) and/or emotions (the projections or display of a feeling) expressed in text (Liu 2010; Munezero et al. 2014).Research in this domain has focused on the problem of sentiment analysis by classifying opinionated text segments (e.g. phrase, sentence or paragraph) in terms of positive or negative polarity, e.g. (Aue and Gamon 2005; Bethard et al. 2004; Breck et al. 2007). The problem with sentiment polarity is that it combines diverse emotions into two classes. For example, negative polarity conflates sadness, fear and anger. Some domains require further differentiation to associate specific emotions with appropriate actions. For example, in monitoring counter-terrorism issues, sadness, fear and anger may require a different targeted response, e.g. counseling, media communication and anti-radicalization. The problem of classifying public reaction to terrorist activities on social media indicates a need for specificity of emotion classification rather than sentiment polarity.

In this study, we compared six emotion classification schemes, including six basic emotions (Ekman 1971), wheel of emotion (Plutchik 1980), Circumplex theory of affect (Watson and

Tellegen 1985), EARL (HUMAINE (Human-Machine Interaction Network on Emotion) 2006), WordNet–Affect (Valitutti et al. 2004) and free text classification scheme. To measure their utility, we investigated their ease of use by human annotators as well as performance of supervised machine learning when such schemes are used to annotate the training data. The study was designed as follows: (1) select a representative set of emotion classification schemes, (2) assemble a corpus of emotionally charged text documents, (3) use a crowdsourcing approach to manually annotate these documents under each scheme, (4) compare the schemes using inter-annotator agreement, and interview a selected group of annotators about their perceptions of each scheme, and (5) compare the classification performance of supervised machine learning algorithms trained on the data annotated under each scheme.

2 Emotion Classification Schemes

The main tension in the literature is whether emotions can be defined as discrete, universal categories of basic emotions, whether they are characterized by one or more dimensions, or whether they are organized hierarchically. Here we discuss five examples of classification schemes, which are summarized in Table 1.

Categorical approaches are usually theory-driven accounts that suggest basic emotions are the functional expression of underlying biological and evolutionary processes (Damasio 2000; Darwin et al. 1998; LeDoux 1998). This view is supported by empirical findings of cross-cultural studies where recognition of facial expressions identified six basic emotions: *anger, disgust, fear, happiness, sadness* and *surprise* (Ekman 1971). Basic emotions provide a simple classification scheme, which has been used in many studies on emotive language analysis, e.g. (Aman and Szpakowicz 2007; Das and Bandyopadhyay 2012; Mohammad 2012; Strapparava and Mihalcea 2008).

Wheel of emotion (Plutchik 1980) (Fig. 1) is a model that uses color to illustrate intensity of emotions and their relationships. At the centre of this model are eight basic emotions: *joy, trust, fear, surprise, sadness, disgust, anger* and *anticipation*. The emotion space is represented so that combinations of basic emotions derive secondary emotions (e.g. *joy* + *trust* = *love, anger* + *anticipation* = *aggression,* etc.). Emotion intensity is represented by color boldness, e.g. *annoyance* is less intense whereas *rage* is more intense than *anger*.

Dimensional approaches represent emotions as coordinates in a multi-dimensional space (Cambria et al. 2012). There is considerable variation among these models, many of which are formed by two or three dimensions (Rubin and Talarico 2009), which incorporate aspects of arousal and valence (e.g. (Russell 1979)), evaluation and activation (e.g. (Whissell 1989)), positive and negative (e.g. (Watson and Tellegen 1985)),

Table 1 A sample of emotion classification schemes

Type	Scheme	Size
Categorical	six basic emotions	6 classes
	wheel of emotion	32 classes
Dimensional	Circumplex	4 dimensions, 8 classes
	EARL	2 dimensions, 10 classes
Hierarchical	WordNet-Affect	6 levels, 1,484 classes

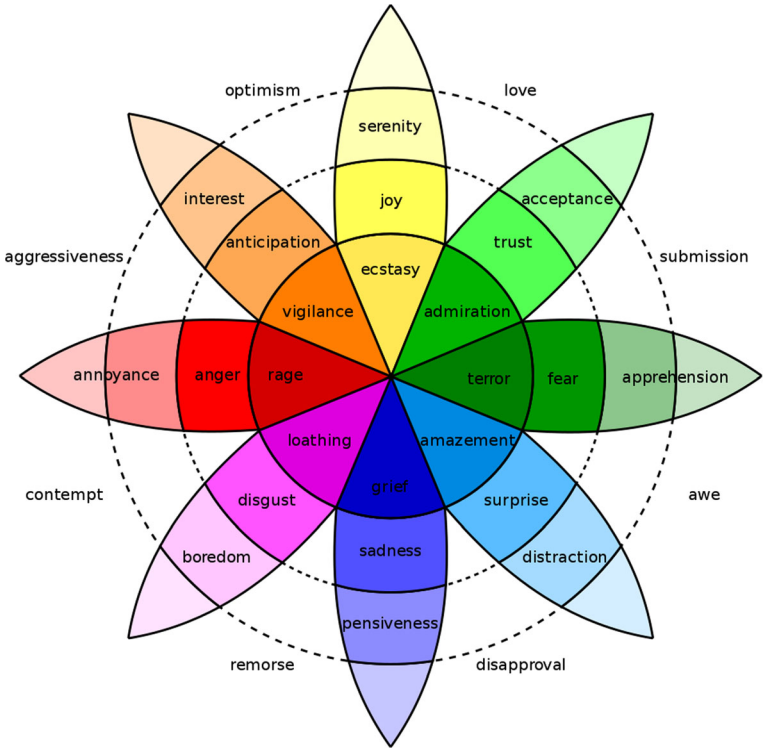

Fig. 1 Wheel of emotion (Plutchik 1980)

tension and energy (e.g. (Thayer 1997)), etc. When faceted information is needed for tasks such as emotive language analysis, dimensional models are appealing because they contain a relatively small set of categories, e.g. (Ovesdotter Alm and Sproat 2005; Strapparava and Mihalcea 2008; Cambria et al. 2012). Circumplex theory of affect (Watson and Tellegen 1985) (Fig. 2) incorporates four dimensions corresponding to *positive affect*, *engagement*, *negative affect* and *pleasantness*, each having two directions: *high* and *low*. Specific emotions are classified into one of eight categories on this scale. For example, *excitement* is classified as having *high positive affect*, *calmness* as having *low negative affect*, etc. Circumplex has been suggested as a useful model for quantifying and qualitatively describing emotions identified in text (Rubin et al. 2004).

EARL is a formal language for representing emotions in technological contexts (HUMAINE (Human-Machine Interaction Network on Emotion) 2006). Unlike schemes derived from psychological theory, EARL has been designed for a wide range of tasks in affective computing, including corpus annotation and emotion recognition. Similarly to Circumplex, EARL organizes emotions as primarily positive and negative, which are further refined based on intensity and attitude. There are five positive and five negative categories, and like Circumplex, specific emotions are given as representative examples for each category. For example, *agitation* is exemplified by *shock, stress* and *tension* (see Table 2).

The capacity of human cognition is dependent on the type and quantity of information stored in working memory. For instance, memory retention is generally shorter for longer words and longer for shorter words (Miller 1956). Humans often use hierarchical approaches

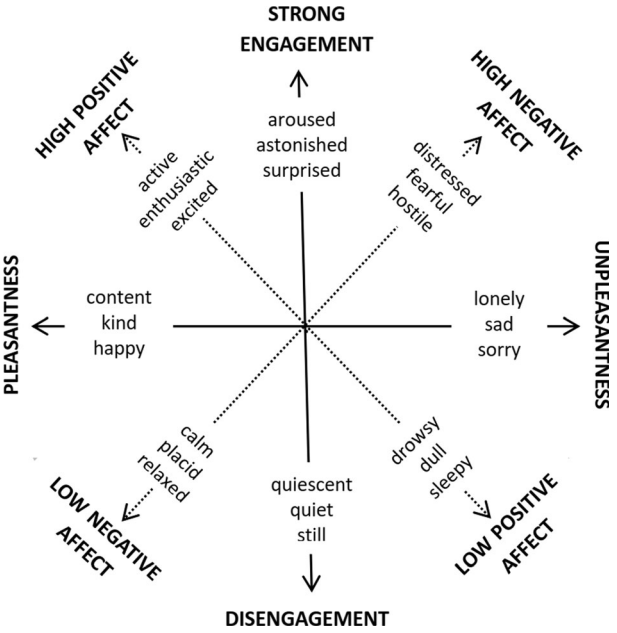

Fig. 2 Circumplex theory of affect (Watson and Tellegen 1985)

to navigate a complex conceptual space by compartmentalizing options at different levels. Unlike other schemes that contain a small, but manageable set of categories, affective hierarchies, e.g. (Laros and Steenkamp 2005; Shaver et al. 1987; Storm and Storm 1987), capture a richer set of emotions, focusing on lexical aspects that can support text mining applications. In affective hierarchies, related emotions are grouped into classes starting with *positive* and *negative* affect as top-level classes. Basic emotions (e.g. *happiness, sadness, love, anger,* etc.) are typically found at the next level of specialization. The lowest level represents instances of individual emotions (e.g. *optimistic, miserable, passionate*) (Russell and Barrett 1999).

WordNet is a lexical database of English nouns, verbs, adjectives and adverbs grouped together into sets of interlinked synonyms known as synsets (Miller 1995). WordNet has been used as a lexical resource for many text mining applications, e.g. (Agarwal et al. 2011; Fast et al. 2015; Sedding and Kazakov 2004). WordNet–Affect (Valitutti et al. 2004) was created specifically as a lexical model for classifying affects, such as moods, situational emotions, or emotional responses, either directly (e.g. *joy, sad, happy*, etc.) or indirectly (e.g. *pleasure, hurt, sorry, etc.*). It was formed by aggregating a subset of WordNet synsets into an affect hierarchy

Table 2 Emotion annotation representation language

Positive category	Examples	Negative category	Examples
Positive & lively	*joy, delight, happiness*	Negative & forceful	*annoyance, anger, contempt*
Caring	*love, affection, empathy*	Negative & not in control	*helplessness, worry, fear*
Positive thoughts	*hope, pride, trust*	Negative thoughts	*doubt, envy, guilt*
Quiet positive	*relaxed, calm, content*	Negative & passive	*sadness, hurt, despair*
Reactive	*politeness, interest, surprise*	Agitation	*shock, stress, tension*

(see Fig. 3). WordNet–Affect has been used as a lexical resource to support many sentiment analysis studies, e.g. (Balahur et al. 2010; Strapparava and Mihalcea 2008).

3 Data Collection

3.1 Text Corpus

Emotive language analysis has been applied to a range of texts from different domains. Studies have focused on emotions expressed in web logs, e.g. (Généreux and Evans 2006; Mihalcea and Liu 2006; Neviarouskaya et al. 2009), fairy tales, e.g. (Ovesdotter Alm and Sproat 2005; Francisco and Gervás 2006), novels, e.g. (Boucouvalas 2002; John et al. 2006), chat messages, e.g. (Zhe and Boucouvalas 2002; Ma et al. 2005), e-mails, e.g. (Liu et al. 2003), Twitter posts, e.g. (Tumasjan et al. 2010; Agarwal et al. 2011), etc. Twitter is a social networking service that enables users to send and read tweets – text messages consisting of up to 140 characters. Twitter provides an open platform for users from diverse demographic groups. An estimated 500 million tweets gets posted each day (Haustein et al. 2016). Information content of tweets varies from daily life updates, sharing content (e.g. news, music, articles, etc.), expressing opinions, etc. The use of Twitter as a means of self-disclosure makes it a valuable source of emotionally-charged text and a popular choice for sentiment analysis studies, e.g. (Go et al. 2009; Pak and Paroubek 2010; Kouloumpis et al. 2011). For these reasons, Twitter was selected as a source of data in the present study.

We assembled a corpus of 500 self-contained tweets, i.e. those that did not appear to be a part of a conversation. More precisely, we excluded re-tweets, replies as well as tweets that contained URLs or mentioned other users to maximize the likelihood of an emotion expressed in a tweet to refer to the tweet itself and not an external source (e.g. content corresponding to a URL). We used four criteria to identify emotionally-charged tweets. They were based on the use of idioms, emoticons and hashtags as well as automatically calculated sentiment polarity. The remainder of this section provides more detail on selection criteria.

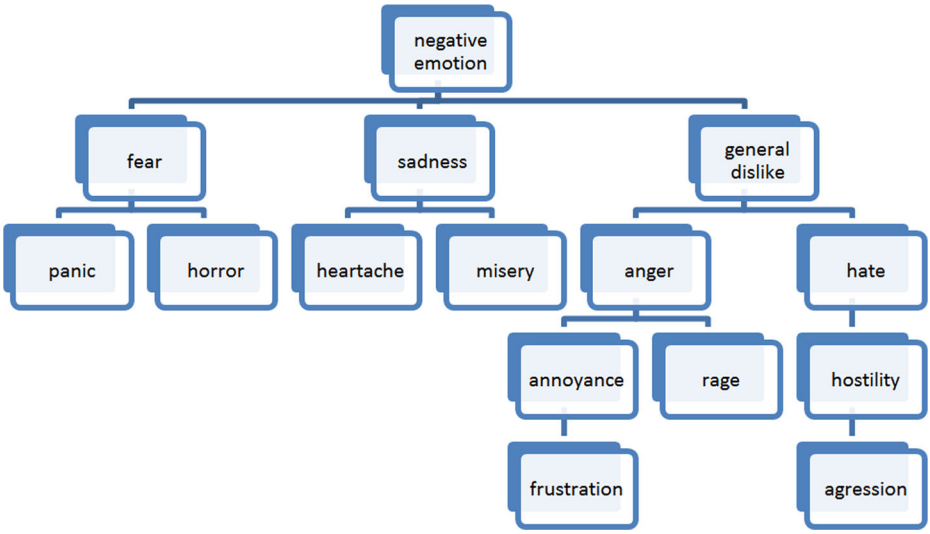

Fig. 3 An excerpt from the WordNet–Affect hierarchy

In previous studies, we demonstrated the value of idioms as pertinent features in sentiment analysis (Williams et al. 2015; Spasić et al. 2017). We found that idiom-based features significantly improve sentiment classification results. Using a set of emotionally-charged idioms described in the original study, we collected 100 tweets containing references to such idioms. The following is an example of a tweet with an idiom presented in italic typeface: "If I see a mouse in this house I will *go ballistic*."

Written online communication has led to the emergence of informal, sometimes ungrammatical, textual conventions (Purver and Battersby 2012) used to compensate for the absence of body language and intonation, which otherwise account for 93% of non-verbal communication (Mehrabian 1972). Emoticons are pictorial representations of facial expressions that seem to compensate for the lack of embodied communication. For example, the smiley face :) is commonly used to represent positive emotions. We collected 100 tweets containing emoticons. Table 3 summarizes the distribution of emoticons across these tweets.

Hashtags, i.e. words or unspaced phrases prefixed with a pound sign (#), are commonly used by Twitter users to add context and metadata to the main content of a tweet, which in turn makes it easier for other users to find messages on a specific topic (Chang 2010). Hashtags are sometimes used to flag the users' emotional state (Wang et al. 2011), which can be seen in the following example: "Sometimes I just wonder. . . I don't know what to think *#pensive*". To systematically search Twitter for emotive hashtags, we used WordNet–Affect as a comprehensive lexicon of emotive words. Our local version of the lexicon consists of 1,484 words including all derivational and inflectional forms of the word senses originally found in WordNet–Affect. We searched Twitter using these surface forms as hashtags to collect 100 tweets. The hashtags were subsequently removed from the original tweets for the following two reasons. First, we wanted the annotators to infer the emotion themselves from the main content. Second, we did not want to skew the inter-annotator agreement in favor of the WordNet–Affect as a classification scheme.

Another strategy for identifying emotionally-charged tweets involved automatically calculated sentiment polarity. We collected 116,903 tweets randomly and processed them with a sentiment annotator distributed as part of the Stanford CoreNLP (Socher et al. 2013), a suite of natural language processing tools. This method uses recursive neutral networks to perform sentiment analysis at all levels of compositionality across the parse tree by classifying a subtree on a 5-point scale: very negative, negative, neutral, positive and very positive. Figure 4 provides an example classified as very negative. We used the sentiment analysis results to select a random subset of 50 very positive and 50 very negative tweets.

Finally, we collected 100 additional tweets at random to include emotionally neutral or ambiguous tweets while correcting for bias towards certain emotions based on the choice of idioms, emoticons and hashtags. Table 4 summarizes the corpus selection criteria and distribution of the corresponding tweets selected for inclusion in the corpus.

Table 3 Distribution of emoticons across 100 tweets

Emoticon	Example	Total
:(	Need to stop having nightmares :(	56
:)	Early finish in work for a change :)	35
:D	So proud of myself right now :D	10
:P	OK... so I have a huge crush! There!! :P	2
<3	I <3 you	2

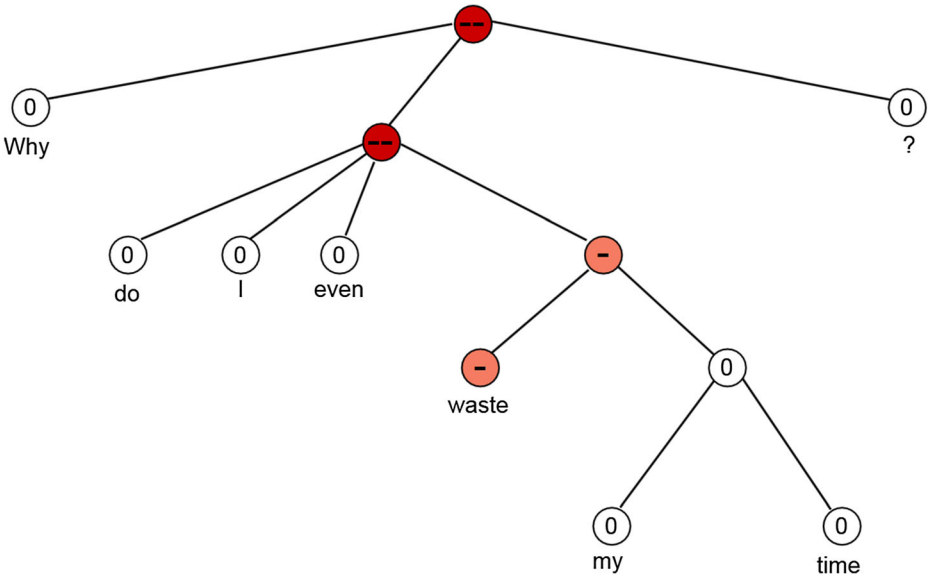

Fig. 4 An example of sentiment analysis results

3.2 Manual Annotation of Emotional Content

Crowdsourcing has become a popular method of quickly obtaining large training datasets to support supervised machine learning approaches for a variety of text mining applications including sentiment analysis, e.g. (Purver and Battersby 2012; Taboada et al. 2011). Web platforms such as CrowdFlower (CrowdFlower 2016) or Mechanical Turk (Amazon 2016) allow users to set up and distribute crowdsourcing jobs to millions of online contributors.

We used CrowdFlower to annotate text documents with respect to their emotional content. A bespoke annotation interface was designed, which consisted of three parts: input text, an annotation menu based on a classification scheme and, where appropriate, a graphical representation of the classification scheme to serve as a visual aid (see Fig. 5 for an example). To mitigate the complexity of WordNet–Affect, we implemented autocomplete functionality, where matching items from the lexicon were automatically suggested as the annotator typed into a free text field. We introduced a *neutral* category into all classification schemes to allow for the annotation of flat or absent emotional response. For example, *"Fixing my iTunes library."* was annotated as *neutral* by 23 of 30 annotators. Similarly, we introduced an *ambiguous* category to allow for annotation of cases where an emotion is present, but indeterminate in the absence of context, intonation or body language. For example, the use

Table 4 Corpus selection criteria and distribution

Criterion	Example	Total
Idiom	If I see a mouse in this house I will *go ballistic*.	100
Emoticon	What a day!!!! *:(*	100
Hashtag	Why are people so mean? *#frustrated*	100
Sentiment polarity	Why do I even waste my time?	100
Random selection	Up this early for work.	100

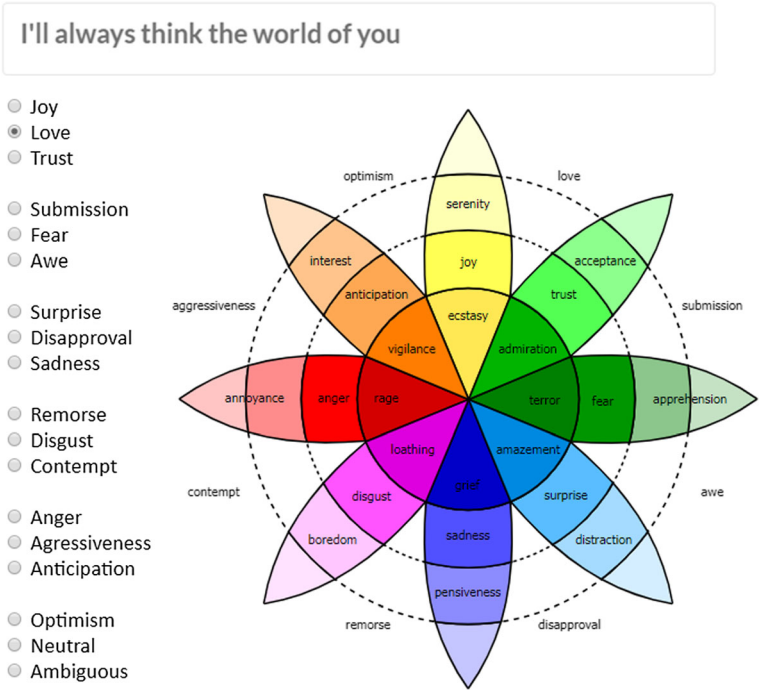

Fig. 5 An annotation example

of punctuation in "*What a day!!!!!*" clearly indicates an emotional charge, but is unclear whether the statement is positive or negative.

In addition to the schemes discussed in Section 2, we also included free text classification, where the choice of annotations was unrestricted. We specifically wanted to investigate whether a folksonomy naturally emerging from annotators' free text choices could give rise to a suitable emotion classification scheme.

Having set up 6 annotation jobs, one for each classification scheme, contributors were asked to annotate each text document with a single class that best described its emotional content. A total of 189 annotators participated in the study. Given a classification scheme, each document was annotated by five independent annotators. In total, 15,000 annotations (500 documents × 6 schemes × 5 annotations) were collected. The distributions of annotations across the schemes are shown in Fig. 6, with WordNet–Affect and free text charts displaying the distributions of the top 20 most frequently used annotations.

4 Utility Analysis: a Human Perspective

4.1 Quantitative Analysis of Annotation Results

4.1.1 Inter-Annotator Agreement

The main goal of this study was to identify an appropriate emotion classification scheme in terms of completeness and complexity, thereby minimizing the difficulty in selecting

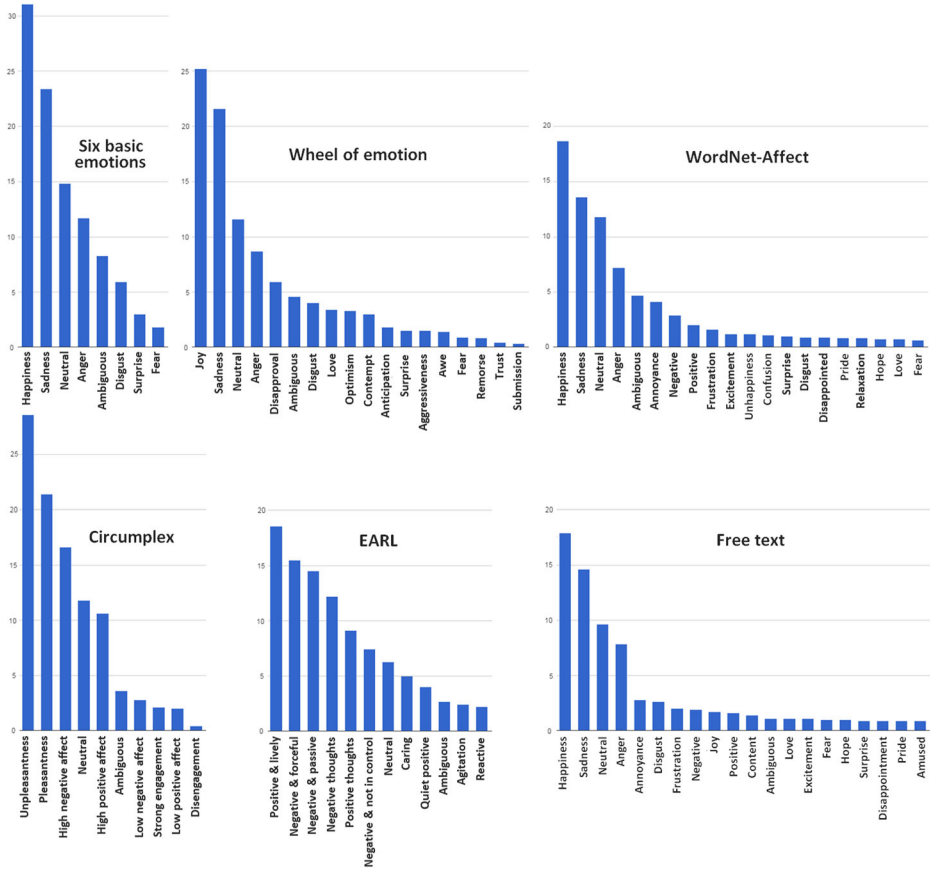

Fig. 6 Distribution of annotations across each scheme

the most appropriate class for an arbitrary text example. We hypothesize that when a correct class is available, unambiguous and readily identifiable, then the likelihood of independent annotators selecting that particular class increases, thus leading to higher inter-annotator agreement (IAA).

We used Krippendorff's alpha coefficient (Krippendorff 2013) to measure the IAA. As a generalization of known reliability indices, it was chosen because it applies to: (1) any number of annotators, not just two, (2) any number of classes, and (3) corrects for chance expected agreement. Krippendorff's alpha coefficient of 1 indicates perfect agreement, whereas 0 indicates chance agreement. Therefore, higher values indicate better agreement. We calculated Krippendorff's alpha coefficient values using an online tool (Geertzen 2016). The results are shown in Fig. 7, which also includes values of adjusted Rand index (Hubert and Arabie 1985; Steinley 2004; Steinley et al. 2016) as an alternative measure of agreement. Krippendorff's alpha coefficient of $\alpha = 0.667$ has been suggested as the trustworthy threshold of data reliability (Krippendorff 2004). With the highest value of 0.483, the IAA results in this study are well below this threshold, which is consistent with other studies on affective annotation (Devillers et al. 2005; Callejas and López-Cózar 2008; Antoine et al. 2014).

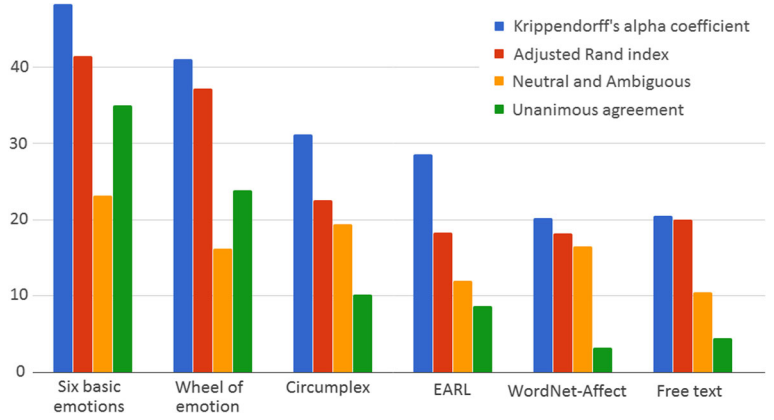

Fig. 7 Inter-annotator agreement results

Here we discuss potential reasons for low IAA. Annotation is a highly subjective process that varies with age, gender, experience, cultural location and individual psychological differences (Passonneau et al. 2008). Additionally, a text document may consist of multiple statements, which may convey different or competing emotional content. For example, there are two statements in the following sentence: "*On train going skating :) Hate the rain :(,"* each associated with a different emotion illustrated clearly by the use of emoticons. Using the wheel of emotion, this sentence received the following five annotations: *sadness, sadness, joy, love*, *ambiguous*. It can be inferred that annotators 1 and 2 focused on the latter statement, whereas annotators 3 and 4 focused on the former statement. Annotator 5 acknowledged the presence of both positive and negative emotions by classifying the overall text *ambiguous*. A genuine ambiguity occurs when the underlying emotion may be interpreted differently in different contexts (e.g. "*Another week off,"* received two *ambiguous* and three *joy* annotations), which leads to inter-annotator disagreement. Other factors such as annotators' skills and focus, the clarity of the annotation guidelines and inherent ambiguity of natural language may have also contributed to low IAA. These factors may explain low IAA, but fail to explain large variation in agreement across the annotation schemes, which ranged from 0.202 to 0.483 with a standard deviation of 11.2. Nonetheless, these results enabled a comparison of different schemes.

Unsurprisingly, given the smallest number of options, the highest IAA ($\alpha = 0.483$) and the highest number of unanimous agreements (175 out of 500, i.e. 35%) were recorded for six basic emotions. An important factor to consider here is that this scheme incurred by far the highest usage of *neutral* and *ambiguous* annotations – 576 out of 2500 (23%). This may imply that the scheme of six basic emotions has insufficient coverage of the emotion space.

Intuitively, one may expect IAA to be higher for schemes with fewer classes, as seen in some empirical studies (Antoine et al. 2014), because fewer choices offer fewer chances for disagreement. However, Krippendorff's alpha coefficient is a chance corrected measure of IAA, which suggests this may not necessarily be the case. Specifically, our study shows higher agreement for a scheme with 18 categories (the wheel of emotion) than it does for schemes of 10 or 12 classes (EARL and Circumplex). With $\alpha = 0.41$, the wheel of emotion recorded the second highest IAA. In comparison to six basic emotions, annotators resorted less frequently to using *neutral* and *ambiguous* annotations (see Fig. 7). It also recorded the second highest number of unanimous agreements (119 out of 500, i.e. 24%).

Dimensional schemes, Circumplex and EARL, both with similar number of classes (12 and 10), recorded similar levels of IAA (α = 0.312 and α = 0.286 respectively). However, an important difference between these schemes was the usage of *neutral* and *ambiguous* annotations. Circumplex incurred the second highest usage of these annotations. On the other hand, EARL had the second lowest usage of these annotations following free text annotations. This implies that with 10 generic categories, this scheme provides better coverage of the emotion space.

Due to the ambiguity and polysemy of natural language, lexical schemes, WordNet–Affect and free text, recorded the lowest IAA (α = 0.202 and α = 0.205 respectively) and incurred the fewest unanimous agreements (16 and 22 out of 500, i.e. 3% and 4% respectively). The lower IAA for WordNet–Affect may be explained by the difficulty of navigating a large hierarchy. With 262 and 260 different annotations recorded, WordNet–Affect and free text covered a wide range of emotive expressions, which provided annotators with the means of referring to a specific emotion when a suitable generic category was not available in other schemes, thus minimizing the use of *ambiguous* annotations.

To determine the significance of the differences in IAA across the schemes, we constructed confidence intervals for the given values. Given an unknown distribution of the Krippendorff's alpha coefficient, the best way to construct confidence intervals by estimation is to use bootstrap (Efron and Tibshirani 1994). We used 1000 replicate re-samples from the 500 tweets. Specifically, we randomly selected instances from the original set of 500 tweets to be included in a sample. The sampling was performed with replacement and, therefore, when a single tweet was included multiple times into the same sample, we re-used the same annotations. We then used the percentage method (Davison and Hinkley 1997) to construct 95% confidence intervals by cutting 2.5% of the replicates on each end. The confidence intervals where as follows: six basic emotions (0.4498, 0.5146), the wheel of emotion (0.3809, 0.4372), Circumplex (0.2871, 0.3348), EARL (0.2602, 0.3073), WordNet Affect (0.1826, 0.2196) and free text (0.1842, 0.2250). Where there is no overlap between the confidence intervals (see Fig. 8), we can assume that there is statistically significant difference on the IAA between the two schemes. Therefore, six basic emotions have the significantly higher IAA than all other schemes and the wheel of emotion has significantly larger agreement than Circumplex, EARL, WordNet Affect and free text. Circumplex and EARL have similar IAA, but significantly larger than WordNet Affect and free text. The last two schemes demonstrated similar IAA.

4.1.2 Establishing the Ground Truth

Emotive language analysis tasks such as subjectivity or sentiment classification can be automated using machine learning, lexicon-based or hybrid approaches (Ravi and Ravi

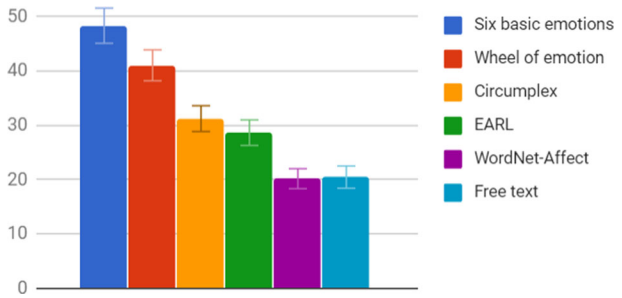

Fig. 8 Confidence intervals for the inter-annotator agreement

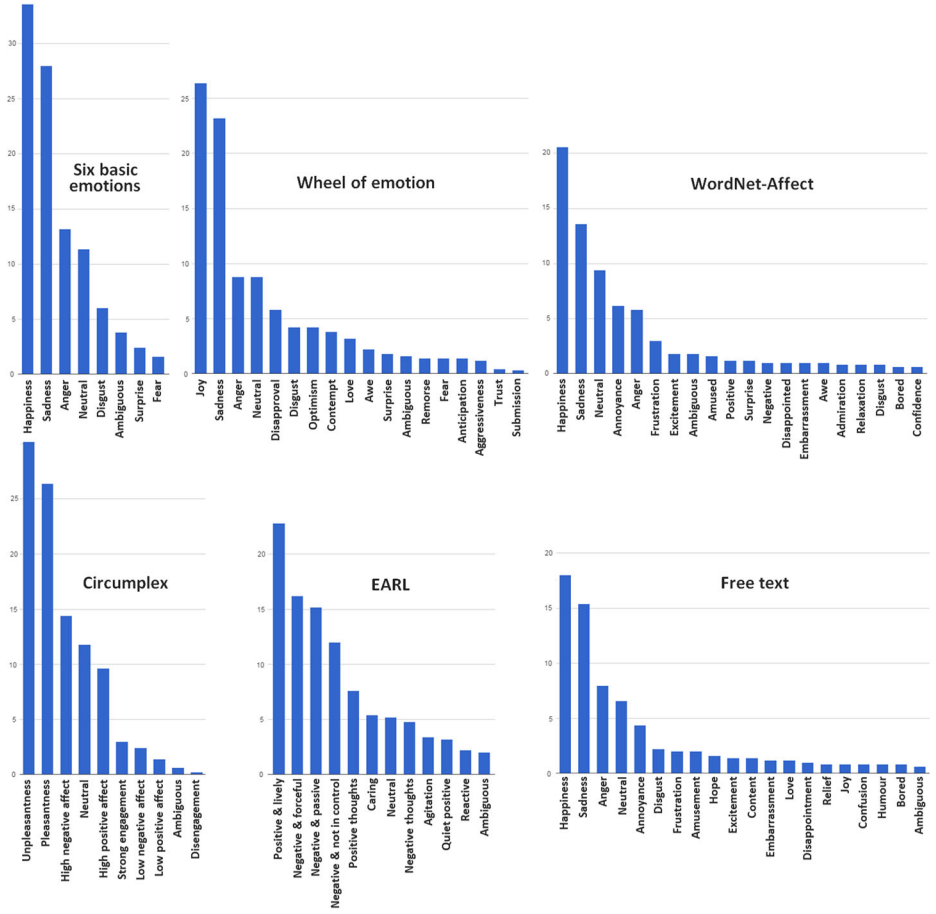

Fig. 9 Distribution of ground truth annotations

2015). The accuracy of such methods is then tested against the ground truth. In addition, supervised learning approaches require the ground truth for training purposes. When the ground truth is not readily available, human experts are asked to annotate the data. The most frequent annotation per data item is then commonly accepted as the ground truth with an expectation for the automated system to behave as the majority of human annotators. In this study, we followed the same approach. For each classification scheme, an annotation agreed by the relative majority of at least 50% was assumed to be the ground truth (see Fig. 9 for distribution of ground truth annotations). For example, using six basic emotions as the classification scheme, the sentence "*For crying out loud be quiet*" was annotated with *anger* four times and once with *disgust*, thus *anger* was accepted as the ground truth.

When no majority annotation could be identified, a new independent annotator resolved the disagreement. Table 5 (across the diagonal) provides the percentage of text instances that required disagreement resolution under each scheme. The remaining values illustrate the overlap of such text instances across the schemes. Overall 18 instances (i.e. 3.6%) required disagreement resolution under all schemes. Instances that required disagreement resolution under many schemes are likely to be genuinely ambiguous. Otherwise, the ambiguity is likely

Table 5 The percentage of instances that required disagreement resolution

	Six basic emotions	Wheel of emotions	Circumplex	EARL	WordNet-Affect	Free text
Six basic emotions	17.6	11.8	7.6	10.2	14.6	15.8
Wheel of emotions	11.8	29.6	11.8	13	22.4	23
Circumplex	7.6	11.8	22	10.2	16.4	15.4
EARL	10.2	13	10.2	37.2	22	23.4
WordNet-Affect	14.6	22.4	16.4	22	48	36.4
Free text	15.8	23	15.4	23.4	36.4	46

to be related to a given annotation scheme. In that sense, we wanted to investigate the relationships between the schemes. We performed multidimensional scaling over the data given in Table 5. Its results suggested that two directions account for 88% of the variation, so we used them to visualize the similarity of classification schemes in terms of underlying ambiguities (see Fig. 10). The first direction (along the x –axis) separates WordNet–Affect and free text from the remaining schemes. The second direction (along the y–axis) separates EARL and the wheel of emotion from the other four schemes. Both directions show the similarity between six basic emotions and Circumplex as well as the similarity between EARL and the wheel of emotion. As expected, WordNet–Affect and free text are far away from the rest and from each other indicating a much higher degree of inter-annotator disagreement.

4.1.3 Correspondence Analysis

To illustrate the difference in the coverage of different schemes, let us consider annotations of the sentence "*I'll always have a soft spot in my heart for this girl,*" (see Table 6). Despite the unanimous agreement under six basic emotions, it is still difficult to interpret the given sentence as an expression of *happiness*. Where *love* or related emotions are available, we can see a strong preference towards choosing such emotions (wheel of emotion, EARL, WordNet- Affect and free text). This point is re-enforced in the case of Circumplex, which similarly lacks a category related to *love*.

To generalize these observations and to explore the relationships between the classes across different schemes, we conducted correspondence analysis (Hirschfeld 1935), a dimension reduction method appropriate for categorical data. It is used for graphical representation of the relationships between two sets of categories. The large number of classes in WordNet–Affect and free text classification makes graphical representation of correspondence analysis involving either of these schemes highly convoluted. We, therefore, only present the results involving the four remaining schemes. For the analysis we used the ground truth annotations (see Section 4.1.2 for more details) and compared them between two schemes at a time. Figures 11, 12, 13, 14, 15, and 16 show the first two dimensions in correspondence analysis between the schemes.

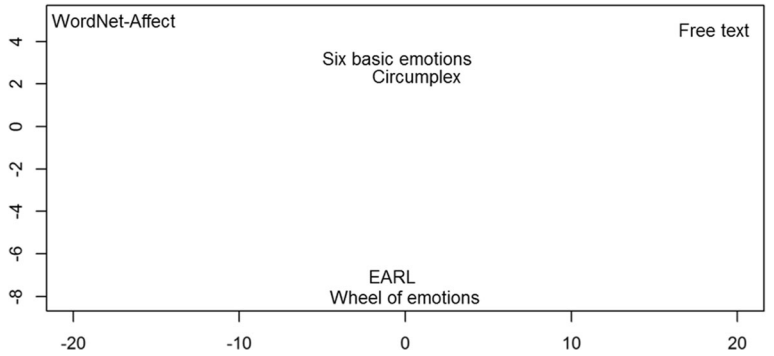

Fig. 10 Multidimensional scaling results

From the correspondence analysis between six basic emotions and Plutchik's wheel (Fig. 11), we can see that the first dimension separates positive emotions (e.g. *happiness*, *love*) on the left from the negative ones (e.g. *anger*, *sadness*) on the right. One can claim that the second direction differentiates between aggressive emotions (e.g. *anger*, *aggressiveness*) and more passive emotions (e.g. *sadness*, *fear*). If we further study the distribution of emotions across the two dimensions, we can see that four emotions from Plutchik's wheel, namely *submission*, *joy*, *love* and *awe*, correspond to a single basic emotion – *happiness*. Emotions that exist in both schemes are located close together in the graph, e.g. *anger* in the basic emotions scheme is close to *anger* in the wheel of emotion and the same applies to *surprise*, *fear*, *sadness* and *disgust*. On the other hand, it seems that emotions like *remorse*, *anticipation*, *optimism*, *disapproval*, *trust* and *aggressiveness*, which exist only in the wheel of emotion, do not correspond closely to a specific basic emotion. This supports evidence that these emotions are not redundant, i.e. cannot be abstracted easily into a basic emotion. Moreover, further analysis of the annotations across the two schemes shows that the annotators often resorted to *happiness* as the only positive basic emotion as a surrogate for a diverse range of emotions found in Plutchik's wheel including *awe*, *submission* and *love*, which do not necessarily imply *happiness*.

From the correspondence analysis between six basic emotions and Circumplex (Fig. 12), we can see that two positive classes in Circumplex, *high positive affect* and *pleasantness*, correspond to the basic emotion of *happiness*. On the other hand, some classes from Circumplex, e.g. s*trong engagement*, *low positive affect* and *low negative affect* are not

Table 6 Examples of annotation preferences

Main emotion	Scheme					
	Six basic emotions	Wheel of emotion	Circumplex	EARL	WordNet-Affect	Free text
happiness	5 × *happiness*	1 × *joy*	5 × *pleasantness* (includes *happy*)	1 × *positive and lively* (includes *joy and happiness*)	1 × *happiness*	2 × *happiness*
love		4 × *love*		4 × *caring* (includes *affection and love*)	1 × *romantic* 1 × *soft-spot* 1 × *affection* 1 × *love*	3 × *love*

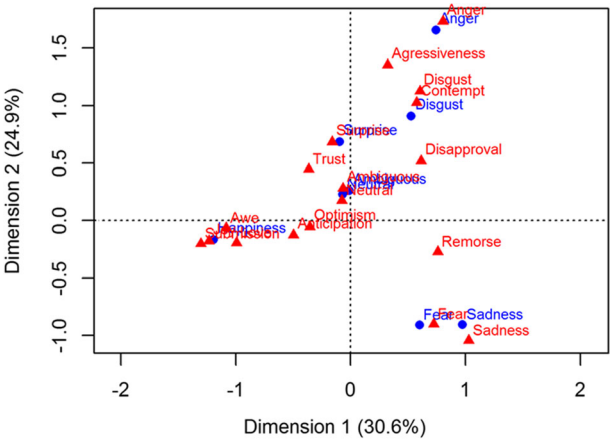

Fig. 11 Six basic emotions (blue) versus the wheel of emotion (red)

particularly close to any basic emotion. Similarly, the correspondence analysis between six basic emotions and EARL (Fig. 13), shows that all positive classes correspond to *happiness*, *negative thoughts* correspond to *fear*, *negative and forceful* corresponds to *anger*, whereas both *negative and passive* and *negative and not in control* correspond to *sadness*.

Figures 14 and 15 show how emotions from Plutchik's wheel relate to classes from Circumplex and EARL respectively. It is clear that even though Circumplex and EARL are richer than six basic emotions, they still do not seem to model emotions from Plutchik's wheel completely and unambiguously. For example, we can see from Fig. 14 that Circumplex does not have a class that corresponds to a number of emotions in Plutchik's wheel, e.g. *optimism*, *trust*, *anticipation*, *remorse* and *disapproval*. Similarly, from Fig. 15 we can see that classes from EARL do not align well against *love, surprise*, *anticipation* and *aggressiveness*. Finally, with few exceptions, Fig. 16 illustrates a clear alignment between classes in Circumplex and EARL, suggesting that they cover and partition the semantic space of emotions in a similar way.

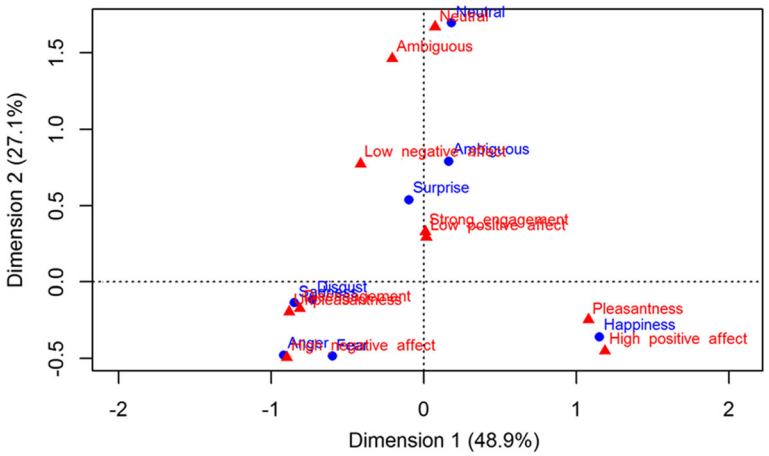

Fig. 12 Six basic emotions (blue) versus Circumplex (red)

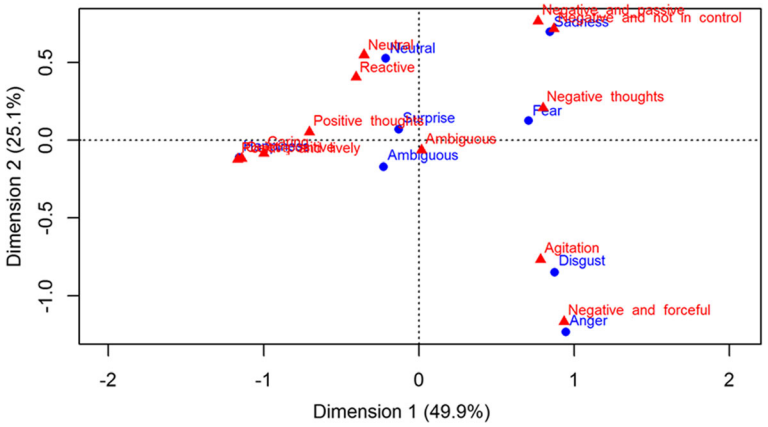

Fig. 13 Six basic emotions (blue) versus EARL (red)

To conclude, both Circumplex and EARL provide generic classes, which align fairly well (see Fig. 16). Not surprisingly, they achieved similar IAA, which is not statistically different (see Fig. 10). When compared with the schemes that use specific emotions rather than generic classes, i.e. six basic emotions and the wheel of emotion, neither of the generic schemes seem to model *surprise* well (see Figs. 12, 13, 14, and 15). In addition, even though EARL explicitly lists *love* as an example of the class *caring*, comparison with the wheel of emotion shows no strong correspondence between the two (see Fig. 15). The best results were seen with schemes that use specific emotions, with six basic emotions demonstrating significantly better IAA agreement than the wheel of emotion. However, six basic emotions require wider range of positive emotions. Figures 11 and 13 indicate that *happiness* is consistently used as surrogate for *love*. We, therefore, suggest expanding six basic emotions with *love* when using it as a classification scheme for emotive language analysis.

4.1.4 Taxonomy Versus Folksonomy

The use of WordNet–Affect and free text gave rise to a relatively large number of distinct annotations, which made their use in correspondence analysis impractical. However, we still

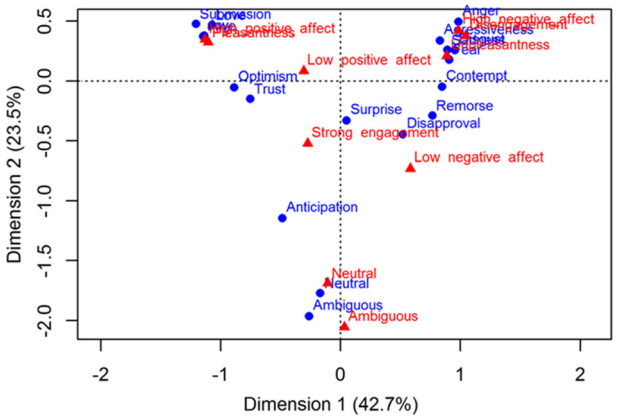

Fig. 14 The wheel of emotion (blue) versus Circumplex (red)

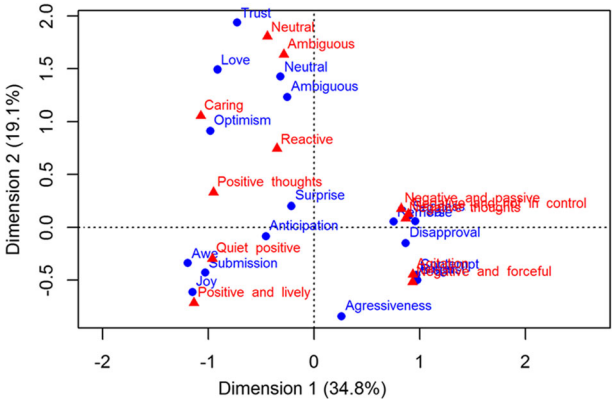

Fig. 15 The wheel of emotion (blue) versus EARL (red)

wanted to explore the difference in the lexical expression of emotion depending on whether their choice was restricted or not. In effect, we can view WordNet–Affect as a taxonomy whose vocabulary is used to ensure consistent annotation. Its hierarchical structure also allows us to compare these annotations in terms of their semantic similarity. However, taxonomies, which are typically defined by domain experts, do not necessarily reflect user vocabulary (Kiu and Tsui 2011). The lack of appropriate taxonomies and the rapidly increasing volume of user-generated information on the Web have given rise to folksonomies. Annotation choices, which are not restricted to a predefined vocabulary, allow folksonomies to emerge in a bottom-up manner (Laniado et al. 2007). Much needed flexibility and freedom for users to annotate information according to their own preferences may make folksonomies inferior to taxonomies in terms of their ability to support search and browse functions mainly because of their flat structure. Therefore, much effort has been put into organizing folksonomies hierarchically (Laniado et al. 2007) or hybridizing them with taxonomies (Kiu and Tsui 2011).

In this study, we included free text annotation to investigate whether a folksonomy naturally emerging from annotators' free text choices could give rise to a suitable emotion classification scheme. To impose a structure on this folksonomy, we aligned it against WordNet–Affect. The two sets of annotations overlapped on a total of 2,107 (42%) individual annotations, which corresponded to 74 distinct words and 63 tree nodes in WordNet–Affect. We extracted the

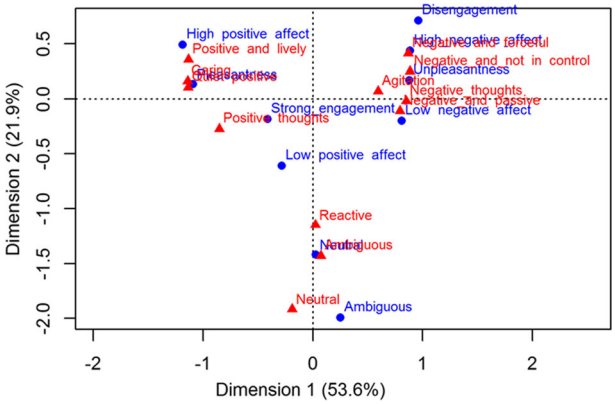

Fig. 16 Circumplex (blue) versus EARL (red)

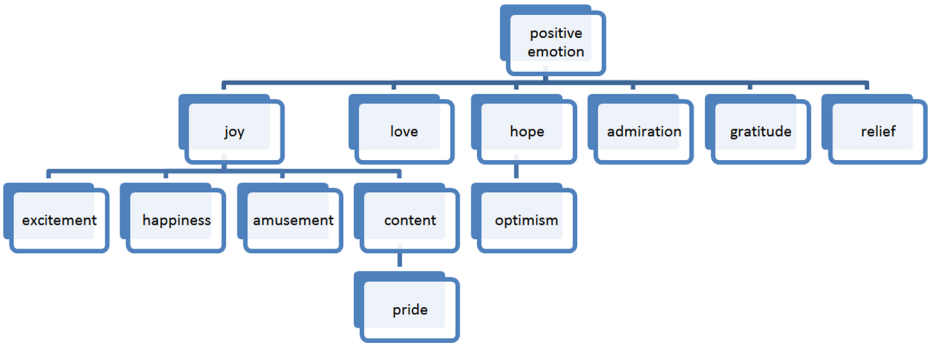

Fig. 17 A folksonomy of positive emotions

corresponding subtree from WordNet Affect and used the frequency with which individual nodes were used to further prune the tree by merging rarely used ones with their closest ancestor. We then analyzed 247 (5%) free text annotations not found in WordNet–Affect, which corresponded to 127 distinct words. Most of these words were only used five times or less and were excluded from further consideration. We analyzed the remaining five words and tried to map them onto previously extracted hierarchy. Two frequently used annotations *humor* and *funny*, were mapped onto an existing node – *amusement*. The other three frequently used annotations, *ill*, *tired* and *exhausted*, were merged into a single concept – *fatigue*, which was then added to the hierarchy. As a result, we organised the folksonomy into a hierarchy of 12 positive and 14 negative emotions (see Figs. 17 and 18). This has reduced the original WordNet–Affect hierarchy from a total of 278 nodes and 11 levels to a manageable hierarchy of 27 nodes and 5 levels.

4.2 Qualitative Analysis of Annotators' Perceptions

In order to gain a qualitative insight into how human annotators interpret and use the schemes, we conducted semi-structured interviews with 6 participants who had an academic background in social sciences. The annotation guidelines were explained to participants. Each participant was given a different sample of five text documents to annotate. They annotated the sample six times, once for each classification scheme. The order in which schemes were used for annotation was randomized for each participant. Their experiences were then discussed in a

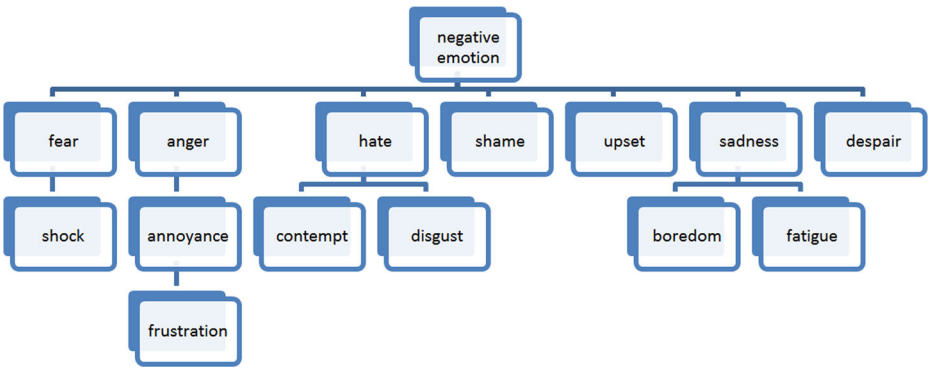

Fig. 18 A folksonomy of negative emotions

Table 7 Semi-structured interview guide

Question	Prompt
How much effort was required to annotate whilst using this scheme?	What factors of this scheme made annotation easy/difficult? What factors of this scheme did you like/dislike?
How did the number of classes affect your annotation choice?	Did more options confuse you? Do you feel restricted with the number of classes on offer? Did it affect how much time you took when making a decision? Do you think this affected the annotation accuracy?
How accurate do you think your annotates were?	Were you annotating with a class most similar to the emotion you had interpreted? Is it fair to say, without neutral and ambiguous, you'd be misclassifying? What was your thought process when an emotion was not available?
Were visual aids helpful during the exercise?	Did the color scheme/ scheme structure influence your annotation or mean anything to you? How could it be improved? Why do you think this would improve it? What in particular was confusing about…?
What was your thought process when the text reflected multiple emotions?	Were you torn between two emotions? Were you certain about your annotation? Can you provide an example? Did you resort to using neutral and ambiguous? Would allowing you to choose two or more emotions be more helpful?

semi-structured interview. Table 7 provides the semi-structured interview guide. The interviews were recorded and transcribed verbatim. We conducted thematic analysis of the transcripts. The extracted themes (see Table 8) were related to annotators (subjectivity and certainty), data (context, ambiguity and multiplicity) and the schemes (coverage and complexity).

Generally, participants found the annotation task difficult, often not feeling confident about their choice. Annotators agreed that features such as punctuation (e.g. !) and words with strong sentiment polarity (e.g. *beautiful, amazing, disgusting* and *horrible*) were strong indicators of an emotion. The annotation choices for utterances that conveyed multiple emotions varied greatly across the annotators. For example, "*My dress is so cute ugh. Praying no one wears the same one or else I will go ballistic,"* was interpreted to express both a positive and a negative emotion.

When context was absent (e.g. *"Please stay away"*), participants required more time to find an appropriate annotation. Annotators found themselves reading the text with different intonation in order to re-contextualize the underlying emotion. Upon failing to identify the context, annotators doubted their original annotation choice, claiming they may have overcompensated for the lack of context.

One significant factor affecting annotation was the number of classes available in a scheme. In particular, for six basic emotions, annotators found that the classes were meaningful or relevant for distinguishing among the polarity of the text, but not the types of emotion expressed. This is consistent with the results of correspondence analysis described in Section 4.1, which decried *happiness* as a poor surrogate for *love*. The insufficient coverage of the emotion space in this scheme significantly restricted the choices, resulting in poor capture of the primary emotion conveyed. It became unsatisfying for participants to annotate with a class that was not fit for purpose, i.e. classes did not map easily onto the emotional content. For example, *"So proud of myself*

Table 8 The summary of thematic analysis

Theme	Definition	Examples
Subjectivity	interpreting text differently	"*Gonna hate being home alone tonight*...I'm imagining it as an everyday situation, she's not scared… it's not sincere, it's not a very strong emotion."
		"To me, 'father' is quite a distant term, so I'm not sure how the person feels about their dad…"
Certainty	doubting their choices	"*Happiness* just doesn't do it. It's not quite there. But if I were to re-annotate, it'd probably be *ambiguous*..."
		"At first I put *pleasure*, I read it, and then thought 'that's not what I mean.' So I changed it to *positive and lively*..."
		"I was torn between *negative thoughts* and *negative and not in control*..."
Context	having insufficient information	"I wasn't sure if it meant what it meant…"
		"I've put *sadness*, but it could also be *love*. It depends on the tone and when it was said…"
		"The context is really important…"
		"The sentence would be different if there was an exclamation mark at the end… the full stop to me means *anger*…"
		"There's nothing in the text that's giving it away…"
Ambiguity	multiple possible interpretations	"I've got *ambiguous*. If it's about a job she could be anxious, or she could be ambitious and raring to go…"
Multiplicity	a range of emotions associated with a single sentence	"There were more options, but I wanted to choose two or three top ones or rate them in order…"
		"I was torn between *disgust* and *contempt*…"
Coverage	how well the scheme covers the emotion space	"I picked what I thought was the best, but I didn't think they fitted that well…"
		"I wanted *excited* or a similar emotion…"
Complexity	the perception of complexity of the scheme including its presentation	"It'd have to be explained to me before annotation, otherwise I'd just gloss over it…"
		"I'd like to have the basic emotions in the middle as it's the starting point, and gradually work out to more specific emotions…"
		"I thought it was a positive statement, but I wasn't sure what kind, so I looked at the words under each category which helped me decide why it was *quiet positive*..."

right now :D," was annotated as *happiness*, but also construes *pride*, an emotion that is distinct from *happiness* (Sullivan 2007).

When faced with the wheel of emotion the annotators found it difficult to choose between related emotions. For example, annotators debated whether *"I'll always think the world of you,"* expressed *love, trust* or *awe*. In this case, annotators agreed that having multiple options in terms of intensity or similarity would be more appropriate. The structure of the wheel received a negative response. Annotators agreed that it contained too much information and was quite complex to understand without additional explanation. There was debate that some emotions in the wheel (e.g. *trust*) are not necessarily emotions, but states, and questioned some emotion combinations (e.g. *anticipation* + *joy* = *optimism, sadness* + *surprise* = *disapproval*). Annotators felt that the wheel of emotion, in comparison to six basic emotions, provides better coverage, but lacks the ability to encode some emotions. For example, annotators required an emotion to represent *discomfort* for "*My throat is killing me,"* but annotated it with *disapproval, sadness* and *neutral* instead.

Annotators felt the categories in both EARL and Circumplex were not distinct. An overlap between some classes (e.g. *negative & not in control* and *negative thoughts*) was named as one

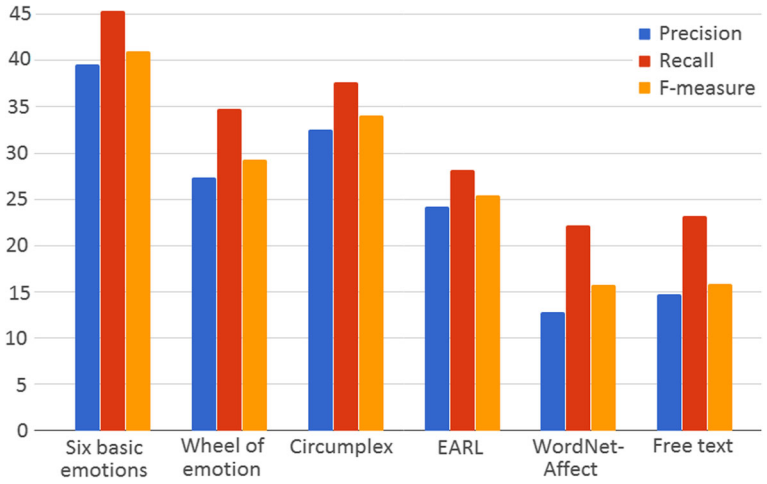

Fig. 19 The results of cross-validation experiments

of the reasons for annotators' disagreement. Although conceptually similar, the dimensional structure of Circumplex and its choice of emotions caused more resistance among annotators, as they misinterpreted the mapping of emotions onto their categories (e.g. they disagreed that *dull, sleepy* and *drowsy* were positive affects). For both EARL and Circumplex, very little attention was paid to the categories themselves. Annotators were in favor of the examples of emotions in each category, and felt that "once they had distinguished" the nuance of emotion being represented, they "had a general feeling as to which category it belonged to". Annotators appreciated having similar emotions clustered together into a generic category. This provided them with useful cues when classifying the general mood of the text, which may be easier than choosing a specific emotion. However, they acknowledged that some information would be lost when annotating with generic categories.

When faced with WordNet–Affect, annotators were able to freely decide on a specific emotion in the hierarchy (e.g. "*I don't feel well ugh,"* received *sick, miserable, unhappy* and *fed-up* annotations). Yet, annotators felt "restricted" as some of the emotions that they had interpreted in the text were not available (e.g. "I was certain it was *relief*, but it only had *relieve*, and they are not the same thing..."). The autocomplete functionality proved insufficient, as annotators continually searched for emotions that were not present in the lexicon. This

Table 9 Misclassification of opposite emotions

Main emotion	Scheme					
	Six basic emotions	Wheel of emotion	Circumplex	EARL	WordNet-Affect	Free text
happiness	39 × *sadness*	34 × *sadness*	42 × *unpleasantness* (includes *sad*)	20 × *negative & passive* (includes *sadness*)	20 × *sadness*	21 × *sadness*
sadness	44 × *happiness*	40 × *joy*	42 × *pleasantness* (includes *happy*)	25 × *positive and lively* (includes *joy and happiness*)	34 × *happiness*	17 × *happiness*

increased the time spent in completing the annotation task. Emotional content was often described in complex terms (e.g. "I chose *aggravated,* because it's stronger than *annoyance*."), or could not be pinpointed (e.g. "You know what it means and you feel it, but you can't find the right word to describe it."). In these situations, annotators would search for the synonyms of their original choices, or search for the basic form of the emotion, until an appropriate substitute was found (e.g. "I wanted *exhausted*, but had to settle for *tired*..."). This may imply that WordNet–Affect is somewhat incomplete. A recommendation for improving the navigation of this scheme is to have a drop down menu of similar emotions in addition to the autocomplete functionality.

Free text classification scheme diminished the confidence in choosing an appropriate annotation (e.g. "There is too much choice now. I think of a word and doubt. Is this what I really mean? Because there is no guideline I doubt. When there is a group, I think 'it definitely fits here'."). Annotators described this scheme as "resembling what we do in everyday life when we read a piece of text. We read something and we feel it." However, when asked to describe an emotion using a particular word, annotators could often not articulate it (e.g. "I had multiple emotions, but could not find a word to describe them all."). Free text annotations accrued a range of lexical representations of emotions with similar valences (e.g. "*My girlfriend disapproves of me :(,"* received *self-disgust, shame* and *disapproval* annotations) and intensities (e.g. "*Don't want to see the hearse coming down my road today. RIP Anna*," received *trepidation, dread, sadness* and *upset* annotations). For both lexical schemes, annotators acknowledged that "regardless of the terms we use, we are all in agreement of the general feeling expressed" in the text.

5 Utility Analysis: A Machine Perspective

Classification performance can be negatively affected by class imbalance and the degree of overlapping among the classes (Prati et al. 2004). To explore how well text classification algorithms can learn to differentiate between the classes within a given scheme, we evaluated the performance of supervised machine learning when the corresponding annotations were used to train the classification model. The ground truth annotations (see Section 4.1.2) were used to create a set of gold standards (one for each classification scheme). We then used gold standard data with Weka (Hall et al. 2009), a popular suite of machine learning software, to perform 10-fold cross-validation experiments. All text documents from the corpus of 500 tweets were converted into feature vectors using a bag-of-words representation. We tested a wide range of supervised learning methods included in Weka. Support vector machines consistently outperformed other methods. We, therefore, report the results achieved by this method. Classification performance was measured using precision (P), recall (R) and F-

Table 10 Misclassification of the neutral category

Main emotion	Scheme					
	Six basic emotions	Wheel of emotion	Circumplex	EARL	WordNet-Affect	Free text
happiness	26	18	22	11	22	11
sadness	22	19	24	6	8	6

Table 11 Misclassification of active and passive negative emotions

Main emotion	Scheme					
	Six basic emotions	Wheel of emotion	Circumplex	EARL	WordNet-Affect	Free text
anger, annoyance or disgust	18	29	23	11	17	26
sadness	7	5	11	6	4	9

measure (F) and results are given in Fig. 19. One may try to assess the significance using bootstrap confidence intervals of the cross-validation results, but it has been reported that such practice may lead to bias estimation and, therefore, should be avoided (Vanwinckelen and Blockeel 2012).

The ranking of the classification schemes with respect to F-measure is similar to the ranking with respect to the IAA with the exception of the wheel of emotion and Circumplex, which swapped places. F-measure ranged from 15.9% to 41.0% with standard deviation of 9.5. Notably, there was less variation in classification performance across the schemes than in IAA. Intuitively, the classification results are expected to be inversely proportional to the number of classes in the scheme. Unsurprisingly, given the smallest number of options, the highest value F-measure (F = 41.0%) was recorded for six basic emotions. However, EARL (10 classes) is ranked behind the wheel of emotion (16 classes). WordNet Affect and free text demonstrated almost identical F-measure, which was found to be at the lower end of the spectrum.To get better insight into the classification performance across the schemes, we analyzed the confusion matrices, which show how the automatically predicted classes compare against the actual ones from the gold standard. For each scheme, confusion often occurred between opposite emotions, *happiness* and *sadness* (see Table 9). These confusions may be explained by the limitations of the bag-of-words approach, which ignores the text structure hence disregarding compositional semantics. Specifically, negation, which can reverse the sentiment of a text expression, was found to contribute to confusion. For example, "*Why do you not love me? Why?*: *(,"* was automatically classified as *pleasantness, caring* or *happiness*, whereas it was annotated as *unpleasantness, negative & passive* and *depression* in the gold standard. Such predictions were largely based on the use of the

Table 12 Confusion matrix for the classification predictions against Circumplex classes

			Predicted									
			a	b	c	d	e	f	g	h	i	j
Actual	Pleasantness	a	69	42	7	6	0	0	0	0	8	0
	Unpleasantness	b	42	87	5	11	0	0	0	0	6	0
	High positive affect	c	26	11	3	4	0	0	0	0	4	0
	High negative affect	d	15	23	3	23	1	1	0	0	6	0
	Low positive affect	e	3	3	0	0	0	0	0	0	1	0
	Low negative affect	f	4	6	1	1	0	0	0	0	0	0
	Strong engagement	g	3	6	2	1	0	0	0	0	3	0
	Disengagement	h	0	0	0	1	0	0	0	0	0	0
	Neutral	i	22	24	4	3	0	0	0	0	6	0
	Ambiguous	j	0	2	0	0	0	0	0	0	1	0

Table 13 Confusion matrix for the classification predictions against EARL classes

			Predicted											
			a	b	c	d	e	f	g	h	i	j	k	l
Actual	Positive & lively	a	55	12	4	11	4	2	2	20	0	2	2	0
	Negative & forceful	b	22	28	3	6	2	4	0	11	0	2	3	0
	Caring	c	5	3	14	2	1	1	0	1	0	0	0	0
	Negative & not in control	d	22	10	2	14	3	1	0	8	0	0	0	0
	Positive thoughts	e	14	4	1	5	5	1	0	7	0	0	1	0
	Negative thoughts	f	5	8	0	1	0	0	0	7	0	1	2	0
	Quiet positive	g	9	2	2	0	0	0	0	2	1	0	0	0
	Negative & passive	h	25	6	3	13	3	1	0	23	0	0	2	0
	Reactive	i	3	1	0	2	0	1	0	4	0	0	0	0
	Agitation	j	3	8	0	2	0	0	0	3	0	1	0	0
	Neutral	k	11	6	2	1	1	0	0	4	0	0	1	0
	Ambiguous	l	4	1	0	0	0	0	0	4	0	0	1	0

word *love*, which represents a text feature highly correlated with the positive classes in the training set. For example, out of 14 mentions of the word *love*, four were used in a negative context. All three negated mentions were found within the negative examples. The remaining negative mention of the word *love* was sarcastic. This example illustrates the need to include negation as a salient feature.

The second largest consistently occurring confusion was related to the *neutral* category (see Table 10), which, in the absence of discriminative features, was typically misclassified as one of two largest classes in the gold standard, i.e. either *happiness* or *sadness* (see Fig. 9). Another trend noticed across all schemes, was misclassification of active negative emotions, *anger, annoyance or disgust*, as *sadness*, and slightly less the other way around (see Table 11). Again, because this behaviour is recorded consistently across all schemes, this phenomenon may be explained by the limitations of the bag-of-words approach. Further investigation is needed to determine whether a richer feature set (e.g. additional syntactic features to differentiate between active and passive voice) would help to better discriminate between these classes.

Whereas the classification confusions discussed above were common across all schemes, it was notable that both dimensional schemes, Circumplex and EARL, demonstrated relatively more confusion across a wider range of classes (see Tables 12 and 13). This suggests that their generic categories may not be sufficiently distinctive, and, therefore, are not the best suited for emotive language analysis.

6 Conclusion

We considered six emotion classification schemes (six basic emotions, wheel of emotion, Circumplex, EARL, WordNet–Affect and free text classification scheme) and investigated their utility for emotive language analysis. We first studied their use by human annotators and subsequently analyzed the performance of supervised machine learning when their annotations were used for training. For both purposes, we assembled a corpus of 500 emotionally charged text documents. The corpus was annotated manually using an online crowdsourcing platform with five independent annotators per document. Assuming that classification schemes with a better balance between completeness and complexity are easier to interpret and use, we expect

such schemes to be associated with higher IAA. We used Krippendorff's alpha coefficient to measure IAA according to which the six classification schemes were ranked as follows: (1) six basic emotions ($\alpha = 0.483$), (2) wheel of emotion ($\alpha = 0.410$), (3) Circumplex ($\alpha = 0.312$), (4) EARL ($\alpha = 0.286$), (5) free text ($\alpha = 0.205$), and (6) WordNet–Affect ($\alpha = 0.202$). Six basic emotions were found to have a significantly higher IAA than all other schemes. However, correspondence analysis of annotations across the schemes highlighted that basic emotions are oversimplified representations of complex phenomena and as such are likely to lead to invalid interpretations, which are not necessarily reflected by high IAA. Specifically, basic emotion of *happiness* was mapped to classes distinct from *happiness* in other schemes, namely *submission*, *love* and *awe* in Plutchik's wheel, high positive affect (e.g. *enthusiastic*, *excited*, etc.) in Circumplex and all positive classes in EARL including caring (e.g. *love*, *affection*, etc.), positive thoughts (e.g. *hope*, *pride*, etc.), quiet positive (e.g. *relaxed*, *calm*, etc.) and reactive politeness (e.g. *interest*, *surprise*, etc.). Semi–structured interviews with the annotators also highlighted this issue. The scheme of six basic emotions was perceived as having insufficient coverage of the emotion space forcing annotators to resort to inferior alternatives, e.g. using *happiness* as a surrogate for *love*. Therefore, further investigation is needed into ways of better representing basic positive emotions by considering those naturally emerging from free text annotations: *love*, *hope*, *admiration*, *gratitude* and *relief*.

In the second part of the study, we wanted to explore how well text classification algorithms can learn to differentiate between the classes within a scheme. Classification performance can be negatively affected by class imbalance and the degree of overlapping among the classes. In terms of feature selection, poorly defined classes may not be linked to sufficiently discriminative text features that would allow them to be identified automatically. To measure the utility of different schemes in this sense, we created six training datasets, one for each scheme, and used them in cross–validation experiments to evaluate classification performance in relation to different schemes. According to the F-measure, the classification schemes were ranked as follows: (1) six basic emotions ($F = 0.410$), (2) Circumplex ($F = 0.341$), (3) wheel of emotion ($F = 0.293$), (4) EARL ($F = 0.254$), (5) free text ($F = 0.159$) and (6) WordNet–Affect ($F = 0.158$). Not surprisingly, the smallest scheme achieved the significantly higher F-measure than all other schemes. For each scheme, confusion often occurred between opposite emotions (or equivalent categories), *happiness* and *sadness*. These confusions may be explained by the limitations of the bag-of-words approach to document representation, which ignores the text structure hence disregarding compositional semantics. Specifically, negation, which can reverse the sentiment of a text expression, was found to contribute to confusion. Another trend noticed across all schemes was misclassification of active and passive negative emotions (e.g. *anger* vs. *sadness*). Again, this phenomenon may be explained by the limitations of the bag-of-words approach. Further investigation is needed to determine whether a richer feature set (e.g. syntactic features) would help to better discriminate between related classes.

The classification confusions discussed above were commonly found across all schemes and, as suggested, represent the effects of a document representation choice rather than specific classification schemes. However, it was notable that both dimensional schemes, Circumplex and EARL, demonstrated higher confusion across a wider range of classes. This suggests that their categories may not be sufficiently distinctive, and, therefore, are not the best suited for emotive language analysis.

To conclude, six basic emotions emerged as the most useful classification scheme for emotive language analysis in terms of ease of use by human annotators and training supervised machine learning algorithms. Nonetheless, further investigation is needed into ways of extending basic emotions to encompass a variety of positive emotions, because *happiness*, as the only representative of positive emotions, is forcibly used as a surrogate for a wide variety of distinct positive emotions.

Acknowledgements Lowri Williams gratefully acknowledges the support of the Engineering and Physical Sciences Research Council (ref: 1511905). Information on the data underpinning the results presented here, including how to access them, can be found in the Cardiff University data catalogue at https://doi.org/10.17035/d.2019.0067889599.

References

Agarwal, A., Xie, B., Ovsha, I., Rambow, O., and Passonneau, R. (2011). Sentiment analysis of twitter data. In *Proceedings of the Workshop on Languages in Social Media*, pp. 30-38.

Aman, S., and Szpakowicz, S. (2007). Identifying expressions of emotion in text. In *Proceedings of the International Conference on Text, Speech and Dialogue*, pp. 196-205.

AMAZON Mechanical TURK (2016). https://www.mturk.com.

Antoine, J.-Y., Villaneau, J. , and Lefeuvre, A. (2014). Weighted Krippendorff's alpha is a more reliable metrics for multi-coders ordinal annotations: Experimental studies on emotion, opinion and Coreference annotation. In *Proceedings of the European Chapter of the Association for Computational Linguistics*.

Aue, A., and Gamon, M. (2005). Customizing sentiment classifiers to new domains: A case study. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*.

Balahur, A., Steinberger, R., Kabadjov, M., Zavarella, V., van der Goot, E., Halkia, M., Pouliquen, B., and Belyaeva, J. (2010). Sentiment analysis in the news. In *Proceedings of the International Conference on Language Resources and Evaluation*.

Bethard, S., Yu, H., Thornton, A., Hatzivassiloglou, V., and Jurafsky, D. (2004). Automatic extraction of opinion propositions and their holders. In *Proceedings of the Association for the Advancement of Artificial Intelligence Spring Symposium*.

Boucouvalas, A.C. (2002). Real time text-to-emotion engine for expressive internet communications. In *Proceedings of the International Symposium on Communication Systems, Networks and Digital Signal Processing*, pp. 305-318.

Breck, E., Choi, Y., and Cardie, C. (2007). Identifying expressions of opinion in context. In *Proceedings of the International Joint Conference on Artificial Intelligence*.

Callejas, Z., & López-Cózar, R. (2008). Influence of contextual information in emotion annotation for spoken dialogue systems. *Speech Communication, 50*, 416–433.

Cambria, E., Livingstone, A., and Hussain, A. (2012). The hourglass of emotions. In *Cognitive Behavioural Systems*, ed. A. Esposito, A.M. Esposito, A. Vinciarelli, R. Hoffmann, and V.C. Müller, Springer, pp. 144-157.

Chang, H.-C. (2010). A new Perspective on twitter hashtag use: Diffusion of innovation theory. *Proceedings of the American Society for Information Science and Technology, 47*, 1–4.

Crowdflower. (2016). https://www.crowdflower.com/

Damasio, A.R. (2000). The feeling of what happens: Body, emotion and the making of Consciousness: Vintage.

Darwin, C., Paul, E., and Phillip, P. (1998). *The Expression of the Emotions in Man and Animals:* Harper Collins.

Das, D., & Bandyopadhyay, S. (2012). Sentence-level emotion and valence tagging. *Cognitive Computation, 4*, 420–435.

Davison, A.C., and Hinkley, D.V. (1997). *Bootstrap Methods and their Application:* Cambridge University Press.

Devillers, L., Vidrascu, L., & Lamel, L. (2005). Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks, 18*, 407–422.

Efron, B., & Tibshirani, R. J. (1994). An introduction to the bootstrap. CRC press.

Ekman, P. (1971). Universals and cultural differences in facial expressions of emotions. In *Proceedings of Nebraska Symposium on Motivation*.

Fast, E., Rajpurkar, P., and Bernstein, M.S. (2015). Text mining emergent human behaviors for interactive systems. In *Proceedings of the Conference on Human Factors in Computing Systems*.

FRANCISCO, V., and GERVÁS, P. (2006). Automated mark up of affective information in English texts. In *Text, Speech and Dialogue*, ed. P. Sojka, I. Kopecek, and K. Pala, Springer, pp. 375-382.

Geertzen, J. (2016). Inter-rater Agreement with Multiple Raters and Variables, https://nlp-ml.io/jg/software/ira/

Généreux, M., and Evans, R. (2006). Distinguishing affective states in weblog posts. In *Proceedings of the Association for the Advancement of Artificial Intelligence Spring Symposium*.

Go, A., Bhayani, R., and Huang, L. (2009). Twitter sentiment classification using distant supervision. Technical report, CS224N, Stanford University.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter, 11*, 10–18.

Haustein, S., Bowman, T. D., Holmberg, K., Tsou, A., Sugimoto, C. R., & Larivière, V. (2016). Tweets as impact indicators: Examining the implications of automated bot accounts on twitter. *Journal of the Association for Information Science and Technology, 67*, 232–238.

Hirschfeld, H. O. (1935). A connection between correlation and contingency. *Mathematical Proceedings of the Cambridge Philosophical Society, 31*, 520–524.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification, 2*, 193–218.

Humaine (Human-Machine Interaction Network on Emotion) (2006). Emotion Annotation and Representation Language (EARL) Version 0.4.0, http://emotion-research.net/projects/humaine/earl/. Available from http://emotion-research.net/projects/humaine/earl/proposal

John, D., Boucouvalas, A., and Xu, Z. (2006). Representing emotional momentum within expressive internet communication. In *Proceedings of the IASTED International Conference on Internet and Multimedia Systems and Applications*.

Kiu, C.-C., & Tsui, E. (2011). TaxoFolk: A hybrid taxonomy–folksonomy structure for knowledge classification and navigation. *Expert Systems with Applications, 38*, 6049–6058.

Kouloumpis, E., Wilson, T., and Moore, J. (2011), Twitter sentiment analysis: The good, the bad and the OMG!. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*.

Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research, 30*, 411–433.

Krippendorff, K. (2013). *Content Analysis - An Introduction to its Methodology*: SAGE Publishing.

Laniado, D., Eynard, D., and Colombetti, M. (2007), Using WordNet to turn a folksonomy into a hierarchy of concepts. In *Proceedings of the Semantic Web Application and Perspectives - Fourth Italian Semantic Web Workshop*.

Laros, F. J. M., & Steenkamp, J.-B. E. M. (2005). Emotions in consumer behavior: A hierarchical approach. *Journal of Business Research, 58*, 1437–1445.

Ledoux, J. (1998). *The Emotional Brain: The Mysterious Underpinnings of Emotional Life*: Simon & Schuster.

Liu, B. (2010). Sentiment analysis and subjectivity. In *Handbook of Natural Language Processing*, ed. N. Indurkhya, and F.J. Damerau, Chapman and Hall, pp. 627-666.

Liu, H., Lieberman, H., and Selker, T. (2003). A model of textual affect sensing using real-world knowledge. In *Proceedings of the International Conference on Intelligent User Interfaces*.

Ma, C., Prendinger, H., and Ishizuka, M. (2005). Emotion estimation and reasoning based on affective textual Interaction. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction*, pp. 622-628.

Mehrabian, A. (1972). *Silent Messages: Implicit Communication of Emotions and Attitudes:* Wadsworth Publishing Company.

Mihalcea, R., and Liu, H. (2006). A Corpus-based approach to finding happiness. In *Proceedings of the Association for the Advancement of Artificial Intelligence Spring Symposium*.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*, 81–97.

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM, 38*, 39–41.

Mohammad, S.M. (2012). #Emotional tweets. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics*.

Munezero, M. D., Suero Montero, C., Sutinen, E., & Pajunen, J. (2014). Are they different? Affect, feeling, emotion, sentiment, and opinion detection in text. *IEEE Transactions on Affective Computing, 5*, 101–111.

Neviarouskaya, A., Prendinger, H., and Ishizuka, M. (2009). Recognition of fine-grained emotions from text: An approach based on the compositionality principle. In *Modeling Machine Emotions for Realizing Intelligence*, ed. T. Nishida, and C. Faucher, Springer, pp. 179-207.

Ovesdotter Alm, C., and Sproat, R. (2005). Emotional sequencing and development in fairy Tales. In *Affective Computing and Intelligent Interaction*, ed. J. Tao, T. Tan, and R.W. Picard, Springer, pp. 668-674.

Pak, A., and Paroubek, P. (2010). Twitter as a Corpus for sentiment analysis and opinion mining. In *Proceedings of the International Conference on Language Resources and Evaluation*.

Passonneau, R.J., Yano, T., Lippincott, T., and Klavans, J. (2008). Relation between agreement measures on human labeling and machine learning performance: Results from an art history image indexing domain. *Computational Linguistics for Metadata Building, 49*.

Plutchik, R. (1980), A general Psychoevolutionary theory of emotion. In *Theories of Emotion*, ed. R. Plutchik, and H. Kellerman, Elsevier, pp. 3-33.

Prati, R.C., Batista, G.E.A.P.A., and Monard, M.C. (2004), Class imbalances versus class overlapping: An analysis of a learning system behavior. In *Advances in Artificial Intelligence*, ed. R. Monroy, G. Arroyo-Figueroa, L.E. Sucar, and H. Sossa, Springer, pp. 312-321.

Purver, M., and Battersby, S. (2012). Experimenting with distant supervision for emotion classification. In *Proceedings of the European Chapter of the Association for Computational Linguistics*.

Ravi, K., & Ravi, V. (2015). A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. *Knowledge-Based Systems, 89*, 14–46.

RUBIN, D. C., & TALARICO, J. M. (2009). A comparison of dimensional models of emotion: Evidence from emotions, prototypical events, autobiographical memories, and words. *Memory, 17*, 802–808.

Rubin, V.L., Stanton, J.M., and Liddy, E.D. (2004). Discerning emotions in texts. In *Proceedings of the Association for the Advancement of Artificial Intelligence Spring Symposium*.

Russell, J. A. (1979). Affective space is bipolar. *Journal of Personality and Social Psychology, 37*, 345–356.

Russell, J. A., & Barrett, L. F. (1999). Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant. *Journal of Personality and Social Psychology, 76*, 805–819.

Sedding, J., and Kazakov, D. (2004). WordNet-based text document clustering. In *Proceedings of the COLING Workshop on Robust Methods in Analysis of Natural Language Data*.

Shaver, P., Schwartz, J., Kirson, D., & O'connor, C. (1987). Emotion knowledge: Further exploration of a prototype approach. *Journal of Personality and Social Psychology, 52*, 1061–1086.

Socher, R. A., Perelygin, J.Y., WU, J., Chuang, C.D., Manning, A.Y., Ng, and Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Spasić, I., Williams, L., and Buerki, A. (2017). Idiom-based features in sentiment analysis: Cutting the Gordian knot. *IEEE Transactions on Affective Computing*.

Steinley, D. (2004). Properties of the Hubert-arable adjusted Rand index. *Psychological Methods, 9*, 386–396.

Steinley, D., Brusco, M. J., & Hubert, L. (2016). The variance of the adjusted Rand index. *Psychological Methods, 21*, 261–272.

Storm, C., & Storm, T. (1987). A taxonomic study of the vocabulary of emotions. *Journal of Personality and Social Psychology, 53*, 805–816.

Strapparava, C., and Mihalcea, R. (2008). Learning to identify emotions in text. In *Proceedings of the ACM Symposium on Applied Computing*.

Sullivan, G. B. (2007). Wittgenstein and the grammar of pride: The relevance of philosophy to studies of self-evaluative emotions. *New Ideas in Psychology, 25*, 233–252.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics, 37*, 267–307.

Thayer, R.E. (1997), *The Origin of Everyday Moods: Managing Energy, Tension, and Stress:* Oxford University Press.

Tumasjan, A., Sprenger, T.O., Sandner, P.G., and Welpe, I.M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Proceedings of the International Conference on Weblogs and Social Media*.

Valitutti, A., Strapparava, C., & Stock, O. (2004). Developing affective lexical resources. *PsychNology Journal, 2*, 61–83.

Vanwinckelen, G., and Blockeel, H. (2012). On estimating model accuracy with repeated cross-validation. In *Proceedings of the 21st Belgian-Dutch Conference on Machine Learning*.

Wang, X., Wei, F., Liu, X., Zhou, M., and Zhang, M. (2011). Topic sentiment analysis in twitter: A graph-based hashtag sentiment classification approach. In *Proceedings of the ACM International Conference on Information and Knowledge Management*.

Watson, D., & Tellegen, A. (1985). Toward a consensual structure of mood. *Psychological Bulletin, 98*, 219–235.

Whissell, C. (1989), The dictionary of affect in language. In *Measurement of Emotions*, ed. R. Plutchik, and H. Kellerman, Elsevier, pp. 113-131.

Williams, L., Bannister, C., Arribas-Ayllon, M., Preece, A., & Spasić, I. (2015). The role of idioms in sentiment analysis. *Expert Systems with Applications, 42*, 7375–7385.

Zhe, X., & Boucouvalas, A. C. (2002). Text-to-emotion engine for real time internet communication. In *Proceedings of the International Symposium on Communication Systems, Networks and DSPs*.