Emotions in text: Dimensional and categorical models

Article *in* Computational Intelligence · January 2012

Impact Factor: 0.67 · DOI: 10.1111/j.1467-8640.2012.00456.x

Rafael A Calvo
University of Sydney

Emotions in text: dimensional and categorical models

Rafael A. Calvo, Sunghwan Mac Kim

School of Electrical and Information Engineering

The University of Sydney

Guest editors: Diana Inkpen and Carlo Strapparava

Text often expresses the writer's emotional state or evokes emotions in the reader. The nature of emotional phenomena like reading and writing can be interpreted in different ways and represented with different computational models. Affective computing (AC) researchers often use a categorical model in which text data is associated with emotional labels. We introduce a new way of using normative databases as a way of processing text with a dimensional model and compare it with different categorical approaches. The approach is evaluated using four data sets of texts reflecting different emotional phenomena. An emotional thesaurus and a bag-of-words model are used to generate vectors for each pseudo-document, then for the categorical models three dimensionality reduction techniques are evaluated: Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (PLSA), and Non-negative Matrix Factorization (NMF). For the dimensional model a normative database is used to produce three-dimensional vectors (valence, arousal, dominance) for each pseudo-document. This 3-dimensional model can be used to generate psychologically driven visualizations. Both models can be used for affect detection based on distances amongst categories and pseudo-documents. Experiments show that the categorical model using NMF and the dimensional model tend to perform best.

1. INTRODUCTION

Emotions and affective states are pervasive in all forms of communication, including text based, and increasingly recognized as important to understanding the full meaning that a message conveys, or the impact it will have on readers. Given the increasing amounts of textual communication being produced (e.g. emails, user created content, published content) researchers are seeking automated language processing techniques that include models of emotions.

Emotions and other affective states (e.g. moods) have been studied by many disciplines. Affect scientists have studied emotions since Darwin (Darwin, 1872), and different schools within psychology have produced different theories representing different ways of interpreting affective phenomena (comprehensively reviewed in Davidson, Scherer and Goldsmith, 2003).

In the last decade technologists have also started contributing to this research. Affective Computing (AC) in particular is contributing new ways to improve communication between the sensitive human and the unemotionally computer. AC researchers have developed computational systems that recognize and respond to the affective states of the user (Calvo and D'Mello, 2010). Affect-sensitive user interfaces are being developed in a number of domains including gaming, mental health, and learning technologies. The basic tenet behind most AC systems is that automatically recognizing and responding to a user's affective states during interactions with a computer, can enhance the quality of the interaction, thereby making a computer interface more usable, enjoyable, and effective.

The AC literature has tried to remain agnostic to the controversies inherent to the different theories (Calvo and D'Mello, 2010). However, ignoring the important debates has significant limitations, because a functional AC application can never be completely divorced from underlying emotion theory. We take here cross-disciplinary approach to compare the implications of different emotion theories in the way we detect emotions in text.

A commonly found difference among emotion, and therefore computer models, is in the way emotions are represented. One popular approach involves the use of a categorical representation, in which emotions consist of labels such as boredom, frustration, anger, etc. This is the approach adopted in most of the models surveyed in (Calvo and D'Mello, 2010). An alternative approach emphasises the importance of the fundamental dimensions of valence and arousal in understanding emotional experience. Dimensional approaches have long been studied by emotion theorists (Russell, 2003) and evidence suggests the existence of at least two fundamental dimensions of emotional experience: valence (i.e. pleasure / displeasure) and arousal (i.e. activation / deactivation). Russell (2003) believed they were universal primitives and called the feeling at any point on this two-dimensional space *core affect*. Other researchers have found 'dominance' a third dimension important to represent emotional phenomena (Bradley and Lang, 1994), particularly in social situations. Dimensional approaches have been proposed in frameworks to study emotions and learning (Kort, Reilly and Picard, 2001) and also been used to build software agents (Zakharov, Mitrovic and Johnston, 2008). Other AC approaches using dimensional models have been reviewed recently (Calvo and D'Mello, 2010).

Research on automated affect recognition in text, the foci of this study, has also focused on categorical approaches (Gupta, Gilbert and Fabbrizio, 2010). In this case, the computational models automatically label documents (or parts of) using supervised and unsupervised techniques. The focus is partially caused because the labels can then be used as metadata that search engines or recommender systems can use. Somewhere in the middle between categorical and dimensional approaches are those that use ordered labels such as the work on sentiment analysis where labels are representations of *sentiments*, for example when movie reviews were studied by Kennedy and Inkpen (2006) who used positive, negative and neutral valence. We discuss this literature further in the next section.

Psychologists on the other hand have most often worked on dimensional models of affect produced by different types of stimulus such as photos (Lang, Greenwald, Bradley and Hamm, 1993) and even text (Bradley and Lang, 1999).

The goal of this study is to evaluate the merits of these two conceptualizations of emotions (a *categorical model* and a *dimensional model*) on textual corpora. The different models are likely to afford different applications. What these applications are and the accuracies that can be expected in each of them are open research questions. For example, the way people normally refer to emotions is using affective labels, dimensional approaches would likely be harder to understand and possibly less useful for designing human-computer interfaces. On the other hand language and personal differences will be part of the problem of 'representation' (Parkinson, 1995) that is the way people internalize the relationship between these emotional categories and emotional dimensions might instead be a more generalizable representation.

This study contributes an evaluation of different computational approaches based on the two models of emotion (categorical and dimensional) mentioned above. The evaluation is based on unsupervised techniques that incorporate three dimensionality reduction methods and two linguistic lexical resources.

The rest of the paper is organized as follows: In Section 2 we present representative research of the emotion models used to capture the affective states of a text. Section 3 describes the techniques of affect classification utilizing lexical resources. More specifically, it describes the role of emotion models and lexical resources in the affect classification. In addition, we give an overview of the dimension reduction methods used in the study. In Section 4 we go over the affective datasets used. Section 5 provides the results of the evaluation, before coming to our discussion in Section 6.

2. BACKGROUND

Emotions in text

Early research trying to link text and emotions includes that by social psychologists and anthropologists trying to find similarities on how people from different cultures communicate (e.g. Osgood, May and Miron, 1975).

Research aimed at understanding how people express emotions through text, or how text triggers different emotions, was conducted by Osgood (Osgood, May and Miron, 1975). Osgood used multidimensional scaling (MDS) to create visualizations of affective words based on similarity ratings of the words provided to subjects from different cultures. The words can be thought of as points in a multidimensional space, and the similarity ratings represent the distances between these words. MDS projects these distances to points in a smaller dimensional space (usually two or three dimensions). Osgood found 'evaluation', 'potency', and ‘activity’ to be the emergent dimensions. *Evaluation* quantifies how a word refers to an event that is pleasant or unpleasant, similar to the hedonic valence dimension used in this project. *Potency* quantifies how a word is associated to an intensity level, particularly strong vs. weak, equivalent to the arousal dimension used here. *Activity* refers to whether a word is active or passive.

Valence-Arousal-Dominance models have rarely been used before in computational approaches to the analysis of emotion in text (Kim, Valitutti and Calvo, 2010, Kim and Calvo, 2010). Rubin et. al. used Watson and Tellegen’s Circumplex Theory of Affect (Rubin, Stanton and Liddy, 2004) on their own dataset and found it useful to classify excerpts into the 8 categories that the model represents in a valence-arousal space. Francisco and Hervás (2007) used ANEW in combination with a Wordnet, but did not use any feature selection technique and the evaluation was restricted to 4 self-annotated stories. Evidence from psychology suggests that they are primary dimensions of affective experience (Barrett, 2006, Russell, 2003). We are even beginning to understand the neural substrates of how words are emotionally perceived in these dimensions (Kensinger and Corkin, 2004). We believe they are essential in a bio-inspired model of emotion. Other dimensional models have been used, but generally with dimensions not driven by psychological theories but the output of corpora based approaches such as Latent Semantic Analysis. Example of this work are (Bellegarda, 2010) and (Valitutti, Strapparava and Stock, 2005).

Other dimensional models have been used in the psychology literature to represent emotions. The Affective Norm for English Words (ANEW) (Bradley and Lang, 1999, Stevenson, Mikels and James, 2007) is one of several projects to develop sets of normative emotional ratings for collections of emotion elicitation objects, in this case English words. This initiative complements others by Bradley and colleagues such as the International Affective Picture System -IAPS (Lang, Greenwald, Bradley and Hamm, 1993), a collection of photographs. These collections provide values for valence, arousal and dominance for each item, averaged over a large number of subjects who rated the items using the Self-Assessment Manikin (SAM) introduced by Lang and colleagues.

Categorical approaches for representing affective states are the most commonly used and are often based on a thesaurus that defines the emotional categories. These models (and their thesauri) are based on the assumption that people using the same language have similar conceptions for different discrete emotions. For example, Wordnet, a lexical database of English terms widely used in computational linguistics research (Miller, Beckwith, Fellbaum, Gross and Miller, 1990) was extended with information on affective terms (Strapparava and Valitutti, 2004). An emotional category corresponds to a Wordnet synset (a collection of affective synonyms). WordNet-Affect (Strapparava and Valitutti, 2004) is one of several affective lexical repositories of words referring to emotional states. WordNet-Affect has an additional hierarchy of affective domain labels.

Other researchers have tried to identify words or lexical structures that are predictive of the affective states of writers or speakers (Cohn, Mehl and Pennebaker, 2004, Kahn, Tobin, Massey and Anderson, 2007, Pennebaker, Mehl and Niederhoffer, 2003). Several of these approaches rely on the Linguistic Inquiry and Word Count (LIWC) (Pennebaker, Francis and Booth, 2001), a validated computer tool that analyzes bodies of text using dictionary-based categorization. LIWC based affect-detection methods attempt to identify particular words that are expected to reveal the affective content in the text (Cohn, Mehl and Pennebaker, 2004, Hancock, Landrigan and Silver, 2007, Kahn, Tobin, Massey and Anderson, 2007). Although computational linguistics has often focused on using content words, discarding pronouns, articles and preposition (function words), Pennebaker and colleagues have found ample evidence that function words have important social psychological functions. The evidence comes from multiple disciplines. For example, the use of first person singular pronouns (e.g., "I", "me") has been shown to be been linked to negative emotions (Chung and Pennebaker, 2007, Weintraub, 1989). In one of these studies students were asked to write about coming to college. The findings showed that currently depressed students used these pronouns more often that either formerly depressed or never depressed students. In other studies Pennebaker and colleagues have studied speeches by political figures (i.e. Rudolph Giuliani former Governor of New York) and showed how the changes in word use (particularly first person singular pronouns) correlated to what is know about their private lives. Increased use of first person singular pronouns increased when he broke up with his wife and when he learned about having cancer, Evidence indicates that there might be a biological substrate to the social use of function words. Two areas of the brain dedicated to language are the Wernicke (left temporal lobe) and the Broca (left frontal lobe). Subjects with neurological damage in Wernicke's area (with Broca intact) use function words (but no content) and communicate socially well. Damage in Broca's area leads to poor social communication and lack of function words.

Text-based affect detection systems have gone a step beyond simple word matching by performing a semantic analysis of the text. For example, Gill et. al. (2008) analyzed 200 blogs and reported that texts judged by humans as expressing fear and joy were semantically similar to emotional concept words (e.g., phobia, terror for fear and delight, bliss for joy). They used Latent Semantic Analysis (LSA) (Landauer, McNamara, Dennis and Kintsch, 2007) and the Hyperspace Analogue to Language (HAL) (Lund and Burgess, 1996) to automatically compute the semantic similarity between the texts and emotion keywords (e.g., fear, joy, anger). Although this method of semantically aligning text to emotional concept words showed some promise for fear and joy texts, it failed for texts conveying six other emotions, such as anger, disgust, and sadness. So it is an open question whether semantic alignment of texts to emotional concept terms is a useful method for emotion detection.

Another approach to textual affect sensing is to construct models from large corpora of world knowledge and apply these models to identify the affective tone in texts (Akkaya, Wiebe and Mihalcea, 2009, Breck, Choi and Cardie, 2007, Liu, Lieberman and Selker, 2003, Pang and Lee, 2008, Shaikh, Prendinger and Ishizuka, 2008). For example, the word 'accident' is typically associated with an undesirable event. Hence, the presence of "accident" will increase the assigned negative valence of the sentence "I was late to work because of an accident on the freeway". The technique of using corpora to extract background knowledge can lead to inaccuracies when this corpora and the corpora to be labeled are statistically different. This approach is sometimes called sentiment analysis, opinion extraction, or subjectivity analysis because it focuses on valence of a textual sample (i.e., positive or negative; bad or good), rather than assigning the text to a particular emotion category (e.g., angry, sad). Sentiment and opinion analysis is gaining traction in the computational linguistics community and is extensively discussed in a recent review (Pang and Lee, 2008).

Supervised and unsupervised machine learning techniques have been used to automatically recognize emotion in text. Supervised techniques have the disadvantage that large annotated datasets are required for training. Since the emotional interpretations of a text can be highly subjective, more than one annotator is needed, and this makes the process of the annotation very time consuming and expensive. For this reason, unsupervised methods are normally preferred in the realm of Natural Language Processing (NLP) and emotions.

The ways supervised and unsupervised techniques can be used to process text have been discussed before. For example, Strapparava and Mihalcea compared a supervised (Naïve Bayes) and four unsupervised techniques (combinations of LSA with Wordnet Affect) for recognizing six basic emotions (Strapparava and Mihalcea, 2008). They found that the different systems have different strengths. NB, for example, was the most accurate (F1) for *Joy* but not for the other 5 emotions. Using the Wordnet Affect lexicon had the highest precision but a low recall. LSA using all the emotion words had the highest recall but the precision is lower.

Techniques for detecting emotions in text have been applied to different application domains. For example, D’Mello and colleagues (D'Mello, Craig, Witherspoon, McDaniel and Graesser, 2008) used LSA but for detecting utterance types and affect in students’ dialogue within an Intelligent Tutoring System. As it is required in a categorical model of emotions, D'Mello proposed a set of categories for describing the affect states in student-system dialogue.

Some researchers have theorized ways to combine the categorical and dimensional models. For example, while considering emotions and learning, Kort (2001) combined the two emotion models, placing categories in a valence-arousal plane. To date, most affective computing researchers have utilized and evaluated supervised methods based on the categorical emotion model.

3. METHODS

This section discusses the techniques for implementing our categorical and dimensional models. The techniques, described in Figure 1, show how in the categorical models for affect detection the vector spaces are produced through mathematical dimensionality reduction techniques. In the dimensionality based model the distances are measured on a psychologically-based three dimensional space (valence, arousal, dominance). For both models we discuss unsupervised techniques were *pseudo-documents or text units* and emotional categories are represented in a common vector space*.* All text units including sentences, paragraphs, documents, subject responses or other text units are referred here as *pseudo-documents*. Distances between both can then be measured and they can be used to label pseudo-documents with their closest category.

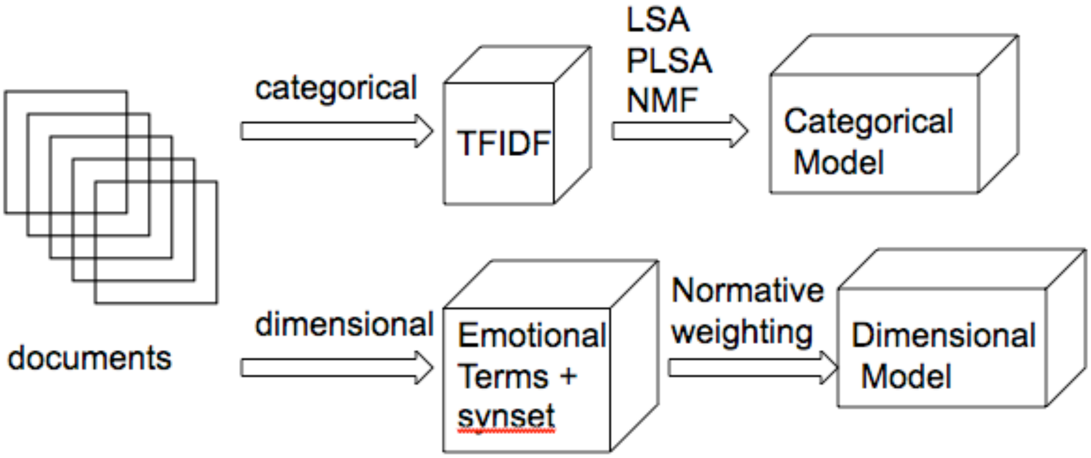

Figure 1: Categorical and Dimensional models

3. 1 Categorical Models

Pseudo-documents (i.e. sentences, paragraphs or responses) in our categorical approach, are converted to a vector space using Term Frequency/ Inverse Document Frequency (*tf-idf*), a weighting scheme developed for Information Retrieval (Baeza-Yates and Neto, 1999). More specifically, terms are encoded as vectors, whose components are co-occurrence frequencies of words in corpora pseudo-documents. Frequencies are weighted according to the log-entropy with respect to the *tf-idf* weighting schema (Baeza-Yates and Neto, 1999). The vector-based model (VSM) representation enables words, sentences, and sets of synonyms (i.e. WordNet synsets) to be represented in a unifying way as vectors. VSM provides a variety of definitions of distance between vectors, corresponding to different measures of semantic similarity.

The VSM representation can be reduced with techniques well known in Information Retrieval: Latent Semantic Analysis (LSA), Probabilistic LSA (PLSA), or the Non-negative

Matrix Factorization (NMF) representations. The dimensions produced by these statistical techniques do not have a direct psychological interpretation, and they are only to be used for the detection (i.e. classification) task that converts pseudo-documents into predefined categories, and this is why we call this approach *categorical*. These dimensionality reduction techniques reduce the computation time and noise in the data.

Latent Semantic Analysis (LSA) is one of the earliest approaches to reduce the dimension of vector representations of textual data (Landauer, McNamara, Dennis and Kintsch, 2007). LSA maps terms or pseudo-documents into a vector space of reduced dimensionality that is the latent semantic space. The mapping of the given terms/pseudo-documents vectors to this space is based on Singular Value Decomposition (SVD), a reliable technique for matrix decomposition. SVD decomposes a matrix as the product of three matrices.

$$A = U\sum V^T \approx U_k \sum_k V_k^T = A_k \qquad (\mathbf{1})$$

where A_k is the closest matrix of rank k to the original matrix. The columns of V_k represent the coordinates for pseudo-documents in the latent space.

Probabilistic Latent Semantic Analysis (PLSA) (Hofmann, 1999) has two key differences with LSA. PLSA defines proper probability distributions and the reduced matrix does not contain negative values. Based on the combination of LSA and probabilistic theories such as Bayes rules. PLSA finds the *latent topics*, the association of pseudo-documents and topics, and the association of terms and topics. In the equation (2), z is a *latent class variable* (i.e. discrete emotion category), while *w* and *d* denote the elements of term vectors and pseudo-document vectors, respectively.

$$P(d,w) = \sum_{z} P(z)P(w|z)P(d|z) \tag{2}$$

where *P(w|z)* and *P(d|z)* are topic-specific word distribution and pseudo-document distribution, individually. The decomposition of PLSA, unlike that of LSA, is performed by means of the likelihood function. In other words, *P(z), P(w|z),* and *P(d|z)* are determined by the Maximum Likelihood Estimation (MLE) and this maximization is performed using the Expectation Maximization (EM) algorithm. For pseudo-document similarities, each row of the *P(d|z)* matrix is used as a low-dimensional representation in the semantic topic space.

Non-negative Matrix Factorization (NMF) (Lee and Seung, 1999) has been successfully applied to semantic analysis. Given a non-negative matrix *A*, NMF finds non-negative factors *W* and *H* that are reduced-dimensional matrices. The product *WH* can be regarded as a compressed form of the data in *A*.

$$A \approx WH = \sum WH \tag{3}$$

W is basis vector matrix and *H* is encoded matrix of basis vectors in the equation (3). NMF solves the following minimization problem (4) in order to obtain an approximation *A* by computing *W* and *H* in terms of minimizing the Frobenius norm of the error.

$$\min_{W,H} \|A - WH\|_F^2, \qquad s.t.\ W,H \geq 0 \tag{4}$$

where $W, H \geq 0$ means that all elements of W and H are non-negative. This non-negative peculiarity is desirable for handling text data that always require non-negativity constraints. The classification of pseudo-documents is performed based on the columns of matrix H that represent the pseudo-documents.

3. 2 Dimensional Model

Experimental psychologists studying stimulus-response phenomena have used dimensional models. Subjects receive a stimulus (e.g. a photo or a text), and then report on the affective experience using a dimensional representation. Often the self-reports of many subjects are recorded producing normative databases of stimulus-response data.

ANEW (Bradley and Lang, 1999) is a set of normative emotional ratings for a collection of English words (N=1,035), where after reading the words, subjects report their emotions in a three dimensional representation. This collection provides the rated values for valence, arousal, and dominance for each word rated using the Self Assessment Manikin (SAM). For each word w, the normative database provides coordinates in an affective space as:

$$\bar{w} = (valence, arousal, dominance) = ANEW(w) \tag{5}$$

The occurrences of these words in a text can be used, in a naïve way, to weight the sentence in this emotional plane. This assumption is frequently used in the literature but naïve since words often change their meaning or emotional value when they are used in different contexts.

3.3 Classification

When a new *pseudo-document* needs to be classified, it is represented as a point in the reduced space of one of the two models (categorical or dimensional) using the approach above. The classification (or emotion detection task) consists of finding the closest emotion category to the pseudo-document.

To measure the distances cosine similarities can be calculated between vectors (i.e. pseudo-documents or emotional categories) in this representation. For the current study the similarities between a pseudo-document and a category need to be above a threshold, otherwise the input is labeled as “neutral”, meaning the absence of emotion. When the similarity is above the threshold, the input is labeled as the emotion with the highest similarity. We used a threshold (t = 0.65) for the purpose of validating a strong emotional analogy between two vectors. Our threshold value was chosen to optimize accuracy and following recommendations from previous studies (Penumatsa, Ventura, Graesser, Franceschetti, Louwerse, Hu and Cai, 2006).

If we define the similarity between a given input text, *I*, and an emotional class, E_j, as sim(I, E_j), the categorical classification result, CCR, is more formally represented as follows:

$$\mathrm{CCR}(I) = \begin{cases} \arg\max\limits_{j}\left(\mathrm{sim}(I, E_j)\right) & \text{if } \mathrm{sim}(I, E_j) \geq t \\ "\mathit{neutral}" & \text{if } \mathrm{sim}(I, E_j) < t \end{cases}$$

One class with the maximum score is selected as the final emotion class.

In the next section the accuracy of classifying pseudo-documents using the two representation spaces is compared. In the categorical space we measure distances between the pseudo-document and the emotional category in the LSA, PLSA and NMF spaces. In the dimensional representation the distances are measured in the VAD space. The VAD value of this sentence is computed by averaging the VAD values of the words:

$$\overline{sentence} = \frac{\sum_{i=1}^{n} \bar{w}}{n} \qquad (6)$$

where *n* is the total number of words in the input sentence (i.e. pseudo-document).

Since the number of words available in this normative database is limited, the chance of co-ocurrence with words in the corpora is low. Instead of only using these words we also used the synset (all synonyms) from WordNet-Affect in order to calculate the position of each emotion category. These emotional synsets are converted to the 3-dimensional VAD space and averaged for the purpose of producing a single point for the target emotion as follows:

$$\overline{emotion} = \frac{\sum_{i=1}^{k} \bar{w}}{k} \qquad (7)$$

where *k* denotes the total number of synonyms in an emotion.

In order to compare the differences amongst techniques we use emotion categories common to most of the corpora used in the evaluation (All except USE). *anger-disgust*, *fear*, *joy*, and *sadness,* found in most corpora used*,* are mapped on the VAD space. Let A_c, F_c, J_c, and S_c be the centroids of these four emotions. Then the centroids, calculated with equation (7), and a 1-9 scale in the valence, arousal and dominance axis, are as follows: A_c = (2.55, 6.60, 5.05), F_c = (3.20, 5.92, 3.60), J_c = (7.40, 5.73, 6.20), and S_c = (3.15, 4.56, 4.00). Apart from the four emotions, we manually define *neutral* to be centered (5, 5, 5). A pseudo-document is tagged with the emotion category closest to its centroid. The centroid of the pseudo-document may be close to an emotional category in the VAD space, even if they do not share any terms in common. We define a distance threshold (empirically set to 4) that must be met before making any dimensional classification.

3.4 Exploratory Visualization

The categorical and dimensional representations of pseudo-documents and emotional categories can also be used to create low dimensional visualizations. The techniques to reduce the vector space in the categorical models require at least six dimensions to represent enough of the variance contained in the original data. The number of dimensions and their lack of a psychological interpretation, makes visualizations based on categorical model not practical.

The VAD models instead are 3-dimensional vectors that can be easily visualized. Figure 2 shows the emotional dimensions and the representation of each sentence in the same space.

4. EMOTION-LABELED DATA

The following four datasets were employed in the evaluation of our dimensional and categorical techniques. The first three (SemEval, ISEAR and Fairy Tales) have 4 emotion categories in common. The fourth (i.e. USE) does not have these categories and is often discussed separately. The number of pseudo-documents in each category is shown in Table 1 and sample texts in Table 2.

News headlines

The first dataset is “Affective Text” from the SemEval 2007 task (Strapparava and Mihalcea, 2007). This dataset consists of news headlines excerpted from newspapers and news web sites. Headlines are suitable for our experiments because headlines are typically intended to evoke emotions that draw the readers’ attention. The dataset has six emotion classes: *anger*, *disgust*, *fear*, *joy*, *sadness* and *surprise*, and is composed of 1,250 annotated headlines. In contrast to other datasets SemEval, allows a sentence to have multiple tags and includes a *neutral* category.

International Survey on Emotion Antecedents and Reactions (ISEAR)

The ISEAR dataset consists of 7,666 sentences (Scherer and Wallbott, 1994), annotated by 1,096 participants with different cultural backgrounds who completed questionnaires about experiences and reactions for seven emotions: *anger*, *disgust*, *fear*, *joy*, *sadness*, *shame* and *guilt*.

Fairy Tales

Sentences in the third dataset were extracted from fairy tales (Alm, 2009) labeled with five emotion categories: *angry-disgusted*, *fearful*, *happy*, *sad* and *surprised*. This data has been used in the literature because emotions are particularly significant elements in the literary genre of fairy tales. The dataset is composed of three stories for children including Grimms’, H.C. Andersen’s, and B. Potter’s stories.

Unit of Study Evaluations (USE)

The USEs are survey instruments used in Australia to assess students' experience of a course, similar to the Student Evaluations of Teaching in the USA. The USE has 12 questions, 8 of which are standardized University-wide and 4 that are selected by each Faculty. It is designed to provide information to those seeking a) to assess the learning effectiveness of a subject, for planning and implementing changes in the learning and teaching environments, and b) to assess the contributions of units or subjects to students' learning experience in their whole degree program. The answers contain a Likert scale and a free text field handwritten by students and later typed-in for this project. The only labels available for this dataset are these ordinals: *strongly agree* and *agree* that corresponds to a positive sentiment, *neutral,* and *disagree* or *strongly disagree* that corresponds to a negative sentiment.

The USEs used in this study were collected from courses taught by two academics over a period of six years were used to create the dataset. After removing responses to question 4 (a question on workload that has a different structure), the dataset contained a total of 909 questionnaires (each with 11 ratings), and out of the possible 9,999, students responded with 3,008 textual responses (each expected to be a description of a rating), a textual response rate of 30.1 %. Out of these we removed internal referencing (e.g. 'see above') and meaningless text (e.g. '?'). The questionnaire and results are described in more detail in (Kim and Calvo, 2010).

Table 1: Number of sentences labeled with each emotion. USE dataset has a different set.

Emotion	SemEval	ISEAR	Fairy tales	Total
Anger-Disgust	62	2,168	218	2,448
Fear	124	1,090	166	1,380
Joy	148	1,090	445	1,683
Sadness	145	1,082	264	1,491

Table 2: Sample sentences in different corpora

Dataset	Sentences tagged with *Sadness/Sad*
SemEval	Bangladesh ferry sink, 15 dead.
ISEAR	When I left a man in whom I really believed.
Fairy tales	The flower could not, as on the previous evening, fold up its petals and sleep; it dropped sorrowfully.
USE	lecturer and tutor was helpful and explained concepts well.

5. EVALUATION OF AFFECT DETECTION TECHNIQUES

The goal of affect detection is to predict a single emotional label given an input sentence. The evaluation in Table 3 shows Majority Class Baseline (MCB) as the baseline algorithm. The MCB is the performance of a classifier that always predicts the majority class. In SemEval and Fairy tales the majority class is *joy*, while *anger-disgust* is the majority emotion in case of ISEAR. The five approaches were evaluated on the dataset of 479 news headlines (SemEval), 5,430 responses to questions (ISEAR), and 1,093 fairy tales' sentences. We define the following acronyms to identify the approaches: LSA-based categorical classification (CLSA), PLSA-based categorical classification (CPLSA), NMF-based categorical classification (CNMF) and Dimension-based estimation (DIM).

Table 3: Emotion Identification results. USE has different labels so cannot be included.

Data set		SemEval			ISEAR			Fairy tales		
Emotion		Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
Anger-Disgust	MCB	0.000	0.000	-	0.399	**1.000**	0.571	0.000	0.000	-
	CLSA	0.089	0.151	0.112	0.468	0.970	**0.631**	0.386	**0.749**	0.510
	CPLSA	0.169	**0.440**	0.244	0.536	0.397	0.456	0.239	0.455	0.313
	CNMF	**0.294**	0.263	**0.278**	0.410	0.987	0.579	**0.773**	0.560	**0.650**
	DIM	0.161	0.192	0.175	**0.708**	0.179	0.286	0.604	0.290	0.392
Fear	MCB	0.000	0.000	-	0.000	0.000	-	0.000	0.000	-
	CLSA	0.434	0.622	0.511	0.633	0.038	0.071	**0.710**	0.583	0.640
	CPLSA	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	CNMF	**0.525**	**0.750**	**0.618**	**0.689**	0.029	0.056	0.704	**0.784**	**0.741**
	DIM	0.404	0.404	0.404	0.531	**0.263**	**0.351**	0.444	0.179	0.255
Joy	MCB	0.309	**1.000**	0.472	0.000	0.000	-	0.407	**1.000**	0.579
	CLSA	0.455	0.359	0.402	0.333	0.061	0.103	**0.847**	0.637	0.727
	CPLSA	0.250	0.258	0.254	0.307	0.381	0.340	0.555	0.358	0.436
	CNMF	**0.773**	0.557	0.648	**0.385**	0.005	0.010	0.802	0.761	0.781
	DIM	0.573	0.934	**0.710**	0.349	**0.980**	**0.515**	0.661	0.979	**0.789**
Sadness	MCB	0.000	0.000	-	0.000	0.000	-	0.000	0.000	-
	CLSA	0.472	0.262	0.337	0.500	0.059	0.106	0.704	0.589	0.642
	CPLSA	0.337	0.431	0.378	0.198	**0.491**	0.282	0.333	0.414	0.370
	CNMF	0.500	**0.453**	**0.475**	0.360	0.009	0.017	**0.708**	**0.821**	**0.760**
	DIM	**0.647**	0.157	0.253	**0.522**	0.249	**0.337**	0.408	0.169	0.240

Precision, Recall, and F-measure

Classification accuracy is usually measured in terms of precision, recall, and F-measure. Table 3 shows the values obtained by five approaches for the automatic classification of four emotions. The best results are marked in bold for each individual class. As can be seen from the table, the performances of each approach hinge on each dataset and emotion category, respectively.

In the case of SemEval, precision, recall and F-measure for CNMF and DIM are comparable. DIM approach gives the best result for *joy*, which has a relatively large number of sentences. In ISEAR, DIM generally outperforms other approaches except for some cases, whereas CNMF has the best recall score after the baseline for the *anger-disgust* category. Figure 2 shows the results of 3-dimensional and 2-dimensional data representation of the ISEAR corpus (other maps are similar, see (Kim and Calvo, 2010) for the one on USEs). Each point in the graph is a text unit. The labels in the graph represent the centroids of the pseudo-documents labeled as such, using equation 7. When the graph is seen in interactive mode we can see how each pseudo-document is placed in relation to other or to the labels.

When it comes to fairy tales, CNMF generally performs better than the other techniques. *Joy* also has the largest number of data instances in fairy tales and the best recall ignoring the baseline and F-measure are obtained with the approach based on DIM for this affect category. CNMF gets the best emotion detection performance for *anger-disgust*, *fear*, and *sadness* in terms of the F-measure.

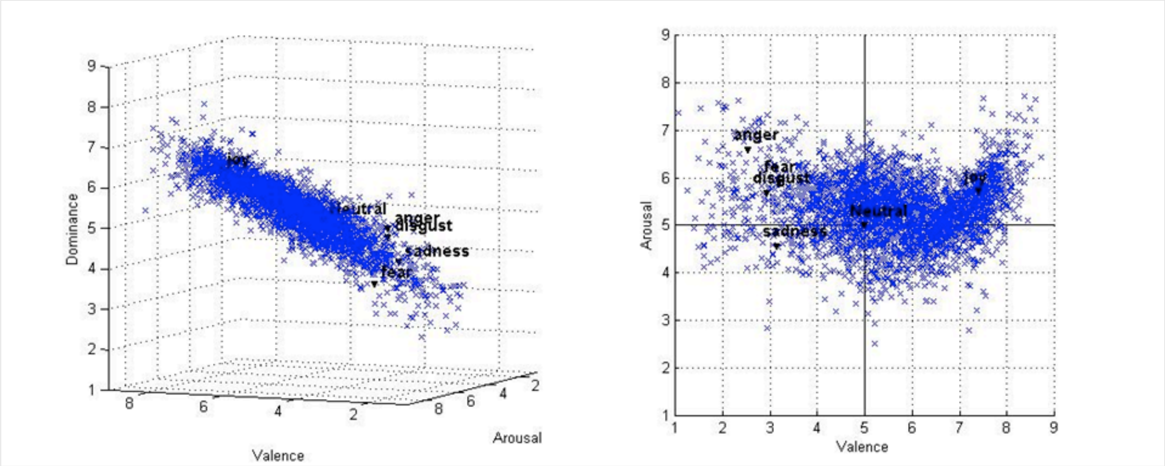

Figure 2: Distribution of ISEAR data in the 3-dimensional and 2-dimensional affective space. The 'x' denotes the location of one sentence corresponding to valence, arousal and dominance. Other maps are similar (c.f. Kim and Calvo, 2010).

Figure 3 and Table 4 display results among different approaches obtained on the different datasets. We describe the classification performance with the macro-averaged F1-measure, that weights every category equally, regardless of how many sentences are assigned to each category. This measurement prevents the results from being biased given the imbalanced data distribution. From this summarized information, we can see that CPLSA performs less effectively with several low performance results across all datasets. CNMF is superior to other methods in SemEval and Fairy Tales datasets, while DIM surpasses the others in ISEAR. Our CPLSA conducted except for ISEAR experiment is inferior to CNMF, DIM as well as CLSA. The result implies that statistical models which consider a probability distribution over the latent space do not always achieve sound performances. In addition, we can infer that models (CNMF and DIM) with only non-negative factors (as opposed to LSA that can have negative loadings) are appropriate for dealing with these text collections.

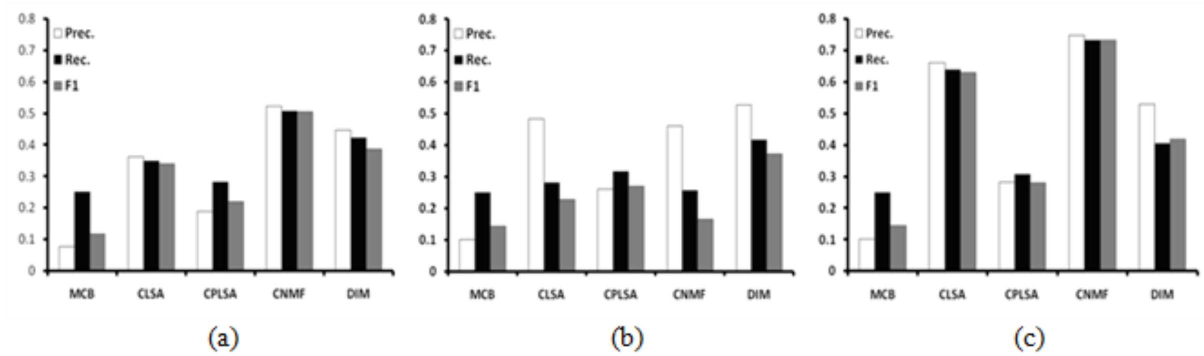

Figure 3: Precision, Recall and F-measure for: a) SemEval, b) ISEAR and c) Fairy Tales

Table 4: F1-measures for the different techniques on the different corpora

	Sem Eval F1	ISEAR F1	Fairy Tales F1	USE
MCB	0.118	0.143	0.145	0.221
CLSA	0.340	0.228	0.630	0.342
CPLSA	0.219	0.270	0.280	0.284
CNMF	0.505	0.166	0.733	0.307
DIM	0.386	0.372	0.419	0.363

Another notable result is that the precision, recall, and F-measure are generally higher in Fairy Tales than in the other datasets. These sentences in the fairy tales tend to have more emotional terms and the length of sentences is longer. The nature of fairy tales makes unsupervised models yield better performance (see Table 2). In summary, of those evaluated the categorical NMF model and the dimensional model show the best emotion identification accuracy as a whole.

It is interesting to compare these results to those in the literature. Danisman and Alpkocak (2008) used the ISEAR collection and vector space models to classify 801 news headlines from SemEval 2007. In their multiclass classification they found that using VSM produced better results (F1 = 0.322) than Naïve Bayes (F1=0.285), SVM (F1=0.286) and ConceptNet (F1=0.221). Regrettably their multiclass classification cannot be directly compared to ours because they separated Disgust and Anger producing 5 classes, rather than 4. Single class classifiers are better for comparison: their study resulted in VSM as the most accurate of all the techniques compared for Anger (F1=0.242), Joy (F1=0.496) and Sad (F1=0.371). SVM was the most accurate for Disgust (F1=0.095). In a different study using the Semeval 2007 data, Strapparava and Mihalcea (2008), showed that Naïve Bayes trained on blog data produced the most accurate results for Anger (F1=0.168) and Joy (F1=0.329) while LSA with synsets produced the best for Sadness (F1=0.231) and LSA single word produced the best for Disgust (F1=0.047) and Fear (F1=0.228). Table 3 shows we had better results for all three: Anger-Disgust (F1=0.278) was better than the results for class Anger or Disgust in either study, and Joy (F1=0.71) and Sadness (F1=0.475) also produced higher F1 scores.

6. DISCUSSION

We contribute a new computational modeling approach of emotions based on data collected using a dimensional model of emotions. The model follows the theory that emotions are better represented in a 3-dimensional space of valence, arousal dominance and this is substantially different to the categorical approach most commonly followed in the affective computing literature.

We show that the dimensional approach can be used as way of visualizing emotions in a psychologically meaningful space rather than a feature space driven by statistics. This might have many practical applications for new ways of searching for emotionally laden content. Furthermore, the dimensional model can also be used in the detection (i.e. classification) of emotion tasks

We compared the performances of three statistically driven dimensionality reduction techniques in the *categorical* representation of emotions with a *dimensional* representation based on psychologically supported data. Both types of representations are based on the naive bag-of-word assumption used in much of the literature, yet they provide good accuracies in the classification tasks. The results show that the NMF-based categorical classification performs best among categorical approaches to classification, and the dimensional approach is similar to NMF.

We have compared the above techniques in four datasets and conclude that the results do not generalize well because the results vary among datasets. This is due to the limitations of the lexical approach and of using background knowledge that affect the accuracy, particularly when the vocabulary of the background knowledge is not close to that of the corpora being modeled (i.e. domain oriented similarity (Strapparava and Mihalcea, 2008)). Future work will aim to investigate further this connection, identifying more effective strategies applicable to generic datasets. We are also developing tools for collecting more representative collections of affective words and possibly relevant n-grams by using a folksonomies approach.

Limitations of the lexical approach

The common bag-of-words assumption used here and in most of the literature is naïve in the sense that the affective meaning is not simply expressed by the lexicon used as the model assumes, it is also an effect of the linguistic structure. For example, we can observe the limitations of this approach in the input sentences below:

"The cook was frightened when he heard the order, and said to Cat-skin, You must have let a hair fall into the soup; if it be so, you will have a ***good*** beating." – which expresses *fear*.

"When therefore she came to the castle gate she saw him, and ***cried*** aloud for joy." – which is the expression for *joy*.

"Gretel was not idle; she ran screaming to her master, and ***cried***: You have invited a fine guest!" – which is the expression for *angry-disgusted*.

It would be interesting to study techniques that combine Natural Language Processing techniques with the use of normative databases.

Affective perspectives: evoke, express or emote

There are different perspectives in which emotions in text can be analyzed (e.g. the writer and the reader), and in a way most of the current AC approaches do not distinguish amongst them. Text can evoke or trigger emotions in those who read it. Text can also reflect or express the emotional state (or its socially acceptable simile) of the person writing it. These are two different functions of emotional text as the one we have discussed here.

It can also be argued that neither of these two perspectives directly reflects the true emotion of a person (either writer or reader). Emotions are often seen as internal states better described by its neural substrates or the subject's physiology or a combination of all the above (c.f. Calvo and D'Mello, 2010).

In future application-driven research it would be important to discriminate these three perspectives more explicitly.

ACKNOWLEDGMENTS

The authors would like to acknowledge Alessandro Valitutti for his contribution to earlier stages of this project. SMK was partially supported by a Norman I. Price Scholarship, University of Sydney.

REFERENCES

AKKAYA, C., J. WIEBE and R. MIHALCEA. Year. Subjectivity Word Sense Disambiguation. *In* Proceedings Conference Subjectivity Word Sense Disambiguation, pp. 190-199.

ALM, C. O. 2009. Affect in Text and Speech. Saarbrücken, VDM Verlag Dr. Müller.

BAEZA-YATES, R. and B. NETO. 1999. Modern Information Retrieval, ACM Press / Addison-Wesley.

BARRETT, L. F. 2006. Are Emotions Natural Kinds? Perspectives on Psychological Science, 1:28-58.

BELLEGARDA, J. 2010. Emotion Analysis Using Latent Affective Folding and Embedding. In Workshop on Computational Approaches to Analysis and Generation of Emotion in Text. NAACL-HLT 2010. Edited by D. Inkpen and C. Strapparava. NAACL Los Angeles, pp. 1-9.

BRADLEY, M. M. and P. J. LANG. 1994. Measuring Emotion: The Self-Assessment Manikin and the Semantic Differential. Journal of behavior therapy and experimental psychiatry, 25:49-59.

BRADLEY, M. M. and P. J. LANG. 1999. Affective Norms for English Words (Anew): Instruction Manual and Affective Ratings. University of Florida The Center for Research in, Technical.

BRECK, E., Y. CHOI and C. CARDIE. 2007. Identifying Expressions of Opinion in Context. In 20th International joint Conference on Artificial Intelligence. Edited by M. Veloso. Morgan Kaufmann Publishers Inc. Hyderabad, India, pp. 2683-2688.

CALVO, R. A. and S. K. D'MELLO. 2010. Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications. IEEE Transactions on Affective Computing, 1:18-37.

CHUNG, C. and J. PENNEBAKER. 2007. The Psychological Functions of Function Words. *In* Social Communication. *Edited by* K. Fielder. Psychology Press New York, pp. 343-359.

COHN, M. A., M. R. MEHL and J. W. PENNEBAKER. 2004. Linguistic Markers of Psychological Change Surrounding September 11, 2001. Psychological Science, 15:687-693.

D'MELLO, S., S. D. CRAIG, A. WITHERSPOON, B. MCDANIEL and A. GRAESSER. 2008. Automatic Detection of Learner's Affect from Conversational Cues. User Modeling and User-Adapted Interaction, 18:45-80.

DARWIN, C. 1872. The Expression of the Emotions in Man and Animals. London, John Murray.

DAVIDSON, R. J., K. R. SCHERER and H. H. GOLDSMITH. 2003. Handbook of Affective Sciences. New York, Oxford University Press, USA.

FRANCISCO, V. and R. HERVÁS. 2007. Emotag: Automated Mark up of Affective Information in Texts. In Doctoral Consortium at the 8th EUROLAN summer school. Edited, pp. 5–12.

GILL, A., R. FRENCH, D. GERGLE and J. OBERLANDER. Year. Identifying Emotional Characteristics from Short Blog Texts. *In* Proceedings Conference Identifying Emotional Characteristics from Short Blog Texts, pp. 2237-2242.

GUPTA, N., M. GILBERT and G. D. FABBRIZIO. 2010. Emotion Detection in Email Customer Care. In Workshop on Computational Approaches to Analysis and Generation of Emotion in Text. NAACL-HLT 2010. Edited by D. Inkpen and C. Strapparava. NAACL Los Angeles.

HANCOCK, J., C. LANDRIGAN and C. SILVER. 2007. Expressing Emotion in Text-Based Communication. In SIGCHI conference on Human Factors in Computing Systems. Edited. ACM Press San Jose, California, USA, pp. 929-932.

HOFMANN, T. 1999. Probabilistic Latent Semantic Indexing. In 22nd annual international ACM SIGIR conference on Research and development in information retrieval. Edited. ACM Berkeley, California, United States, pp. 50-57.

KAHN, J., R. TOBIN, A. MASSEY and J. ANDERSON. 2007. Measuring Emotional Expression with the Linguistic Inquiry and Word Count. American Journal of Psychology, 120:263-286.

KENNEDY, A. and D. INKPEN. 2006. Sentiment Classification of Movie Reviews Using Contextual Valence Shifters. Computational Intelligence, 22:110-125.

KENSINGER, E. and S. CORKIN. 2004. Two Routes to Emotional Memory: Distinct Neural Processes for Valence and Arousal. Proceedings of the National Academy of Sciences of the United States of America, 101:3310.

KIM, S., A. VALITUTTI and R. A. CALVO. 2010. Evaluation of Unsupervised Emotion Models to Textual Affect Recognition. In Workshop on Computational Approaches to Analysis and Generation of Emotion in Text. NAACL-HLT 2010. Edited by D. Inkpen and C. Strapparava. NAACL Los Angeles, pp. 62-70.

KIM, S. M. and R. A. CALVO. 2010. Sentiment Analysis in Student Experiences of Learning. In Third International Conference on Educational Data Mining (EDM2010). Edited by A. Merceron, P. Pavlik and R. Baker Pittsburgh, USA.

KORT, B., R. REILLY and R. W. PICARD. Year. An Affective Model of Interplay between Emotions and Learning: Reengineering Educational Pedagogy-Building a Learning Companion. *In* Proceedings Conference An Affective Model of Interplay between Emotions and Learning: Reengineering Educational Pedagogy-Building a Learning Companion, pp. 43-46.

LANDAUER, T. K., D. S. MCNAMARA, S. DENNIS and W. KINTSCH. 2007. Handbook of Latent Semantic Analysis, Laurence Erlbaum and Associates.

LANG, P. J., M. K. GREENWALD, M. M. BRADLEY and A. O. HAMM. 1993. Looking at Pictures: Evaluative, Facial, Visceral, and Behavioral Reactions. Psychophysiology, 30:261-273.

LEE, D. D. and H. S. SEUNG. 1999. Learning the Parts of Objects by Non-Negative Matrix Factorization. Nature, 401:788-791.

LIU, H., H. LIEBERMAN and T. SELKER. Year. A Model of Textual Affect Sensing Using Real-World Knowledge. *In* Proceedings Conference A Model of Textual Affect Sensing Using Real-World Knowledge, pp. 125-132.

LUND, K. and C. BURGESS. 1996. Producing High-Dimensional Semantic Spaces from Lexical Co-Occurrence. Behavior Research Methods Instruments & Computers, 28:203-208.

MILLER, G., R. BECKWITH, C. FELLBAUM, D. GROSS and K. MILLER. 1990. Introduction to Wordnet: An on-Line Lexical Database. Journal of Lexicography, 3:235-244.

OSGOOD, C. E., W. H. MAY and M. S. MIRON. 1975. Cross-Cultural Universals of Affective Meaning. Urbana, University of Illinois Press.

PANG, B. and L. LEE. 2008. Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval, 2:1-135.

PARKINSON, B. 1995. Ideas and Realities of Emotion. London, Routledge.

PENNEBAKER, J., M. FRANCIS and R. BOOTH. 2001. Linguistic Inquiry and Word Count (Liwc): A Computerized Text Analysis Program. Mahwah NJ, Erlbaum Publishers.

PENNEBAKER, J., M. MEHL and K. NIEDERHOFFER. 2003. Psychological Aspects of Natural Language Use: Our Words, Our Selves. Annual Review of Psychology, 54:547-577.

PENUMATSA, P., M. VENTURA, A. C. GRAESSER, D. R. FRANCESCHETTI, M. LOUWERSE, X. HU and Z. CAI. 2006. The Right Threshold Value: What Is the Right Threshold of Cosine Measure When Using Latent Semantic Analysis for Evaluating Student Answers. International Journal of Artificial Intelligence Tools, 12:257-279.

RUBIN, V., J. STANTON and E. LIDDY. 2004. Discerning Emotions in Texts. In AAAI-EAAT, 2004. Edited. AAAI.

RUSSELL, J. A. 2003. Core Affect and the Psychological Construction of Emotion. Psychological Review, 110:145-172.

SCHERER, K. R. and H. G. WALLBOTT. 1994. Evidence for Universality and Cultural Variation of Differential Emotion Response Patterning. Journal of Personality and Social Psychology, 66:310-328.

SHAIKH, M., H. PRENDINGER and M. ISHIZUKA. 2008. Sentiment Assessment of Text by Analyzing Linguistic Features and Contextual Valence Assignment. Applied Artificial Intelligence, 22:558-601.

STEVENSON, R. A., J. A. MIKELS and T. W. JAMES. 2007. Characterization of the Affective Norms for English Words by Discrete Emotional Categories. Behavior Research Methods, 39:1020-1024.

STRAPPARAVA, C. and R. MIHALCEA. 2007. Semeval-2007 Task 14: Affective Text. In the 4th International Workshop on Semantic Evaluations. Edited. Association for Computational Linguistics Prague, Czech Republic, pp. 70-74.

STRAPPARAVA, C. and R. MIHALCEA. 2008. Learning to Identify Emotions in Text. In the 2008 ACM symposium on Applied computing. Edited. ACM Fortaleza, Ceara, Brazil, pp. 1556-1560.

STRAPPARAVA, C. and A. VALITUTTI. 2004. Wordnet-Affect: An Affective Extension of Wordnet. In Proceedings of LREC. Edited Lisbon, pp. 1083-1086.

VALITUTTI, A., C. STRAPPARAVA and O. STOCK. 2005. Lexical Resources and Semantic Similarity for Affective Evaluative Expressions Generation. In Affective Computing and Intelligent Interaction. Edited, pp. 474-481.

WEINTRAUB, W. 1989. Verbal Behavior in Everyday Life. New York, Springer.

ZAKHAROV, K., A. MITROVIC and L. JOHNSTON. 2008. Towards Emotionally-Intelligent Pedagogical Agents. Intelligent Tutoring Systems, 5091:19-28.