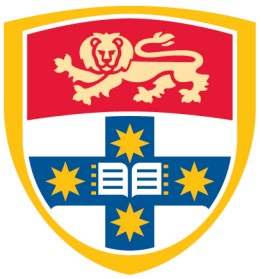
Recognising Emotions and Sentiments in Text



THE UNIVERSITY OF SYDNEY

This thesis is submitted in fulfilment of the requirements for the degree of Master of Philosophy in the School of Electrical and Information Engineering at The University of Sydney

Sunghwan Mac Kim

April 2011

Abstract

The increasing amount of textual information means that it is paramount to find ways of extracting valuable information from it. In many genres of writing emotions and sentiments are most important (e.g. product reviews or personal evaluations). But the nature of emotional phenomena in text (and other media) is very complex and can be interpreted in different ways and be represented by different computational models. Psychologists and affective computing researchers often use a categorical model in which text data are associated with emotional labels or a dimensional model where data is represented with coordinates in a two or three dimensional model Our study focuses on a new approach of combining these two representations using normative databases. These databases were developed by showing a stimuls (a word) to people and asking them to annotate them with sets of normative emotional ratings using the Self-Assessment Manikin (SAM). In this study, 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 emotion synset and input 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 Factorisation (NMF). In the dimensional model a normative database is used to produce three-dimensional vectors (valence, arousal, dominance) for each emotion synset and input document. The 3-dimensional model can be used to generate psychologically driven visualisations. Both models can be used for affect detection using distances amongst categories and input documents. To evaluate the performance measure we compare three methods based on a categorical model with a dimensional model-based method by using precision, recall, F-measures, and Cohen's Kappa. Experiments show that the categorical model using NMF and the dimensional model tend to perform best. I also perform a detailed analysis of the results including comparison with other systems and extraction of frequently occurring words.

Acknowledgements

First of all, I would like to thank my supervisor, Associate Professor Rafael A. Calvo for his direction, encouragement, and support with respect to this thesis as well as my research. I never would have achieved the fruits of my study without him.

Many thanks to my fellow Learning and Affect Technologies Engineering (LATTE) group members: Ming Liu, Stephen O'Rourke, Omar Alzoubi, Payam Aghaeiour, Sazzad Md Hussain, and Vilaythong Southavilay (Toto) for all the discussions and time.

I am grateful to Dr. Alessandro Valitutti, who is a co-author of my first paper for his valuable feedback and comments.

I also appreciate the invaluable advice and information about postgraduate life given by Dr. Young Choon Lee.

I would like to express my gratitude and appreciation to my parents for their support and love.

I thank my in-laws, especially Ewha Shin, for their assistance and trust.

Lastly, I am grateful to my wife Jhina and to my lovely daughters Yoobin and Davynn more than to anyone else for their endless sacrifice and patience during my study years.

Preface

The outcomes of my project include the following three peer-reviewed publications and CD:

- S.M. Kim and R.A. Calvo. Sentiment Analysis in Student Experiences of Learning. The 3rd International Conference on Educational Data Mining, pp. 111-120, Pittsburgh, USA, June 2010.

- S.M. Kim, A. Valitutti, and R.A. Calvo. Evaluation of Unsupervised Emotion Models to Textual Affect Recognition. Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, Association for Computational Linguistics, pp. 62-70, Los Angeles, California, June 2010.

- R.A. Calvo and S.M. Kim. Emotions in Text: Dimensional and Categorical Models. Accepted in Computational Intelligence.

- The CD contains all codes and datasets used in my experiments.

Table of Contents

1 Introduction **1**
1.1 Motivation 1
1.2 Contributions 3
1.3 Thesis Structure 4

2 Computational Models of Emotion **6**
2.1 Categorical Emotion Models 6
2.2 Dimensional Emotion Models 8
2.3 Two Emotion Models and Mappings 12
2.4 Summary 15

3 Emotions in Text **16**
3.1 Machine Learning Approaches 19
3.1.1 Supervised Learning 20
3.1.2 Unsupervised Learning 22
3.2 Text Mining Approaches 24
3.2.1 Corpus-based Approach 24
3.2.2 Thesaurus (Knowledge)-based Approach 26
3.2.3 Fusion Approach 27
3.3 Other Related Research 28
3.4 Summary 29

4 Methods **33**
4.1 Categorical Emotion Classification 34
4.1.1 WordNet-Affect 36
4.1.2 Vector Space Model 38
4.1.3 Dimension Reduction Methods 47

4.2 Dimensional Emotion Estimation 50
4.2.1 ANEW 51
4.2.2 Three-Dimensional Estimation 54
4.3 Summary 55

5 Evaluation **57**
5.1 Methodologies 58
5.1.1 Precision, Recall and F-Measure 58
5.1.2 Cohen's Kappa 61
5.2 Emotion-Labeled Data 62
5.2.1 SemEval: News headlines 63
5.2.2 ISEAR 63
5.2.3 Fairy Tales 64
5.2.4 Unit of Study Evaluation (USE) 65
5.3 Evaluation of Unsupervised Emotion Models to Textual Affect Recognition 68
5.3.1 Comparison with other systems 73
5.3.2 Frequently occurring words 75
5.4 Sentiment Analysis in Student Experiences of Learning 77
5.5 Summary 81

6 Conclusion **82**

A Detailed Results **84**

B TMG: A Matlab Toolbox for Generating Term-Document Matrices from Text Collections **90**

C EmotionML **94**

D SAM and Feeltrace **96**

E A Categorical Annotation Scheme for Emotions **99**

Bibliography **110**

List of Figures

Figure 2.1: Multidimensional scaling of Russell's circumplex model of emotion (Russell, 1980) 9
Figure 2.2: Thayer's two-dimensional model of emotion (Thayer, 1989) 9
Figure 2.3: Plutchik's emotion wheel (Plutchik, 1980) 10
Figure 2.4: Kort's Affective Model of Interplay Between Emotions and Learning (Kort, et al., 1990) 11
Figure 2.5: Hevner's adjective circle (Hevner, 1936) 12
Figure 2.6: Mapping six basic emotions onto Russell's circumplex model 15
Figure 4.1: System flow 34
Figure 4.2: Emotional hierarchy 37
Figure 4.3: The construction of a Term-by-Sentence Matrix (The 11 × 4 term-by-sentence matrix where the element $a_{i,j}$ is the number of times term i appears in sentence j) 41
Figure 4.4: Representation of sentences in a 3-dimensional vector space 42
Figure 4.5: The representation of a term, a sentence, and a synset on VSM 43
Figure 4.6: Cosine similarity between documents. $\text{sim}(d_1, d_2) = \cos\theta$ 46
Figure 4.7: Two-dimensional affective map of ANEW terms (Bradley & Lang, 1999) .. 52
Figure 4.8: Distribution of emotions and Fairy Tales sentences in the sentiment space .. 56
Figure 5.1: 2-by-2 confusion matrix 58
Figure 5.2: The Venn diagram representation of precision and recall 59
Figure 5.3: High precision and high recall 60
Figure 5.4: Unit of Study Evaluation 66
Figure 5.5: Distribution of the ISEAR dataset in the 3-dimensional and 2-dimensional sentiment space. The blue 'x' denotes the location of one sentence corresponding to valence, arousal, and dominance 71
Figure 5.6: Comparisons of Precision, Recall, and F-measure: (a) SemEval; (b) ISEAR; (c) Fairy tales 71
Figure 5.7: Comparisons of Mean Kappa: (a) SemEval; (b) ISEAR; (c) Fairy tales 72
Figure 5.8: Distribution of the USEs dataset in the 3-dimensional (left) and 2-dimensional (right) sentiment space. The 'x' denotes the location of one comment corresponding to valence, arousal, and dominance. 78
Figure B.1: Processing steps in TMG 93
Figure C.1: Emotion-labelled sentence and emotional state annotation tags 95
Figure D.1: The SAM tool on three emotion dimensions. 97
Figure D.2: The Feeltrace tool on two emotion dimensions. 98
Figure E.1: The Project Management pane 105
Figure E.2: Annotation Window 105
Figure E.3: Annotation Window (The choice of emotion type) 107
Figure E.4: Annotation Window (The choice of intensity type) 108
Figure E.5: Annotation Window (The choice of confidence type) 108
Figure E.6: Annotation Examples (based on EmotionML 1.0) 109

List of Tables

Table 3.1: Sentiment Analysis research 31
Table 3.2: Emotion Detection research 32
Table 4.1: A-Labels and examples 36
Table 4.2: Stative and Causative terms 38
Table 4.3: Formulas for local/global term weights and document normalisation 44
Table 4.4: Example ANEW terms from All Subjects 52
Table 4.5: Example ANEW terms from Male Subjects 53
Table 4.6: Example ANEW terms from Female Subjects 53
Table 5.1: Interpretation of Kappa values (Landis & Koch, 1977) 62
Table 5.2: Number of sentences for each emotion 64
Table 5.3: Sample sentences labelled with sadness/sad from the datasets 64
Table 5.4: Number of comments and sample comments for each sentiment 67
Table 5.5: Emotion identification results 69
Table 5.6: Overall average results 72
Table 5.7: Comparison results with other systems 74
Table 5.8: Most frequent 10 words from fairy tales 76
Table 5.9: Sentiment identification results 78
Table 5.10: Sample feedbacks from misclassified results. (Positive values are those rates 4 as 5, neutral as 3 and negative 1 or 2) 79
Table 5.11: Overall average results 80
Table A.1: Detailed results from SemEval 2007 84
Table A.2: Detailed results from ISEAR 85
Table A.3: Detailed results from Fairy tales 87
Table A.4: Detailed results from USE 89
Table B.1: Parameters and values used in TMG toolbox for pre-processing 91
Table E.1: Five emotions used in annotation 102

List of Abbreviations

ANEW	Affective Norms for English Words
AP	Adjacency Pair
DA	Dialog Acts
DAL	Dictionary of Affect of Language
EKD	Emotional Keyword Dictionary
EM	Expectation Maximisation
GEMEP	Geneva Multimodal Emotion Portrayal
GI	General Inquirer
HCI	Human-Computer Interaction
IESs	Intelligent Educational Systems
ISEAR	International Survey on Emotion Antecedents and Reactions
ITSs	Intelligent Tutoring Systems
ITSPOKE	Intelligent Tutoring SPOKEn dialogue system
KBANN	Knowledge-Based Artificial Neural Network
LSA	Latent Semantic Analysis
MLE	Maximum Likelihood Estimation
MPQA	Multi-perspective Question Answering
NLP	Natural Language Processing
NMF	Non-negative Matrix Factorisation
OMCS	Open Mind Common Sense
PAD	Pleasure-Arousal-Dominance
PLSA	Probabilistic Latent Semantic Analysis
PMI	Pointwise Mutual Information
SAM	Self-Assessment Manikin
SVD	Singular Value Decomposition
SVM	Support Vector Machine
TDM	Term-by-Document Matrix
TMG	Text to Matrix Generator
TSM	Term-by-Sentence Matrix
UI	User Interface
USE	Unit of Study Evaluations
VAD	Valence-Arousal-Dominance
VSM	Vector Space Model
WSJ	Wall Street Journal

CHAPTER 1
Introduction

"*Emotion has four aspects. (i) Cognition: a situation a must be perceived, related to past experiences, and evaluated .... (ii) Expression: Emotion is expressed outwardly in the form of somatic and autonomic activities .... (iii) Experience... the 'inward aspect of emotion'.., psychologists once divided emotion into two categories, those accompanied by pleasant affect and those which are unpleasant .... (iv) Excitement...when we experience certain emotions we look and feel excited.*"

- Theodore C. Ruch, 1962 (Kleinginna & Kleinginna, 1981).

1.1 Motivation

The interrelation of text and emotions has been a captivating topic for centuries. What makes people *feel* what they read? How is the writer's emotion conveyed in a text? How can we write to communicate an emotional message more clearly? A number of researchers have attempted to obtain answers to these questions for a long time and there is an enormous amount of literature on techniques and devices for emotion detection (Bloom, Garg, & Argamon, 2007; Hancock, Landrigan, & Silver, 2007; Zhang, Barnden, Hendley, & Wallington, 2006).

Computers have enabled researchers to study large amounts of text in a systematic way, and concrete applications that explore the emotional meaning of text are now possible, for example, a search engine uses affect information from song lyrics to find appropriate music corresponding to users' mood (Cho & Lee, 2006). However, automatically understanding the affective meaning of the same text is still a challenging research domain.

The importance of text as a form of communication has increased with the use of computers. A great deal of today's social interaction happens online and in text form (e.g.

email and social networks). This has increased the need to improve the human computer interface and the way computers contribute to people's text-based communications. Successful social interaction means successful affective communication, which makes use of users' emotions and plays a vital role in effective social interaction (Ma, Prendinger, & Ishizuka, 2005; Neviarouskaya, Prendinger, & Ishizuka, 2007; Parlade et al., 2009).

Researchers have recognised the significance of affect with respect to Human-Computer Interaction (HCI) (Cowie et al., 2001), and the affective states of users can be recognized using intelligent User Interface (UI) (H. Liu, Lieberman, & Selker, 2003a). In order to support individuals in their writing or reading activities, or groups communicating, researchers have tried to develop techniques for sensing users' affective states. Many researchers have tried to detect the emotional states of users through various sensor channels on UI such as facial expressions, speech, and text. Among them, text is a particularly important modality for identifying emotions because in this global era most human knowledge is transmitted via text especially over the Internet. Hence, studying the relationship between natural language and affective information, and dealing with its computational treatment is becoming an important field. However, previous research does not cover the wide variety of affective phenomena in natural language, and many interesting and challenging aspects still need to be investigated. Another challenge is that there is not a single accepted theory of emotion and different approaches stress different aspects and deal with different data and different situations. The background chapter of this thesis and the several literature review articles on affect sensing published recently (Calvo & D'Mello, to appear; B. Pang & L. Lee, 2008) provide a complete review of this literature.

Two attempts to measure emotions are based on two different models: dimensional and categorical. In the categorical model emotions are labelled. We say that a person is "*happy*" or "*sad*" and people get a sense of what we mean. In the dimensional model the representation is based on a set of quantitative measures using multi-dimensional scaling (e.g. "pleasant-unpleasant", "excitement", and "yielding-resisting"). The first model is closer to every-day understanding but carries many theoretical assumptions that contradict actual data, as will be discussed later. Computational models

must incorporate these psychological theories. So far most research has focused on categorical models of emotion, and not on systematic comparisons of the two models.

In the affective computing domain, supervised learning techniques are preferred due to strong performance. However, a challenge to using supervised techniques is the need for corpora with text that has been annotated with emotion labels. These are time-consuming and expensive to produce. Unsupervised techniques do not have these requirements but are often less precise.

This thesis investigates unsupervised computational approaches to affect detection in text. In addition, two computational models of emotion are compared and this leads to new insights that make emotion models computationally feasible and scalable to various text genres and collections.

1.2 Contributions

The research in this thesis contributes to an important ongoing topic, namely emotion modelling applied to affect detection. Various techniques for representing emotions are evaluated in this study. Both categorical and dimensional models of emotion have been applied to affect classification of text. This work represents the first systematic evaluation of a technique to combine these two emotion models under consistent conditions and evaluation methodologies.

Another key contribution of the research in this thesis is the practical application of two emotion models using unsupervised and thesaurus-based approaches. This allows the valid exploration of existing lexical resources used for affect detection in text. This requires a clear understanding of how both text and emotions are represented in a computer in order to classify text by emotion. Such accurate understanding is essential in the design of textual affect detection methods. The relationship between emotion and text is of significance when mapping various textual data to an emotion space. The classification systems perform more efficiently and precisely by extracting and analysing the relation information.

There are different perspectives from which emotions in text can be analysed. Text can evoke or trigger emotions in those who read it and text can also reflect or express the emotional state of the person writing it. These are two different functions of emotional text but we have discussed only one here. Our approaches do not distinguish amongst them in this thesis.

The crucial objective of this research is to review the wide range of applications of these emotion models. The implemented emotion models are gained to be feasible to apply to various kinds of datasets and the performance results are compared with respect to the datasets. In addition, our best approaches significantly outperform the existing works of the "Affective Text" task in SemEval 2007 which focuses on classification of emotions and valences in text (C. Strapparava & R. Mihalcea, 2007).

1.3 Thesis Structure

The literature supporting the theoretical models compared in this thesis is reviewed in Chapters 2 and 3. Chapter 2 presents the research underlying the two emotion models used to represent the affective states. It starts with an introduction to the categorical emotion model in Section 1. Section 2 details the dimensional emotion model used in the experiment and describes how individual models can be mapped onto each other in Section 3.

Chapter 3 describes machine learning and text mining techniques used in affect detection. This chapter also discusses the supervised and unsupervised learning techniques most commonly used. It presents the corpus-based and thesaurus-based approaches frequently used in the literature, before coming to a discussion of approaches that combine these two. Section 3 gives a brief survey on research related to our work.

Chapter 4 presents in more detail the specific techniques that we have developed. The techniques to detect emotions in text are explored as an aspect of text classification problems utilizing lexical resources. More specifically, it describes the role of emotion models and lexical resources in affect classification. Section 1 describes the categorical emotion classification which derives from a categorical emotion model and WordNet-

Affect. Section 2 presents a dimensional emotion estimation based on a dimensional emotion model and Affective Norms for English Words (ANEW).

Chapter 5 presents our results for the evaluation of unsupervised emotion models in terms of textual affect recognition. It explores the feasibility of sentiment analysis in students' responses to Unit of Study Evaluations (USE). Section 1 reviews the methodologies that are relevant to the evaluation. In Section 2 we go over the affective datasets that are used in the experiment. Section 3 and Section 4 provide the results of the evaluation and a detailed analysis of the outputs from experiments. Finally, the works presented in the sections are concluded and discussed, respectively.

Chapter 6 concludes with a summary of the work presented in the preceding chapters. It provides insight into future improvements and ideas towards the direction of future work.

CHAPTER 2

Computational Models of Emotion

"*Emotion refers to the process whereby an elicitor is appraised automatically or in an extended fashion, an affect programme may or may not be set off, organized responses may occur, albeit more or less managed by attempts to control emotional behavior.*"

- Paul Ekman, 1977 (Kleinginna & Kleinginna, 1981).

There are two significantly different models for representing emotions: *the categorical model* and *the dimensional model*. Each type of model helps to convey a unique aspect of human emotion and both of them can provide insight into how emotions are represented and interpreted within the human mind. The categorical model and dimensional models have two different methods for estimating the actual emotional states of a person. In the former, a subject is usually required to choose one emotion out of a set of emotions that best represents the feeling conveyed. The latter exploits rating scales for each dimension by using tools like the Self Assessment Manikin (SAM) (Lang, 1980) or Feeltrace (Cowie, Douglas-Cowie, Savvidou, et al., 2000). SAM consists of pictures of manikins, to estimate the static degree of a dimension at a fixed moment. In contrast, Feeltrace is able to track emotional information continuously over time. A detailed explanation of the two tools can be found in Appendix D.

2.1 Categorical Emotion Models

The most straightforward way of identifying emotions is the use of emotion-denoting words, or category labels. The categorical model either assumes that there are discrete

emotional categories such as six basic emotion categories namely *anger*, *disgust*, *fear*, *joy*, *sadness* and *surprise* (Ekman, 1992), or uses domain-specific expressive categories. There are both primary and unrelated emotions in the model. Each emotion is characterised by a specific set of features, expressing eliciting conditions or responses. Most work in affective computing has focused on the six basic emotions. However, many researchers have argued that different sets of emotions are required for different domains, for instance, in the field of teaching and education. As a specific example D'Mello, Picard and Graesser (2007) proposed five categories (*boredom*, *confusion*, *delight*, *flow*, and *frustration*) for describing affect states in the student-system dialogue. Learners rarely feel fear or disgust, whereas they typically experience boredom or delight which is an argument for the need for domain-specific categories. The advantage of categorical representation is that it represents human emotions intuitively with easy to understand emotion labels.

However, there are several shortcomings with the categorical model of emotions due to the limited number of labels. For example, the emotional categories consist of discrete elements, and a great variety of emotions within each discrete category can be frequently observed. The categories do not cover all emotions adequately because numerous emotions are grouped together under one category. Furthermore, the same affective states can be expressed by means of different emotional categories owing to cultural, environmental, linguistic or personality differences, which leads to poor agreement among emotional categories. These findings indicate that emotional categories may not represent distinct affective states although the set of emotion categories is defined. In addition, this problematic conceptualisation may result in non-optimal or inefficient affect-detection. First, it can lead to a forced-choice identification problem, which is that subjects are likely to discriminate among presented categories rather than to identify an emotion label themselves. This can force the subjects to choose an irrelevant category. The second problem is more serious and related to the first one. It is occasionally not possible for subjects to select an appropriate category since it does not exist in the label set. Therefore, a categorical model has the limitations of an identification task in attempting to identify the precise emotional states perceived by people. For instance, subjects cannot help selecting one of six basic emotions (e.g. *anger*,

disgust, *fear*, *joy*, *sadness*, and *surprise*) even though they feel *neutral* and want to choose that category.

Nevertheless, the categorical model has been dominant and there are many variations of the model due to its simplicity and familiarity. The only way that the categorical models might differ is in terms of how many categories they list.

2.2 Dimensional Emotion Models

A second approach to identifying emotions is a dimensional model, which represents affects in a dimensional form. Emotional states in this model are related to each other by a common set of dimensions and are generally defined in a two or three dimensional space. Each emotion occupies a location in this space.

A variety of dimensional models has been studied and we will review each model in brief. Russell's model of affect is introduced (Russell, 1979, 1980; Russell, Lewicka, & Niit, 1989) as a reference circumplex through a figure with a setting of points representing the emotions (Figure 2.1). Emotion-related terms are organised in a circumplex shape which enables a subject to choose a position anywhere between two discrete emotion-related terms. Numerical data are obtained from the relative position of the points in the two-dimensional bipolar space (*valence-arousal*). The valence dimension indicates *positive* and *negative* emotions on different ends of the scale. The arousal dimension differentiates *excited* vs. *calm* states. Scherer's affect model also appears as a reference circumplex for the binary classification experiments presented in Généreux and Evans (2006). The proximity of two emotion categories in the circumplex represents conceptual similarity of the two categories. Mehrabian's model is based on a three-dimensional PAD (*Pleasure-Arousal-Dominance*) representation (Mehrabian, 1996). In this model, the dominance dimension is used to distinguish whether the subject feels in control of the situation or not. The pleasure dimension corresponds to the valence of Russell's model.

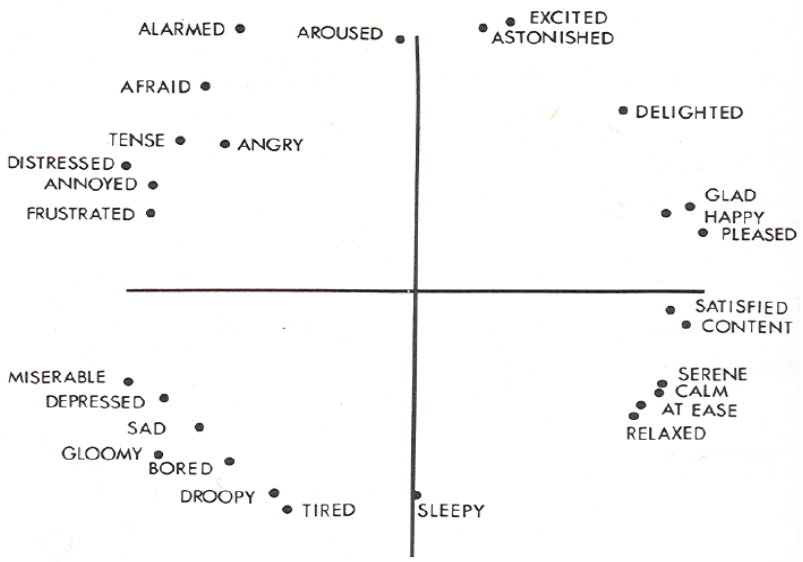


Figure 2.1: Multidimensional scaling of Russell's circumplex model of emotion (Russell, 1980)

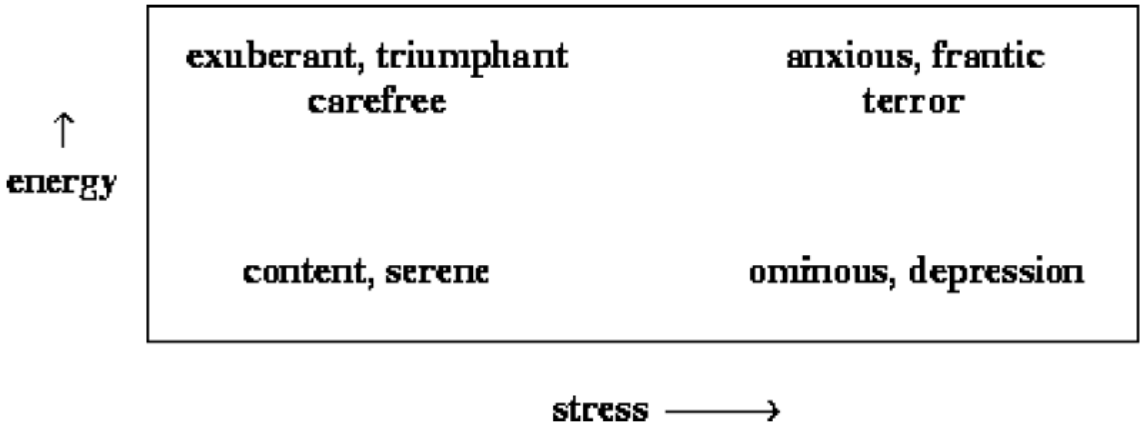


Figure 2.2: Thayer's two-dimensional model of emotion (Thayer, 1989)

Thayer (1989) utilises the two dimensions of *energy* and *stress* shown in Figure 2.2. In the energy-stress model, *contentment* is located in low energy/low stress, *depression* in low energy/high stress, *exuberance* in high energy/low stress, and *anxious/frantic* in high energy, high stress, respectively. Whissell (1989) and Plutchik (1980) have both presented an affect model located on an activation-evaluation space. However, each model has a completely different approach to represent the location of an emotion. Whissell proposed two numerical values to show how emotions can be related to *activation* and *evaluation*. On the other hand, Plutchik used angles on the emotion

circle. The model is called the "*emotion wheel*", which is shown in Figure 2.3. Kort and colleagues (1990) have proposed a model relating phases of learning to emotions in a valence-arousal plane. Figure 2.4 is used in a fully automated computer program designed to recognise a learner's emotion. Finally, the basic two and three-dimensional models have also been expanded to circular models, such as Russell's circumplex model (Russell, 1980) and Hevner's adjective circle (Hevner, 1936). In these approaches, a list of adjectives, 28 terms for Russell, and 67 terms for Hevner, are mapped to their respective quadrant. In particular, Hevner studies the affective values of six features derived from music and their relations with emotion. Figure 2.5 shows that these features are mapped to a circular model of affect encompassing eight different emotional categories.

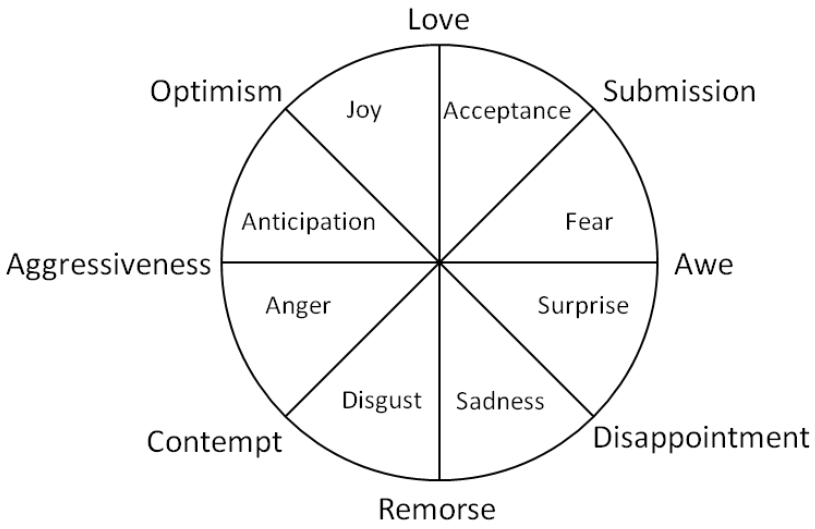


Figure 2.3: Plutchik's emotion wheel (Plutchik, 1980)

The description of emotional states by means of emotion dimensions has some advantages. A major benefit of dimensional models is that they are not correlated to a certain emotional state (e.g. *angry* or *happy*). Two or three dimensions of emotional meaning are commonly identified by means of rating. Due to their gradual nature, emotion dimensions are able to capture subtle emotion concepts that differ only slightly

in comparison with broad emotion categories. Emotion dimensions can represent very specific identification and a large range of people's emotion concepts. In particular, a dimensional description is well-suited for the task of measuring the full defined emotional states.

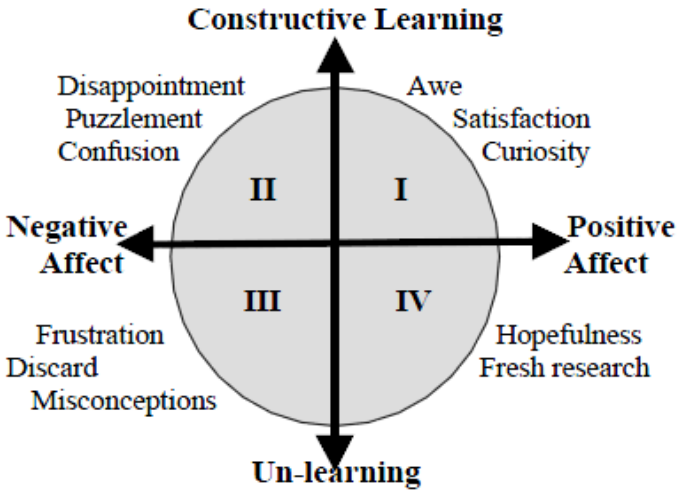


Figure 2.4: Kort's Affective Model of Interplay Between Emotions and Learning (Kort, et al., 1990)

In addition, emotional states are related to each other on a dimensional space, which is a significantly different approach from the categorical model. A dimensional model provides a means for measuring the degree of similarity between emotion categories. Adjacent categories on the space are very similar, while opposite categories are distinctly different from each other. In summary, a dimensional emotion model is a useful representation capturing all relevant emotions and provides a means for measuring similarity between affective states.

Figure 2.5: Hevner's adjective circle (Hevner, 1936)

2.3 Two Emotion Models and Mappings

The question of which of the models is more suitable for representing and measuring emotions has been studied. One experiment showed that both the categorical and the dimensional models are valuable to explore emotion data by Ritz et al. (2005). However, they found that it is difficult to differentiate between two positive emotions by means of 2-dimensions (*arousal* and *valence*). This leads to either the necessity of adding another dimension to the dimensional model or supplementary considering a categorical model to discriminate between two particular emotions. Francisco and Gervás (2006) attempted to present a system for the automated mark up of affective information in texts by using Emotag. Emotag is an approach that uses two representations of emotions: both a categorical model and a dimensional model. While considering emotions and learning, Kort, Reilly and Picard (1990) proposed (but provided no empirical evidence for) a model that combines two emotion models, placing categories in a valence-arousal plane. This mixed approach has also been used in other domains such as blog posts where Aman and Szpakowicz (2007) studied how to identify emotion categories as well as emotion intensity. Barrett (1998) suggested that the applicability of one of two models might differ individually based on *valence* focus and *arousal* focus. A categorical model is

appropriate for capturing the affective states in the lower valence and higher arousal focus. By contrast, a dimensional model is better when emotions are high in valence focus and low in arousal focus.

In addition, judging the suitability of emotion representations should take into account modalities such as speech, video, and text. In general, the dimensional representation is utilised to detect emotional content in the case of speech. On the other hand, video components tend to use the categorical model with intensity information for studying facial expressions. A categorical model is also adopted for most emotion recognition in text. The popularity of a categorical model is shown in Liu et al. (2005) and Nigam and Hurst (2004) with respect to textual affect recognition.

An accurate understanding of how emotions are represented both in the human mind and in the computer is essential in the study of affect detection. The relationship between emotion and text is also important when mapping various textual information to an emotion space. In general, the study of emotions in written text is conducted from two opposite points of view. The first is the viewpoint of a writer. This is concerned with how emotions influence a writer of a text in choosing certain words and/or other linguistic elements. The second point of view is concerned with how a reader interprets the emotion in a text, and what linguistic clues are used to infer the emotion of the writer. In this thesis, the second point of view is taken because we are interested in the way people infer emotions. In the remainder of this chapter two emotion models will be reviewed with respect to their relevance to this project.

It may be necessary or should be possible to map between different emotion representations. For example, a multi-modal generation system requires a mapping mechanism so its components using different representations work together (Krenn et al., 2002). These different emotion representations are not independent and they have just been created for different purposes with different capture methodologies. Emotion categories can be located on emotion dimensions by means of a rating test (Cowie et al., 1999). In that case mapping from a categorical model to a dimensional model is a simple task if the coordinates of the emotion category have been calculated. However, it is not possible to convert data from one emotion representation to another at all times. The

mappings from dimensions to categories are not fully possible because a dimensional model can only capture the most essential aspects of an emotion concept and provide an underspecified description of an emotional state. For instance, the coordinates for *anger*, *fear*, and *disgust* may be very close because the three categories share the same valence/arousal properties (see Figure 2.6). The features distinguishing the three categories cannot be represented through a dimensional model. For this reason, the corresponding region on the space can only be mapped to a coarse or generic category like *anger-fear-disgust* rather than a specific category, which is not very rigorous (a lossy mapping). Adding one dimension like dominance improves the classification result, in particular, regarding neighbouring emotions. In our experiment, we make use of the three-dimensional space of affective states in order to classify these vague emotions into a specific category more accurately as aforementioned. In summary, mappings between the currently existing emotion representations are imperfect. The mapping of emotional words is particularly inaccurate in the case of dealing with text since the self-interpretation of an emotional word like "love" can be different from its interpretation in the context when mapping it between two models. Thus, a full description of the emotion encompassing all context-related information is ultimately required to clarify the emotion mappings. Figure 2.6 illustrates the emotional wheel with the verbal semantic mapping of basic emotions.

The mapping was attempted in the NECA project (Krenn, 2003), in which an affective reasoning component is used for determining the appropriate emotion in a given dialogue situation, represented as a combination of emotion category and intensity. This representation is mapped onto emotion dimensions, using the intensity value to linearly interpolate between the neutral state and the coordinates of the fully developed emotional state in the dimensional space.

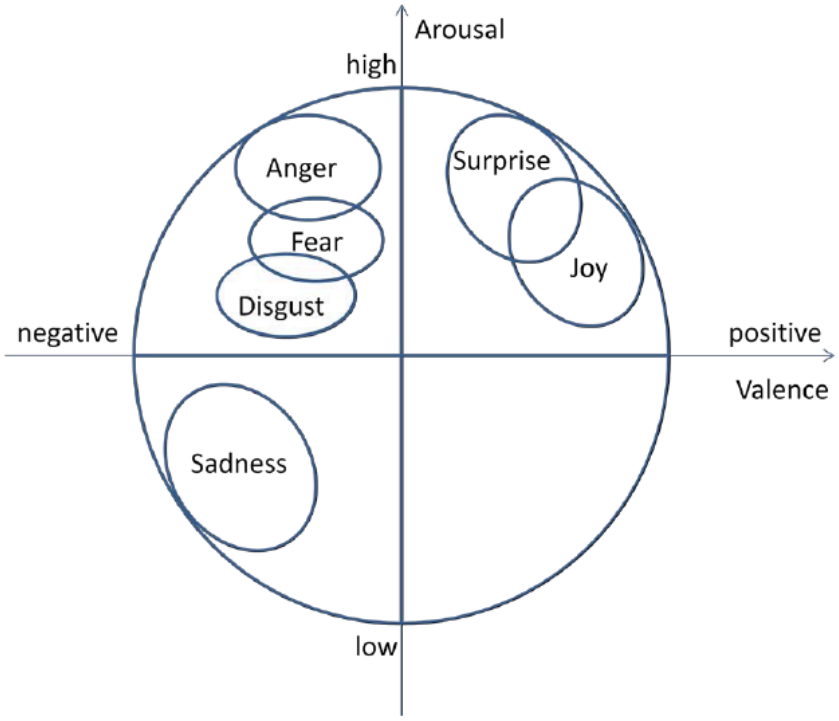


Figure 2.6: Mapping six basic emotions onto Russell's circumplex model

2.4 Summary

Two emotion models and the related academic literature are presented: a categorical model representing emotions as distinct categories and a dimensional model representing essential emotional properties in terms of dimensions. Each type of model helps to convey a unique aspect of human emotion and the models can provide insight into how emotions are represented and interpreted within human mind. Categorical emotion models use the straightforward and intuitive description of emotions such as *happy* or *sad*. Dimensional emotion models are particularly well suited for representing emotion concepts. The performance of two emotion models is evaluated and compared in our work.

CHAPTER 3

Emotions in Text

"Emotion. A complex subjective, psychological process, which may be induced by environmental stimuli and mediated by physiological variables; it may have the power to motivate an organism to action. It is a felt tendency toward stimuli appraised as good, and away from those appraised as bad."

- Philip G. Zimbardo, 1980 (Kleinginna & Kleinginna, 1981).

The aim of textual affect detection is to understand how people express emotions through text, or how text triggers different emotions (Osgood, May, & Miron, 1975). Affect detection in text can identify expressions of emotion such as *happiness*, *sadness*, *anger*, etc. Emotions in text are subjective expressions that describe the feelings of people about entities, events, and their properties. In most cases, emotions are hidden behind the text and it is difficult for a reader to extract relevant sentences with emotions. Emotion or affect identification is normally used in the field of cognitive science (Ortony, Clore, & Collins, 1988) and there are some connections to *affective computing*, where the goals include enabling computers to recognize and express emotions (Picard, 1997).

Most research in textual sentiment or emotion sensing has targeted documents, which is commonly known as document-level affect detection (see Tables 3.1 and 3.2). This approach considers the whole document as the basic information unit. Affect detection at the document-level assumes that the document expresses a single emotion and the same emotion holds throughout the document. A problem of document-level occurs when multiple emotions are expressed in the same document. A single category or a set of dimension values cannot be used to describe different emotions in a document. On the other hand, genres such as fiction and blogs benefit from a finer-grained level of

analysis since there is often a dynamic progression of emotions in narrative texts (Cecilia Ovesdotter Alm, Roth, & Sproat, 2005; Aman & Szpakowicz, 2007). To increase sentiment or emotion detection the analysis is applied to individual sentences, and is called sentence-level affect detection (see Tables 3.1 and 3.2). The assumption of sentence-level affect detection is that the sentence expresses a single emotion from a single emotion holder. This assumption makes only simple sentences with a single emotion appropriate for this type of analysis. For this reason, sentence-level analysis is not suitable for a compound sentence, which may express more than one emotion by providing both *anger* and *joy*, or a mixed emotion.

As mentioned earlier, affect detection is related to cognitive science and affective computing. Affective computing researchers have been trying to develop computational systems that can recognise and respond to the emotional states of the user (Calvo & D'Mello, to appear). In general, the channels such as behaviour, physiology and language are used to investigate communication between the highly emotional human and the emotionally challenged computer. Behaviour and physiology have contributed to much of the work in affect detection using bodily sensors that monitor facial expressions, gross body language, acoustic-prosodic vocal features, and physiological measures such as heart rate monitors, electromyography, skin conductance, etc (Anttonen & Surakka, 2005; Chuang & Wu, 2004; Grings & Dawson, 1978; Lichtenstein, Oehme, Kupschick, & Jürgensohn, 2008; Rainville, Bechara, Naqvi, & Damasio, 2006). On the other hand language is a less frequently explored channel. Researching languages in affective computing refers to finding the relationship between emotions and texts, and it particularly addresses the task of sensing emotions in human-computer natural language dialogue (S. K. D'Mello, Craig, Witherspoon, McDaniel, & Graesser, 2008). The ultimate goal of affective computing is enhancing the quality of the interaction with a computer to make a computer interface more usable, enjoyable, and effective by means of automatically recognizing and responding to a user's affective states during the interactions. In addition, affective computing in the field of education is a crucial component of Intelligent Tutoring Systems (ITSs) (S. D'Mello, et al., 2007). The ITSs with animated pedagogical agents have been designed to recognize, assess, and react to a learner's cognitive and affective states. Affective sensitive ITSs help learners

comprehend explanations by interacting with them in natural language and this is considered to have a positive impact on learning (S. D'Mello & Graesser, 2007). Therefore, the requirement to detect affect in the dialogue between ITSs and learners grows out of this need in the education domain.

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 the text (Abbasi, Chen, & Salem, 2008; Akkaya, Wiebe, & Mihalcea, 2009; Breck, Choi, & Cardie, 2007; H. Liu, Lieberman, & Selker, 2003b; B. Pang & L. Lee, 2008; Shaikh, Prendinger, & Ishizuka, 2008; Whitelaw, Garg, & Argamon, 2005a). 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". This approach is called sentiment analysis because it focuses on the 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 or opinion analysis is a popular research area in the computational linguistics community and is extensively discussed in a recent review (B. Pang & L. Lee, 2008). In one of the earliest Natural Language Processing (NLP) studies in sentiment analysis, Hatzivassiloglou and McKeown (1997) separate adjectives into positive or negative groups by using conjunctions on the semantic orientation of the combined adjectives. Many other different fields that make use of automatic sentiment analysis include recognising sentiments in on-line text (Boiy, 2007; S.M. Kim & Hovy, 2006), extracting investor sentiment from stock message boards (Das & Chen, 2001), distinguishing editorials from news articles (J. Wiebe, Wilson, Bruce, Bell, & Martin, 2004; Yu & Hatzivassiloglou, 2003), analysing public comments for positive and negative responses (Popescu & Etzioni, 2005), determining support and opposition in congressional debates (Thomas, Pang, & Lee, 2006). In the past few years, it has been possible for more and more people to make their opinions available and to share their experiences with others due to the advent of the information era. On the Web particularly sentiment analysis has proved to be very useful applied to product and movie reviews (Pang & Lee, 2004; Turney, 2002). This topic is dealt with in more detail in Chapter 5.

Sentiment analysis also helps universities or educational institutions to analyse students' sentiments with respect to education policies, education services or education issues. Opinions found within comments, feedback or critiques provide useful indicators for many different purposes. Therefore, an automated opinion detection system is very useful. For instance, questionnaires such as the Unit of Study Evaluation (USE) can be used in sentiment analysis in order to analyse students' satisfaction toward the courses (Sunghwan Mac Kim & Calvo, 2010). These USEs contain textual and quantitative responses about a student's subjective experience of learning. Sentiment analysis enables an interpretation of the reasoning behind the evaluation. Consequently, academics or institutions are able to provide students with a better education by incorporating the results of sentiment analysis.

In this respect, an affect detection task is considered as a text classification problem using categories to represent an emotion. It is also a challenging natural language processing or text mining problem. Due to its tremendous value for practical applications, there has been an explosive growth of both research in academia and applications in industry.

3.1 Machine Learning Approaches

Supervised and unsupervised approaches have been used to automatically recognise expressions of emotion in text such as *happiness*, *sadness*, *anger*, etc. The following sections give a basic background for supervised and unsupervised machine learning. In supervised learning, an algorithm is provided with a label for every example, and this information is used to learn a mapping from examples to labels. In unsupervised learning, no labels are provided at all in advance and consequently no training is provided. Semi-supervised learning falling between supervised and unsupervised learning is mentioned in brief. The semi-supervised learning approach first trains a model on labelled data, then labels some unlabelled data by that model. The model is iteratively trained with new and larger data until no more unlabelled examples exist. A bootstrapping approach is proposed to classify the polarity of adjectives and their target nouns in blogs (Suzuki, Takamura, & Okumura, 2006). Sindhwani and Melville (2008) use a semi-supervised

sentiment prediction algorithm that exploits lexical knowledge in conjunction with unlabelled examples.

Supervised learning 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 for emotion-related work in the realm of NLP (Morinaga, Yamanishi, Tateishi, & Fukushima, 2002; Strapparava, Valitutti, & Stock, 2007). We take advantage of the unsupervised learning technique in the experiments of this thesis.

The goal of machine learning is to learn the following simple function

$$\mathcal{F}: \mathcal{X} \rightarrow \mathcal{Y}$$

where $\mathcal{X}$ is a set of examples $\{x_1 \ldots x_n\}$. Each example may be associated with a label from the universe of possible labels $\mathcal{Y}$ with pairs <*x*, *y*>. When the set of labels $\mathcal{Y}$ is discrete and finite, the labels are called target classes. Each example x_i has one or more properties, which are called features. These features describe the properties of the examples, and can be used in learning as predictors of the target class. The features related to our experiments are explained in Chapter 4.

3.1.1 Supervised Learning

Supervised learning is a technique in which an algorithm uses predictor and target attribute value pairs to learn the predictor and target value relation. In other words, the goal of supervised learning is reasoning from externally supplied instances (a training dataset) to make predictions about future instances (a test dataset). The training data consist of pairs of predictor and target values. Each predictor value is tagged with a target value. If the algorithm can predict a categorical value for a target attribute, it is called a classification function. In the case of text categorization, supervised learning can be applied when a set of predefined topic categories, such as “business, sports, entertainment”, are provided as well as a set of documents labelled with those categories. The resulting classifier is used to assign class labels to the testing instances. Various

supervised machine learning classification techniques have been applied to automated affect detection, such as Naïve Bayesian, Support Vector Machines (SVM), Neural Networks, etc. These techniques have been exploited to classify movie reviews into two classes, *positive* and *negative* (Dave, Lawrence, & Pennock, 2003; Li, Bontcheva, & Cunningham, 2007; Pang, Lee, & Vaithyanathan, 2002; J. M. Wiebe, Bruce, & O'Hara, 1999). In addition, supervised and unsupervised techniques have been compared before. Strapparava and Mihalcea (2008) describe the comparison between a supervised (Naïve Bayes) and an unsupervised (Latent Semantic Analysis - LSA) method for recognising six basic emotions.

A corpus (plural corpora) or dataset is a large and structured set of texts. A process known as annotation is required to make naturally unlabelled corpora labelled for the training of supervised learning models. A set of documents labelled with predefined categories is the outcome of annotation. The level of annotation mostly corresponds to the level of analysis. In general, learning from annotated data (supervised learning) leads to better results than learning from raw data (unsupervised learning). Unfortunately, the supervised classification learning method has several drawbacks for sentiment analysis. First, supervised learning needs a training dataset, which is a fundamental and essential resource for the purpose of building smart classifiers. However, annotation (making a training dataset) is a tedious and time-consuming task. Moreover, identifying sentiments or emotions in text is difficult, error prone and a highly subjective task. Therefore, adding a good annotation scheme is required for building an objective training set (one annotation scheme is give in Appendix E). Second, supervised learning techniques treat sentiment analysis as a standard classification problem and thereby feature selection is of significance in the learning process. Feature selection is critical to achieving good performance in text categorisation problems. However, it is not easy to identify good features since the process is usually language dependent. In addition, feature selection is task dependent; different features are required for different classification tasks. Furthermore, there are a number of possible features to consider in sentiment analysis and general feature approaches do not always succeed well. Finally, supervised learning techniques do not tend to consider the deeper structure of the natural language, which is an important characteristic in terms of sentiment analysis in text.

3.1.2 Unsupervised Learning

The above techniques require enough labelled data in order to train the classifiers. However, this is not always available in reality. In other words, there is, initially, very limited training data available when building new classification systems. The unsupervised learning technique is capable of overcoming this problem by grouping data into related sets. Unsupervised learning is a technique in which the algorithm uses only the predictor attribute values. There are no target attribute values and the learning task is to gain some understanding of relevant structure patterns in the data. Examples of unsupervised learning are as follows: finding clusters (e.g. k-means), dimensionality reduction (e.g. LSA), building topographic maps (e.g. elastic networks), finding the hidden causes or sources of the data, and modeling the data density (Kanayama, Tetsuya, & Hideo, 2004; S.-M. Kim & Hovy, 2004; Nasukawa & Yi, 2003; J. Wiebe, et al., 2004).

In particular, it is not easy to interpret affective texts since judging dominant emotional words and phrases can be highly subjective for affect detection. Thus, using unsupervised learning would be quite reasonable and natural. However, unsupervised affect detection has to be able to draw satisfactory inferences from the test data without access to labelled data or prior linguistic information regarding the emotion indicators. For these reasons, a number of unsupervised learning approaches take advantage of predefined emotion lexicons, which are sets of affect indicative words or phrases, and then determine the degree of emotion of a text unit via these lexicons. There are various methods to estimate similarity between text units and terms within the lexicons such as cosine angle. Therefore, the selection of the lexicons is a crucial component in applying this type of technique. In this thesis, WordNet-Affect and ANEW are utilised as the lexical repositories.

In addition, unsupervised learning does not assume domain-specific knowledge. This indicates that an unsupervised technique depends on general cues and can ignore potentially strong emotion evidence drawn from the topic-specific documents. Thus, it exhibits quite different performances for different domains with regard to related affect classification problems (Turney, 2002). For this reason, D'Mello et al. (2008) specifically suggested a specialised affect classifier for detecting utterance types and emotions in

students' dialogue within Autotuor. They particularly used LSA in order to improve Intelligent Educational Systems (IESs).

The unsupervised technique used by Turney (2002) had three steps. It performs classification based on some fixed syntactic phrases that are likely to be used to express emotions. In the first step, it extracts phrases containing adjectives or adverbs. The reason for doing this is that research has shown that adjectives and adverbs are good indicators of subjectivity and emotions. However, although an isolated adjective may indicate subjectivity, there may be an insufficient context to determine its emotion orientation. Therefore, the algorithm extracts two consecutive words, where the first member of the pair is an adjective/adverb and the second provides context. In the next step, the orientation of the extracted phrases is estimated using the *Pointwise Mutual Information* (PMI) measure. A PMI measure reflects the statistical association between two words. More specifically, PMI is the degree of statistical dependence when one of the words is observed in the presence of the other. The affect orientation of a phrase is computed based on its association with the positive reference word and its association with the negative reference word. Lastly, the algorithm calculates the average affect orientation of all phrases in the review. If the average is a positive value, the review shows an affirmative recommendation. Otherwise, it is a negative recommendation.

Popescu and Etzioni (2005) make use of an unsupervised classification technique called relaxation labelling (Hummel & Zucker, 1983) to recognize the contextual polarity of words that are at the heads of selected opinion phrases. They take an iterative approach, using relaxation labelling first to determine the contextual polarities of the words, then again to label the polarities of the words with respect to their targets. A third stage of relaxation labelling then is used to assign final polarities to the words, taking into consideration the presence of other polarity terms and negation. Popescu and Etzioni (2005) use features that represent conjunctions and dependency relations between polarity words. Apart from these, many other unsupervised methods exist (Efron, 2004; Fei, Liu, & Wu, 2004; Yi, Nasukawa, Bunescu, & Niblack, 2003).

3.2 Text Mining Approaches

Emotion recognition uses a variety of channels such as facial features, posture/gesture and conversational cues. To develop a system that can detect users' affective states by means of dialogue, we have to provide a suitable knowledge of affective concepts and lexicon. In addition, we should provide the association information between an input concept and the most appropriate emotional category. There are two types of data source that are taken into consideration for the purpose of sensing emotion in text. One is a corpus that is a large and structured set of texts and the other is a thesaurus (knowledge) that contains synonyms and antonyms. It is difficult to obtain well-designed emotional data and this fact motivates researchers to try various methodologies and protocols for the creation of an affective database. Banziger and Scherer (2007) describe a new corpus that is called The Geneva Multimodal Emotion Portrayal (GEMEP) Corpus. The GEMEP consists of more than 7,000 audio-video emotional portrayals, representing 18 emotions, portrayed by 10 different actors with the assistance of a professional theatre director. Moreover, Zara and colleagues (Zara, Maffiolo, Martin, & Devillers, 2007) have established EmoTaboo which is the name of a protocol designed to collect a corpus of multimodal expressions of emotion during human-human interaction. On the other hand, WordNet-Affect (Strapparava & Valitutti, 2004), which is the core of thesaurus-based approaches, is developed and extended from WordNet. WordNet-Affect is an affective lexical hierarchical resource and allows us to identify sentences containing emotional words corresponding to affective synsets. The following two approaches are categorised based on the above two types of data source: Corpus and Thesaurus.

3.2.1 Corpus-based approach

In the corpus-based approach, researchers take into account a variety of raw corpora pertinent to their experiments (Bo Pang & Lillian Lee, 2008). Emotion or polarity is decided by using co-occurrence in a corpus. Wilson et el. (2005) proposed a new approach to phrase-level sentiment analysis. In contrast, most of the work on sentiment analysis has been done at the document level. Wilson, Wiebe, and Hoffmann make use of

a two-step process that classifies each phrase as neutral or polar and then disambiguates all phrases marked as polar into *positive*, *negative*, *both*, or *neutral*. This two-step classifier is named Neutral-Polar Classification, which utilises the Multi-perspective Question Answering (MPQA) Opinion Corpus. 15,991 subjective expressions from 425 documents (8,984 sentences) are annotated in the corpus. The classifier is developed by using the BoosTexter AdaBoost.HM (Schapire & Singer, 2000) machine learning algorithm.

New approaches are demonstrated in order to sense textual affect in online communication such as AOL Instant Messenger, and MSN Messenger. In Holzman and Pottenger (2003), textual chat messages are automatically converted into speech phonemes since people can often detect the emotions of others easily by sensing the tone of their voice rather than by reading written text. The primary advantage is that this approach is robust for noise in chat data due to the features of speech phonemes. To be specific, the characteristics of many online messages contain misspellings, do not adhere to grammar rules, and do not stick to complete words. The corpus data for this method are obtained from two sources. The first is a set of conversations between one of the authors and another individual over a 2-month period. The second is another set of dialogues held by two separate individuals. In addition, k-nearest neighbour instance based learning (IBk) is employed for this application and is compared with four off-the-shelf machine learning methods (Naïve Bayes, One R, Decision Table, j48)

On the other hand, Litman and Riley (2004) make use of acoustic-prosodic features and lexical features for predicting student emotions in computer-human tutoring dialogues. The corpus consists of student conversations with ITSPOKE (Intelligent Tutoring SPOKEn dialogue system) (Litman & Silliman, 2004) and was collected from November 2003 to April 2004. While lexical features outperform acoustic-prosodic features, combining lexical and speech features does not improve performance as indicated in performing machine learning experiments (boosted decision trees).

Liu, Lieberman, and Selker (2003a) establish a powerful approach that makes use of a large real-world commonsense database. This approach tackles the problems and limitations that four categories (keyword spotting, lexical affinity, statistical natural

language processing, hand-crafted models) of existing methods garner. Open Mind Common Sense (OMCS) is selected as a real world corpus of 400,000 facts about the everyday world. Commonsense is represented by English sentences that are divided into 20 or so sentence patterns. Four commonsense affect models (Subject-Verb-Object-Object Model, Concept-Level Unigram Model, Concept-Level Valence Model, Modifier Unigram Model) are generated from OMCS and are combined for the purpose of classifying the affect of text. Moreover, four smoothing models like decay, interpolation, global mood, and meta-emotion are applied in order to make the transition of emotions from one sentence to the next fluid. In this application, the system is integrated into a client's email to provide feedback to the user. However, some emotions cannot be fully expressed due to limitations of pre-drawn facial expressions.

3.2.2 Thesaurus (Knowledge)-based approach

The thesaurus-based (or knowledge)-based approach uses synonyms or glosses of lexical resources in order to determine the emotion or polarity of words, sentences or documents. Yi, Nasukawa, Bunescu, and Niblack (2003) apply NLP techniques to sentiment analysis that is composed of feature term extraction, sentiment detection, and subject and sentiment association by relationship analysis. In terms of feature term extraction, two feature term selection algorithms are developed and tested based on a mixed language model and likelihood ratio. Sentiment analysis in the second stage uses two linguistic resources: the sentiment lexicon and the sentiment pattern database. The sentiment lexicon contains the sentiment definition of individual words from General Inquirer (GI), Dictionary of Affect of Language (DAL), and WordNet. The sentiment pattern database includes sentiment extraction patterns for sentence predicates in which sentiment verbs are collected from GI, DAL, and WordNet. Approximately 120 sentiment predicate patterns are currently stored in the database. The experiment result is compared with the collocation algorithm and the algorithm of ReviewSeer classifier (Dave, et al., 2003).

According to Neviarouskaya et al. (2007), the rule-based Affect Analysis Model was developed to handle correctly written text as well as an informal style of writing.

Since informal messages are particularly and frequently written in abbreviated or expressive manner in online communication, the peculiarity of this communication medium is taken into account. This model takes advantage of 1627 words including adjectives, nouns, verbs, and adverbs as the source of affective lexicon excerpted from WordNet-Affect. Moreover, 364 emoticons and 337 most popular acronyms and abbreviations are collected in order to support the handling of abbreviated language and the interpretation of affective features of emoticons, abbreviations, and words. Interjections and modifiers are also included in the affect database. The Affect Analysis Model consists of five sequential stages: symbolic cue analysis, syntactical structure analysis, word-level analysis, phrase-level analysis, and sentence-level analysis. Furthermore, an avatar is created in order to reflect the detected affective information in connection with text input.

3.2.3 Fusion approach

The fusion approach is a kind of hybrid method that makes use of both corpus- and thesaurus-based approaches to overcome the disadvantages of both. Seol and colleagues (2008) demonstrate the hybrid system that is decomposed into keyword-based and machine learning methods. If the input sentence has emotional keywords, the keyword-based approach is applied. In other cases, the system uses a machine learning method, Knowledge-Based Artificial Neural Network (KBANN) to infer emotions from sentences with no emotional keywords. The keyword-based approach is based on EKD (Emotional Keyword Dictionary) that consists of words that have emotional meaning. On the other hand, KBANN network uses 3,200 sentences in the emotion-tagged corpus that come from a script of drama, novel and public web diary. This emotion recognition system combines a keyword-based approach and a KBANN machine learning approach and evaluates eight emotions (*anger*, *fear*, *hope*, *sadness*, *happiness*, *love*, *thank*, *neutral*) by separate modules.

Somasundaran, Ruppenhofer and Wiebe's (2007) studies suggest that classifiers with lexical and discourse knowledge give the best performance in regard to detecting

opinions in multi-party conversations. Sentiment and arguing are annotation categories as well as two opinion types. For this work, 7 scenario-based team meetings (6504 sentences) from the AMI corpus (Carletta et al., 2006) where the participants had to design a new TV remote control are annotated. Furthermore, Somasundaran, Ruppenhofer and Wiebe take into account not only sentiment and arguing lexicons as knowledge sources, but also Dialog Acts (DA) and Adjacency Pair (AP) to capture the flow of discourse. The hypothesis that knowledge sources, DA, and AP are useful indicators of opinion expression in conversational data is verified by using SVM.

The experiments performed by Strapparava and Mihalcea (2008) describe the comparison between knowledge (thesaurus)-based and corpus-based methods for six basic emotions: *anger*, *disgust*, *fear*, *joy*, *sadness* and *surprise*. In particular, five different systems for emotion analysis are implemented by using two approaches. *WordNet-Affect Presence*, *LSA Single Word*, *LSA Emotion Synset* and *LSA All Emotion Words* are based on the thesaurus-based method and *Naïve Bayes Trained On Blogs* is devoted to the corpus-based method. As far as knowledge-based emotion annotation is concerned, WordNet-Affect is considered as an affective words database. In contrast, blog entries from LiveJournal.com are taken into account when it comes to corpus-based emotion annotation. Moreover, Latent Semantic Analysis (LSA) is implemented in regard to representing word sets and texts in a knowledge-based approach, while Naïve Bayes classifier is trained by means of the blogposts in the corpus-based approach. The evaluation results show that the WordNet-Affect Presence provides the best precision and the LSA All Emotion Words leads to the highest recall and F-measure.

3.3 Other Related Research

Some researchers focus on more specific tasks such as just looking at words (Hatzivassiloglou & McKeown, 1997) and subjective expressions (S.-M. Kim & Hovy, 2004; Wilson, et al., 2005). However, most researchers have been studying the assignment of sentiments or emotions to text. Various datasets have been used ranging from the Wall Street Journal (WSJ), new articles, movie reviews, blog posts, MPQA, customer feedback, children's fairy tales, etc. In addition, different techniques have been

applied to automatically capture the sentiments or emotions of text. These are largely based on NLP, machine learning algorithms, and unsupervised learning.

Tables 3.1 and 3.2 list some existing work in sentiment analysis and emotion detection respectively and show the different types of categories along with the associated emotion models, approaches, domains, techniques, etc. Table 3.1 presents studies related to sentiment analysis, most of which use categorical models and corpus-based approaches. In addition, we could infer from the table that sentiment analysis is useful in reviews. Most of emotion detection also uses categorical models but they are applied to various domains such as dialogues or fairy tales. Sentiment analysis tends to be used at the document-level, whereas emotion detection is generally performed at the sentence-level.

3.4 Summary

This chapter has given an overview of two approaches regarding machine learning: supervised learning and unsupervised learning. Supervised learning techniques are applied to emotion detection or sentiment analysis with training data. However, the emotion-related annotation, which is the task of making the training data, is highly subjective and error prone. For this reason, unsupervised-learning has a preference for capturing emotions or sentiments in text.

In addition to machine learning approaches, there are three text mining approaches that are correspondent with two types of data sources to sense emotions or polarities in text. Corpus-based approach makes use of raw corpus pertinent to detecting affective states. Emotion or sentiment detection is performed through using co-occurrence in a corpus. Thesaurus-based approach exploits lexical resources such as GI, DAL, WordNet and WordNet-Affect to detect emotions or sentiments from textual information. Fusion approach is a hybrid of corpus-based and thesaurus-based approaches. Fusion approach is used in order to strengthen the strength and make up for the weakness of each approach.

Papers	Categories	Emotion Model	Approaches	Domains	Level	Dataset	Features	Techniques
(J. M. Wiebe, et al., 1999)	Subjective/Objective	Categorical	Corpus-based	News	Sentence	Wall Stree Journal Treebank (1,004 sentences)	Syntactic, Stylistic	NB
(Riloff & Wiebe, 2003)	Subjective/Objective	Categorical	Corpus-based	News	Sentence	News from FBIS (37,947 sentences)	Semantic, Stylistic	NB
(Yu & Hatzivassiloglou, 2003)	Fact/Opinion(Positive,Negative,No Orientation,Mixed Orientation, Uncertain Orientation)/Uncertain	Categorical	Corpus-based	News	Sentence /Document	Wall Stree Journal (8,000 articles)	Syntactic, Semantic	NB, Similarity Score
(J. Wiebe, et al., 2004)	Subjective/Objective/Unsure	Categorical	Corpus-based	News	Sentence	Wall Stree Journal (1,289,006 words), Newsgroup (103,623 words)	Syntactic, Stylistic	KNN
(Chesley, Vincent, Xu, & Srihari, 2006)	Objective/Positive/Negative	Categorical	Fusion-based	News, discourse	Document	Web documents (1,152 documents)	Syntactic, Stylistic	SVM
(Ng, Dasgupta, & Arifin, 2006)	Positive/Negative	Categorical	Corpus-based	Reviews	Document	Reviews (2,000) & Non-reviews (2,000) from Movie domain	Syntactic	SVM
(Li, et al., 2007)	Opinionated/Non-opinionated	Categorical	Corpus-based	News	Sentence	MPQA (535 documents), NTCIR-6 English Corpus (439 documents)	Syntactic	SVM
(Pang, et al., 2002)	Positive/Negative	Categorical	Corpus-based	Reviews	Document	Movie Reviews from IMDb (2,053 reviews)	Syntactic	NB, ME, SVM
(Gamon, 2004)	Positive/Negative	Categorical	Corpus-based	Reviews	Document	Global Support Services survey (11,399 documents), Knowledge Base survey (29,485 documents)	Syntactic, Stylistic	SVM
(Mullen & Collier, 2004)	Positive/Negative	Categorical	Corpus-based	Reviews	Document	Movie Reviews from IMDb (1,380 reviews)	Syntactic, Semantic	SVM
(Pang & Lee, 2004)	Subjective/Objective	Categorical	Corpus-based	Reviews	Sentence /Document	Movie Reviews from a site (5,000 snippets), Movie Reviews from IMDb (5,000 sentences)	Syntactic, Semantic	SVM, NB
(Whitelaw, Garg, & Argamon, 2005b)	Positive/Negative	Categorical	Fusion-based	Reviews	Document	Movie Reviews from IMDb (2,000 reviews)	Syntactic, Semantic	SVM(SMO)
(Cui, Mittal, & Datar, 2006)	Positive/Negative	Categorical	Corpus-based	Reviews	Document	Electronic Products Review from some sites (321,434 reviews)	Syntactic, Stylistic	SVM(PA), LM, Winnow
(Abbasi, et al., 2008)	Objective/Subjective(Positive, Negative)	Categorical	Corpus-based	Reviews, discourse	Document	Movie Reviews from IMDb (2,000 reviews), Messages from US and Middle East Forum (1,000 for each)	Syntactic, Stylistic	SVM(SMO)

(Morinaga, et al., 2002)	Positive/Negative	Categorical	Rule-based	News	Sentence	Product fields from internet (2,319 records)	Syntactic	Similarity Score
(Turney, 2002)	Positive/Negative	Categorical	Rule-based	Reviews	Phrase	Reviews from Epinions (410 reviews)	Syntactic, Semantic	Semantic Orientation
(Agrawal, Rajagopalan, Srikant, & Xu, 2003)	Positive/Negative	Categorical	Corpus-based	Discourse	Document	3 topics from Usenet (13,642 postings for Abortion, 12,029 postings for Gun Control, 10,285 postings for Immigratioin)	Syntactic, Link-based	SVM, NB, Link analysis
(Dave, et al., 2003)	Positive/Negative	Categorical	Corpus-based	Reviews	Sentence	Product Reviews from C net (31,574 reviews), Product Reviews from Amazon (5,920 reviews)	Syntactic	SVM, NB, Similarity Score
(Nasukawa & Yi, 2003)	Positive/Negative	Categorical	Rule-based	Reviews	Sentence	Camera reviews from web pages (2,000 cases)	Syntactic, Semantic	Semantic Relationship
(Yi, et al., 2003)	Positive/Negative/Neutral	Categorical	Thesaurus-based	Reviews, News	Sentence	Digital camera reviews from some sites (2,323 reviews), Music reviews from Epinions (2,639 reviews)	Syntactic, Semantic	NLP
(Beineke, Hastie, & Vaithyanathan, 2004)	Positive/Negative	Categorical	Fusion-based	Reviews	Document	Movie Reviews from Pang (27,886 reviews)	Semantic	NB, Semantic Orientation
(Kanayama, et al., 2004)	Favorable/Unfavorable/Question/Request	Categorical	Rule-based	Reviews	Sentence	Reviews from bulletin boards (200 sentences)	Syntactic, Semantic	Similarity Score
(S.-M. Kim & Hovy, 2004)	Positive/Negative/Neutral	Categorical	Fusion-based	Discourse	Sentence	Web discourses from DUC 2001 corpus (100 sentences)	Semantic	Similarity Score
(Nigam & Hurst, 2004)	Positive/Negative	Categorical	Corpus-based	Discourse	Message /Sentence	Web discourses from online resources (34,000 messages)	Syntactic, Semantic	Winnow
(B. Liu, et al., 2005)	Positive/Negative	Categorical	Fusion-based	Reviews	Sentence	Reviews of 15 electronic products from web pages	Syntactic, Semantic	Similarity Score
(Wilson, et al., 2005)	Positive/Negative/Neutral/Both	Categorical	Fusion-based	News	Phrase	MPQA (8,984 sentences)	Syntactic, Semantic	BoosTexter AdaBoost.HM

Table 3.1: Sentiment Analysis research

(Strapparava & Mihalcea, 2008)	Anger/Disgust/Fear/Joy/Sadness/Surprise	Categorical	Fusion-based	News	Sentence	News headlines from newspapers (1,250 headlines)	Semantic	LSA, NB
(S. D'Mello & Graesser, 2007)	Boredom/Confusion/Delight/Flow/Frustration	Categorical	Corpus-based	Dialogue	Sentence	Dialogues from 4 people (4,298 sentences)	Semantic	NB, NN, KNN, DT, LR
(Seol, et al., 2008)	Anger/Fear/Hope/Sadness/Happiness/Love/Thank/Neutral	Categorical	Fusion-based	Discourse	Sentence	Scripts from web pages (3,200 sentences)	Semantic	NN
(Somasundaran, et al., 2007)	Sentiment/Arguing/Utterance Arguing	Categorical	Fusion-based	Dialogue	Sentence	AMI Corpus (6,504 sentences)	Syntactic, Stylistic	SVM
(H. Liu, et al., 2003a)	Happy/Sad/Angry/Fearful/Disgusted/Surprised	Categorical	Corpus-based	Fact	Sentence	OMCS Corpus (about half a million sentences)	Syntactic, Semantic, Stylistic	Linguistic Models
(Holzman & Pottenger, 2003)	Neutral/Angry/Sad/Afraid/Disgusted/Ironic/Happy/Surprise	Categorical	Corpus-based	Dialogue	Sentence	Conversation (1,201 messages)	Syntactic	NB, One R, D Table, Ib1, j48, Ib20
(Neviarouskaya, et al., 2007)	Anger/Disgust/Fear/Guilt/Interest/Joy/Sadness/Shame/Surprise, Intensity	Hybrid	Rule-based	Dialogue	Sentence	Blog post corpus (160 sentences)	Syntactic, Stylistic	Affect Analysis Model
(Cecilia Ovesdotter Alm, et al., 2005)	Angry/Disgust/Fearful/Happy/Sad/Positively Surprised/Negatively Surprised	Categorical	Fusion-based	Fairy tale	Sentence	Children stories (185 tales)	Syntactic, Semantic, Stylistic	Winnow
(Scott & Matwin, 1998)	Livestock/Gold, Corn/Wheat, Murder/Marriage, Political/Religion, Microbiology/Neuroscience	Categorical	Fusion-based	Discourse	Document	Reuters-21578 (537 texts), DigiTrad (856 texts), UseNet (529 texts)	Syntactic, Semantic	Ripper
(Danisman & Alpkocak, 2008)	Anger/Disgust/Fear/Joy/Sad	Hybrid	Fusion-based	News	Sentence	ISEAR (7,467 sentences), News (801 headlines)	Syntactic, Semantic	VSM, ConceptNet, NB, SVM
(Aman & Szpakowicz, 2007)	Happiness/Sadness/Anger/Disgust/Surprise/Fear/Mixed emotion/No emotion, Intensity	Hybrid	Fusion-based	Discourse	Sentence	Blog post corpus (5,205 sentences)	Semantic	NB, SVM
(Subasic & Huettner, 2001)	83 affect categories, Centrality, Intensity	Hybrid	Thesaurus-based	News, Reviews	Sentence /Document	News report (a train crash in London), Moview review (Matrix)	Syntactic	NLP, Fuzzy Logic
(Strapparava, et al., 2007)	Joy/Fear/Surprise/Anger/Sadness	Categorical	Thesaurus-based	News	Sentence	News titles from News sites (50 titles)	Semantic	LSA
(R. Mihalcea & Liu, 2006)	Happy/Sad	Categorical	Fusion-based	Discourse	Phrase	Blog post corpus from LiveJournal.com (10,000 blogposts)	Syntactic, Semantic	NB, SVM, Rocchio

Table 3.2: Emotion Detection research

CHAPTER 4
Methods

"*Emotion is a way of feeling and a way of acting. It may be defined as a tendency of an organism toward or away from an object, accompanied by notable body alterations. There is an element of motivation-an impulsion to action and an element of alertness, a hyperawareness or vividness of mental processes. There is of course the opposite, a depression of movement.*"

- A. R. Vonderahe, 1944 (Kleinginna & Kleinginna, 1981).

In our study emotions or affects are detected in text based on the aforementioned two models: the *categorical model* and the *dimensional model*. Categorical classification is the basis of the categorical model. Likewise, a dimensional model is the foundation of practical dimensional estimation. Categorical classification is carried out with features derived from WordNet-Affect, whereas dimensional estimation is performed with features from ANEW.

Both methods go through pre-processing steps: stopwords listing and stemming. These steps help the significant linguistic components of text to be focused and considered by removing unimportant features. Most languages are full of structural words that provide little meaning to text.

In the following, the explanation of each classification method is given in more detail.

4.1 Categorical Emotion Classification

Figure 4.1 gives a system flow overview of categorical classification. The system consists of the following four major components: TMG (Text to Matrix Generator), Dimension Reduction, WordNet-Affect, and Vector Space Model (VSM). TMG is a Matlab toolbox for text mining and the detailed explanation is given in Appendix B.

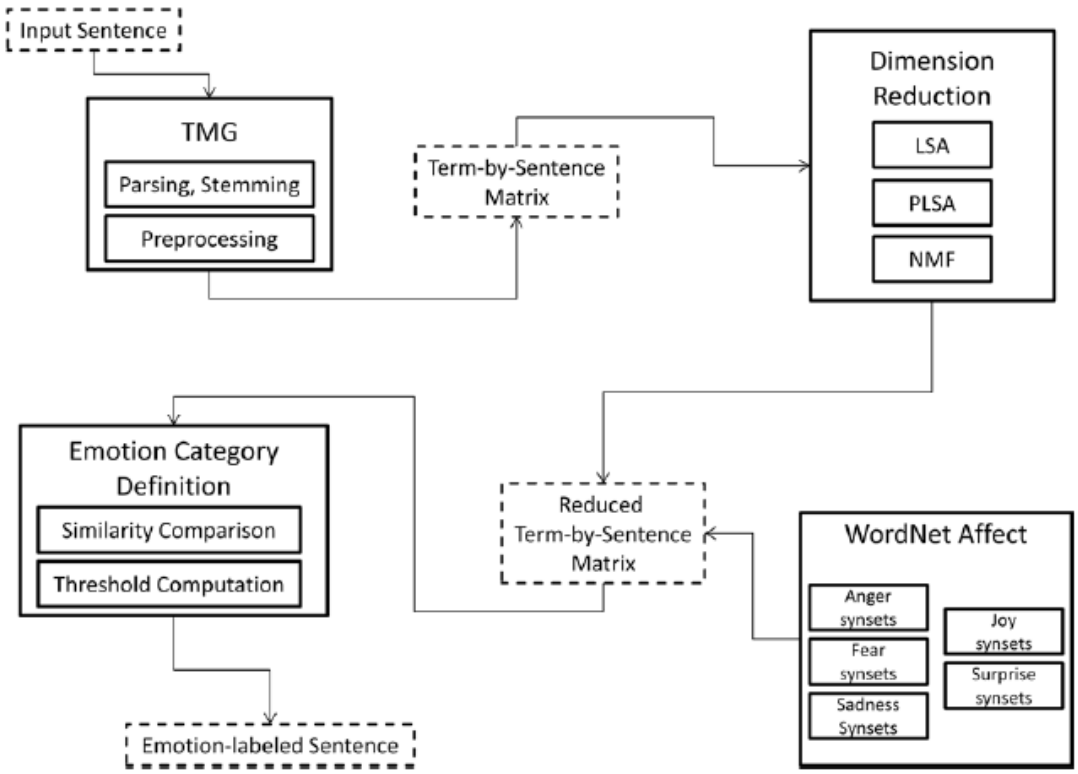


Figure 4.1: System flow

One of the most important and difficult characteristics in text classification problems is the high dimensionality of space. The original space consists of unique terms like words and the documents that contain the terms. This space can contain tens or hundreds of thousands of terms for even a moderate size text corpus. In addition, most of these dimensions are not related to text classification which means some dimensions can

have some noise and irrelevant data. These data have a negative effect on classification results. Another remarkable characteristic is the sparsity of space. These peculiarities are the causes of the increase in computation time. For these reasons, it is highly desirable to reduce the dimensionality of space without sacrificing the semantic meaning of text. Dimensionality reduction refers to the process of granting each word weights that correspond to its importance in the context. Dimension reduction is also able to enhance computational efficiency as well as classification precision. There are various kinds of automatic text dimensionality reduction techniques such as information gain, mutual information and chi-squared but we present the following three well-known methods: Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (PLSA), and Non-negative Matrix Factorisation (NMF).

WordNet-Affect (Strapparava & Valitutti, 2004) is an affective lexical repository of words referring to emotional states. WordNet-Affect extends WordNet by assigning a variety of affect labels to a subset of synsets representing affective concepts in WordNet (emotional synsets). In addition, WordNet-Affect has a hierarchy of affective domain labels. There are publicly available lists relevant to the six basic emotion categories extracted from WordNet-Affect and we used those lists of emotional words for our experiment.

In addition to WordNet-Affect, the Vector Space Model (VSM) is exploited in which terms and textual documents can be represented through a term-by-document matrix. More specifically, terms are encoded as vectors, whose components are co-occurrence frequencies of words in corpora documents. Frequencies are weighted according to the log-entropy with respect to a *tf-idf* weighting schema (Baeza-Yates & Ribeiro-neto, 1999). Finally, the number of dimensions is reduced through dimension reduction methods. Vector-based representation enables words, sentences, and sets of synonyms (i.e. WordNet synsets) to be represented in a unifying way using vectors. VSM provides a variety of definitions of distance between vectors, corresponding to different measures of semantic similarity. In particular, we take advantage of the cosine angle between an input vector (input sentence) and an emotional vector (i.e. the vector representing an emotional synset) as a similarity measure to identify which emotion the sentence connotes.

4.1.1 WordNet-Affect

According to Strapparava, Valitutti, and Stock (2006), affective words are classified into two groups. There are words such as "fear" and "cheerful" which refer directly to emotional states. These words are called *direct affective words*. On the other hand, *indirect affective words* have an indirect reference that depends on the context (e.g. "monster", "cry"). WordNet-Affect is an affective lexical resource that is essential for affective computing, computational humour, text analysis, etc. and it particularly has a lexical repository of direct affective words. WordNet-Affect is an extension of WordNet by means of selecting and labelling of synsets representing affective concepts. Besides, WordNet-Affect has an emotional hierarchy of *affective domain labels* (A-Labels) with which the synsets representing affective concepts are annotated (See Table 4.1 and Figure 4.2). In the hierarchy, additional A-labels are provided, hierarchically organised starting form a-label emotion with respect to WordNet-Affect. The hierarchy consists of approximately 1637 words and 918 synsets.

A-Label	Examples of Synsets
Emotion	noun "anger#1", verb "fear#1"
Mood	noun "animosity#1", adjective "amiable#1"
Trait	noun "aggressiveness#1", adjective "competitive#1"
Cognitive State	noun "confusion#2", adjective "dazed#2"
Physical State	noun "illness#1", adjective "all_in#1"
Hedonic Signal	noun "hurt#3", noun "suffering#4"
Emotion-Eliciting Situation	noun "awkwardness#3", adjective "out_of_danger#1"
Emotional Response	noun "cold_sweat#1", verb "tremble#2"
Behavior	noun "offense#1", adjective "inhibited#1"
Attitude	noun "intolerance#1", noun "defensive#1"
Sensation	noun "coldness#1", verb "feel#3"

Table 4.1: A-Labels and examples

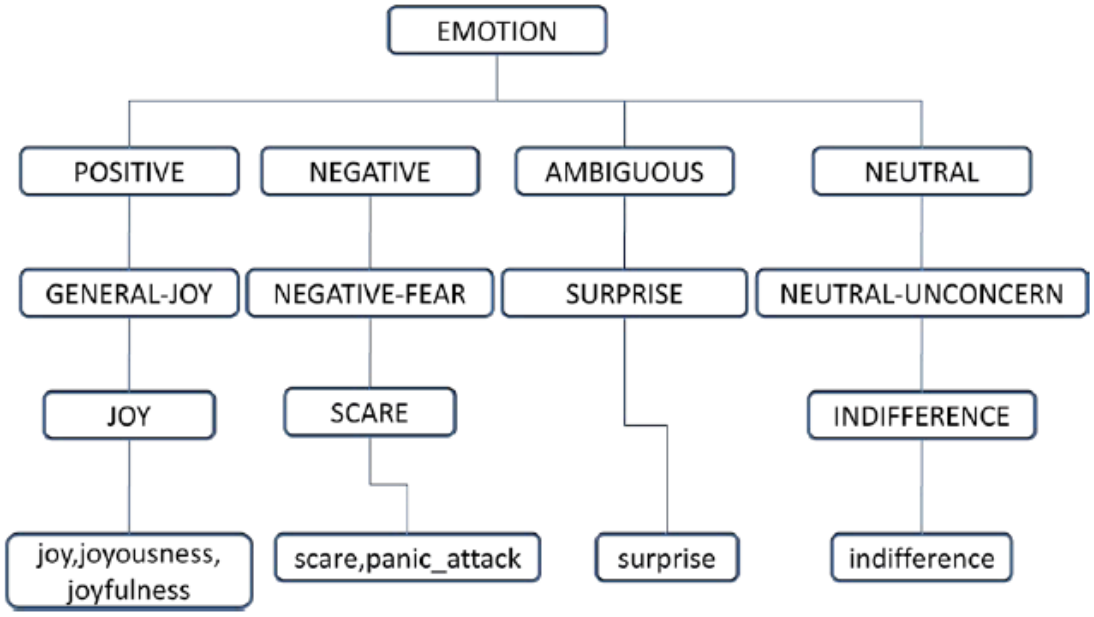


Figure 4.2: Emotional hierarchy

In addition to the specialisation of the emotional hierarchy and the A-label emotions, WordNet-Affect contains two types of tagging information as new extensions. Firstly, *stative/causative tagging* is concerned mainly with the adjectival interpretation. An emotional adjective is called *stative* if it refers to some emotion owned or felt by the subject denoted by the modified noun (e.g. "**cheerful** boy"). *Causative* means that emotion is caused by the entity represented by the modified noun (e.g. "**amusing** movie"). Stative/causative terms are concerned with the subjective/objective quality of the text. Stative or causative labels are marked on emotional synsets (Adjectives, Verbs, and Adverbs) of the hierarchy, which is shown in Table 4.2. There are about 450 stative synsets and 250 causative synsets in WordNet-Affect. The second tagging is *valence* that provides *positive* or *negative* information. *Valence* tagging distinguishes synsets into the following four categories according to emotional valence: *positive*, *negative*, *ambiguous*, and *neutral*. The synsets *enthusiasm#1* and *horror#1* are examples of *positive* and *negative* emotions, respectively. *Surprise#1* is an *ambiguous* type of synset when the valence depends on the context. Lastly, the valence is tagged as *neutral* when the synset

is considered affective but not characterised by valence (*indifference#1*). Valence tagging can be represented on the emotional hierarchy and it is located in the second level of Figure 4.2.

Emotional category	Stative terms	Causative terms
Disgust	disgusted	disgusting, disgust, disgustingly
Encouragement	encouraged, backed_up	encouraging, encourage, encouragingly

Table 4.2: Stative and Causative terms

WordNet-Affect, a linguistic resource for the lexical representation of affective knowledge, has been used with the aim of supporting applications relying on language recognition and generation. For instance, Strapparava, Valitutti and Stock (2007) described automatic textual emotion recognition and its visualisation by kinetic typography (text animation). In order to analyse affective content, the authors were using not only affective words from WordNet-Affect, but also an affective lexicon derived from the evaluation of the semantic similarity between generic terms and affective concepts. The latter methodology will be explained in more detail in the following section.

4.1.2 Vector Space Model

There are various methodologies by which terms and documents can be represented to capture the relative importance of the terms in a document or to measure the similarity between terms or documents. The most popular and common family of text presentation techniques is based on the VSM in which textual documents can be represented through a term-by-document matrix. This methodology was introduced by Gerard Salton (Berry & Browne, 2005; C D Manning, Raghavan, & Schütze, 2009), often referred to as the father

of modern information retrieval on account of his methodology's influence as a general modern form of language. A set of documents can be expressed as term vectors or document vectors in VSM and this space model is fundamental to information retrieval such as document classification and document clustering. Both terms and documents are encoded as vectors in k-dimensional space. The choice k can be based on the number of unique terms, topics, or classes related to the text corpus. In that way, each vector is used in order to reflect the significance of the corresponding term, topic, or class in representing the semantics of a document. This technique is referred to as dimensionality reduction or feature selection since the total number of dimensions is decreased to k. Three representative reduction methods are introduced in the following section. VSM can exploit geometric relationships between document (and term) vectors so as to explain similarities and differences in classes. As mentioned before, WordNet-Affect is concerned with direct affect words. However, a technique is also necessary for evaluating the affective weight of *indirect affective words*. The mechanism is based on similarity between generic terms and affective lexical concepts in the VSM. VSM has the following three characteristic factors: a local term weight, global term weight, and a similarity measure. These characteristics will be depicted in more detail in the following sections.

Term-by-Document (TDM) and Term-by-Sentence Matrix (TSM)

A collection of n documents which is comprised of m terms can be represented as an $m \times n$ Term-by-Document Matrix (TDM). The n column vectors correspond to the n documents, whereas the m row vectors represent the m terms in the matrix. The former are interpreted as the document vectors and the latter are considered the term vectors in VSM. TDM tends to normally have several tens of thousands of rows and columns even for a collection of moderate size. Terms are generally stemmed before indexing which causes reduced dimensions and it is a kind of feature selection. This reduced matrix is useful in manipulating the contextual matrix. The matrix element assigned to a cell means the importance of the term in representing the meaning of the document and the value a_{ij}

is the weighted frequency at which the *i*-th term occurs in the *j*-th document. Some weighting schemes are discussed in the next section.

The definition of documents in TDM is not confined to only the pure meaning of document. A document can be decomposed into a number of hierarchical components such as paragraphs and sentences. The hierarchy level of reasonable decompositions depends on the granularity of the semantic analysis, that is, the purpose of the application. Emotion detection can occur at several levels: the term level, the sentence level, the passage level, and the document level. For instance, the document level is enough for identifying the whole sentiment of a document. The level of granularity is restricted by the level of annotated corpus for the evaluation of classification results. The annotated level of corpus should correspond to the results of computing the performance. Sentence-based emotion analysis is performed and each document is decomposed into non-overlapping and sequential sentences, which correspond to the document, and the divided sentences are stored into the column of Term-by-Sentence Matrix (TSM). For convenience, the terminology "document" is considered the same as "sentence" in this thesis.

Figure 4.3 demonstrates how each sentence can be represented as an 11 × 4 TSM in a VSM from a collection of ISEAR (International Survey on Emotion Antecedents and Reactions) dataset. In this simple example, each column defines a document, while each row corresponds to a unique term or keyword in the text corpus. Because a stopword list contains words like *was*, *in*, etc. that are not taken into account, only a bolded subset of all the words used in the four sentences are selected as terms for the purpose of indexing. In general, standard pre-processing steps are performed, namely, stopwords removal and word stemming before creating a TSM. In the matrix example shown in Figure 4.3, not all the words are considered to describe the content of sentences. Only words like "*traffic*" and "*friends*" are selected and which words to index and which words to discard are determined in the pre-processing stages. In the construction of a TSM, terms are usually identified by their word stems after getting rid of common words. For example, the words "*passing*" and "*pass*" are treated as the same term and they are considered as a single dimension (one single term). Stemming reduces the number of rows in the TSM and this

reduction is certainly a significant ingredient for large collections of documents in terms of data storage.

The values stored in each matrix element show the frequency in which a term occurs in a document. For instance, a term "*exam*" appears once in the document *S3* but not in the other three documents. Notice that a particular term occurs only once in any given document. Figure 4.4 shows how each row and column of the 11 × 4 matrix in Figure 4.3 can be represented as a vector in the reduced 3-dimensional space. The dimension reduction techniques are explained in Section 4.1.3.

Sentences from ISEAR dataset

Sentence1 (S1): When I was **involved** in a **traffic accident**.

Sentence2 (S2): **Friends** who **torture animals**.

Sentence3 (S3): **Passing** an **exam** I did not **expect** to **pass**.

Sentence4 (S4): When someone **stole** my **bike**.

		Documents			
		S1	S2	S3	S4
Terms	accid	1	0	0	0
	anim	0	1	0	0
	bike	0	0	0	1
	exam	0	0	1	0
	expect	0	0	1	0
	friend	0	1	0	0
	involv	1	0	0	0
	pass	0	0	2	0
	stole	0	0	0	1
	tortur	0	1	0	0
	traffic	1	0	0	0

Figure 4.3: The construction of a Term-by-Sentence Matrix (The 11 × 4 term-by-sentence matrix where the element $a_{i,j}$ is the number of times term i appears in sentence j)

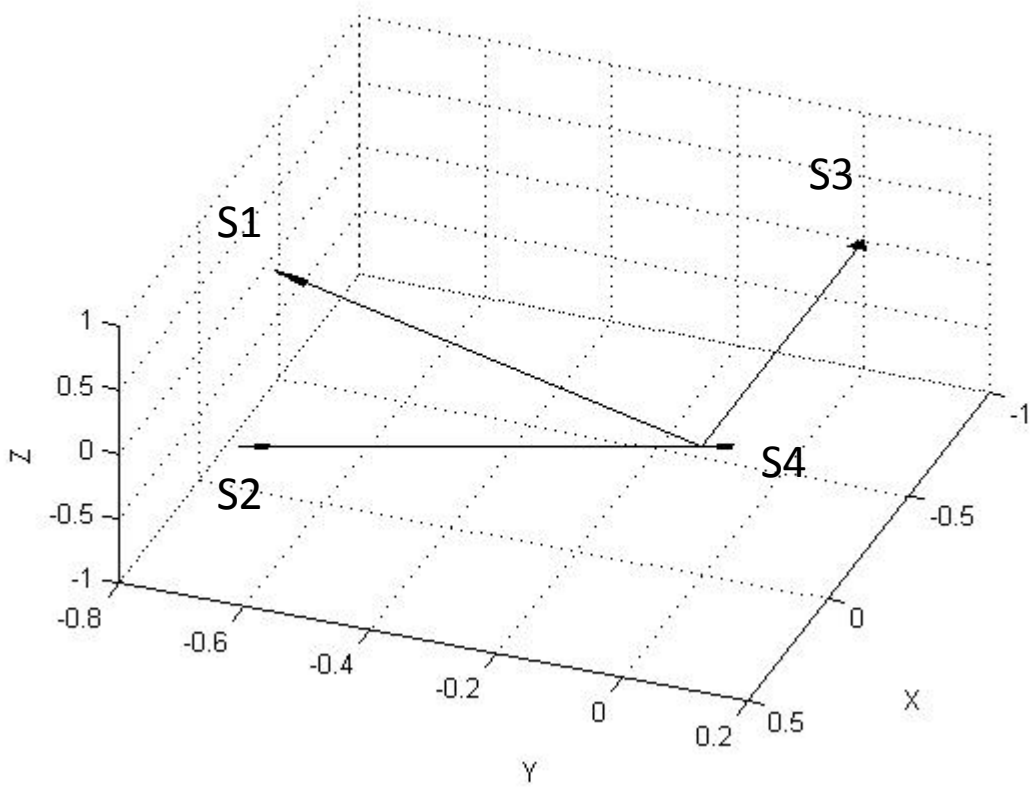


Figure 4.4: Representation of sentences in a 3-dimensional vector space

All the words, texts, and synsets can be represented homogeneously on the VSM. More specifically each text or synsets can be represented in the VSM exploiting a variation of the pseudo-document methodology by summing up the normalized vectors of all the terms contained in it as in the following representation. Figure 4.5 displays a synset existing in a 3-dimensional space together with a term and a sentence. In particular, an emotional synset vector for each emotion consists of the synonyms of the emotion from WordNet-Affect. For example, *wrath*, *pique*, and *chafe* can be *w1*, *w2*, and *w3* in an *anger* synset vector, respectively.

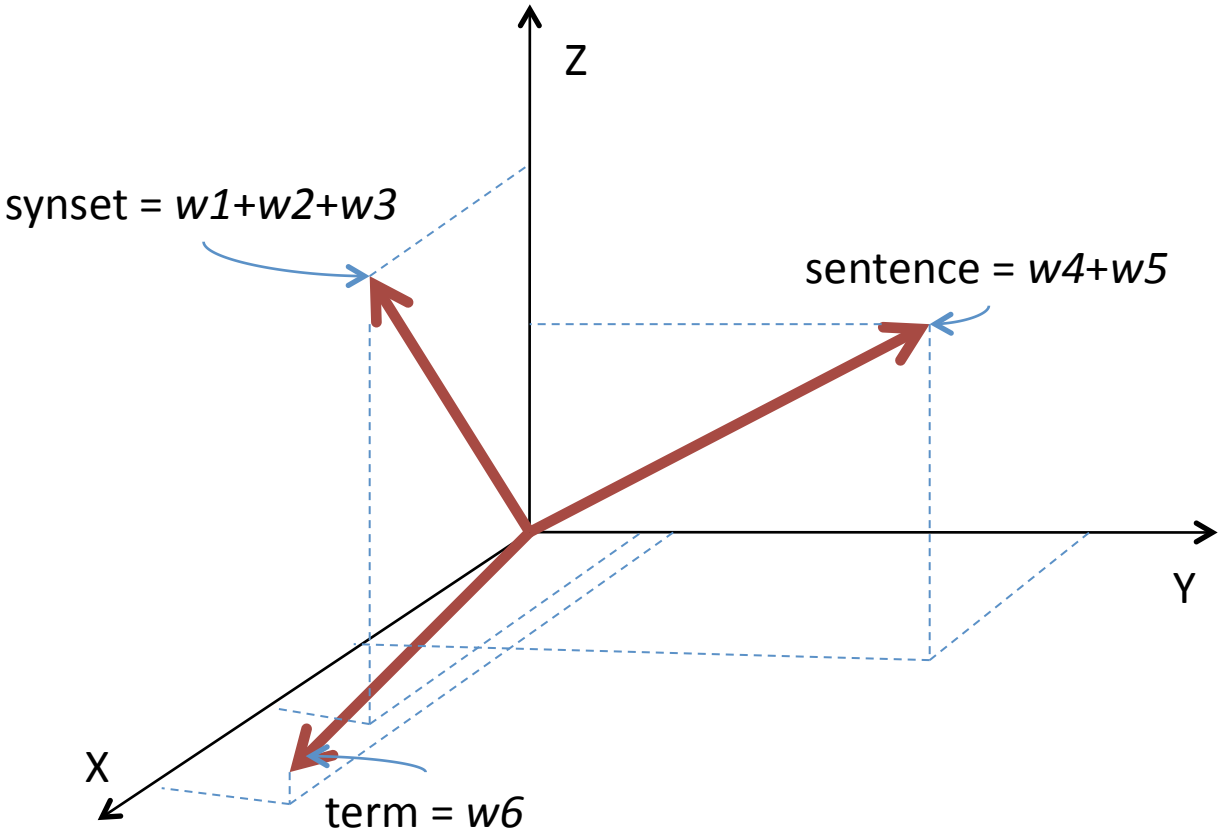


Figure 4.5: The representation of a term, a sentence, and a synset on VSM

Term Weighting

Each element of TSM denotes term frequencies that refer to how many times a term occurs in a document. This is one type of weighting scheme which helps account for both the significance of terms in the particular text units such as sentences as well as the degree to which the term carries information in the specific domain in general. The main purpose of term weighting is to improve retrieval performance. Performance means the ability to retrieve relevant information and dismiss irrelevant information. Term weighting is one approach commonly used to enhance the performance of the information retrieval system. If each element a_{ij} is defined in a TSM, the cell component can be represented as follows:

$$a_{ij} = l_{ij} \times g_i \times d_j$$

where l_{ij} indicates the local weight for term i occurring in the document j, g_i is the global weight for term i in the collection, and d_j is a document normalisation factor which specifies whether or not the columns of a TSM (i.e. the documents) are normalised.

The following table shows the popular weight schemes when manipulating text collections in VSM. For convenience, define

$$\psi(r) = \begin{cases} 1, & r > 0 \\ 0, & \text{otherwise} \end{cases}$$

tf_{ij} denotes the number of times or frequency that term i appears in document j and let $p_{ij} = \frac{tf_{ij}}{\sum_j tf_{ij}}$.

Local Term Weight (l_{ij})		Global Term Weight (g_i)		Normalisation (d_j)	
Binary (b)	$\psi(tf_{ij})$	None (x)	1	None (x)	1
Logarithm (l)	$1 + \log(tf_{ij})$	Entropy (e)	$1 + \left(\sum_j \left(p_{ji} \log p_{ij}\right)/\log n\right)$	Cosine (c)	$\frac{1}{\sqrt{\sum_i \left(g_i l_{ij}\right)^2}}$
Augmented (n)	$0.5 + \frac{0.5 \times tf_{ij}}{\max_i tf_{ij}}$	IDF (f)	$\log \frac{n}{\sum_j \psi(tf_{ij})}$		
Frequency (t)	tf_{ij}	GfIdf (g)	$\frac{\sum_j tf_{ij}}{\sum_j \psi(tf_{ij})}$		
		Normal (n)	$\frac{1}{\sqrt{\sum_j tf_{ij}^2}}$		
		Probabilistic inverse (p)	$\log \frac{n - \sum_j \psi(tf_{ij})}{\sum_j \psi(tf_{ij})}$		

Table 4.3: Formulas for local/global term weights and document normalisation

Therefore, a term-weighting scheme can be simply expressed through the combination of a three-letter string corresponding to the desired local, global, and normalisation factors. Term weight, for example, $a_{ij} = t \times g \times c$, stands for the usages of

Frequency for local term weight, GfIdf for global term weight, and Cosine for normalisation.

Log(l)-Entropy(e) in the VSM has traditionally been one of the best known and the most effective weighting schemes. Moreover, the Log(l)-Entropy(e) term weighting function has been taken into account in most of the literatures in order to improve the conceptual discrimination of documents and the global importance of a keyword across the collection. For these reasons, the Log(l)-Entropy(e) weighting schema is utilised in our experiment for the purpose of computing the cell value of TSM.

Text Semantic Similarity Measures

A number of different methods have been devised and compared for evaluating the semantic similarity of documents such as *Dice*, *Jaccard*, *overlap*, *Lin*, *α-Skew* and *Js-Div* (Lin, 1998; Christopher D. Manning & Schütze, 2000; Pereira, Tishby, & Lee, 1993). We can represent documents as vectors through VSM. The vector-based semantic representation of documents leads to a lot of benefits to take advantage of various linear algebra operations. On top of that, terms, sentences, paragraphs, and documents can be represented in a uniform way on the vector space (Strapparava & Valitutti, 2004). Similarity between sentences is computed in one of two ways. The first technique is the dot-product calculating the Euclidean distance between the two vectors and the second method is the cosine similarity taking a measure of the angle between the vectors. The latter technique emphasises the directions of vectors rather than the lengths of them. Euclidean distance is a generally used measure of similarity in the area of image data, while cosine similarity is widely exploited in the realm of text data (Bingham & Mannila, 2001). In particular, cosine similarity is commonly used in some supervised learning algorithms for document categorization. For instance, given a new document, cosine similarity is used to find the most relevant documents whose categories are used to assign categories to a new document.

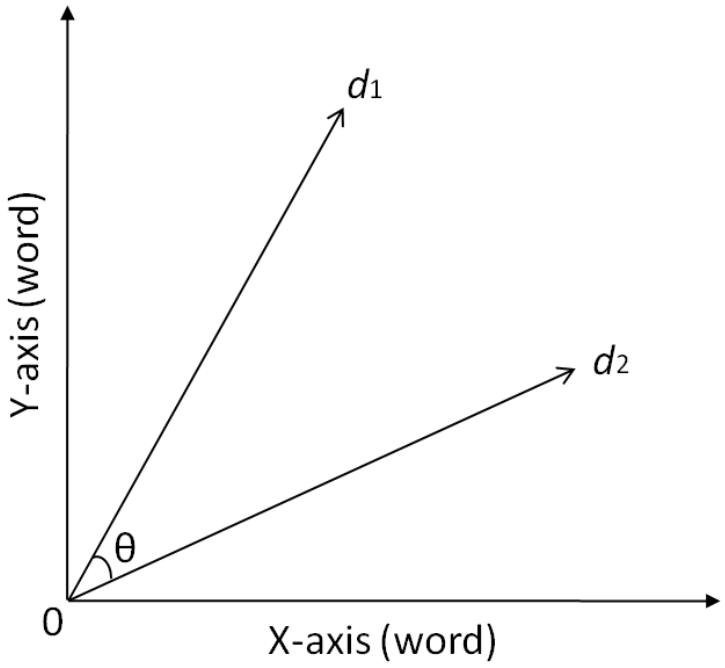


Figure 4.6: Cosine similarity between documents. sim(d_1, d_2) = cos θ

According to Mihalcea, Corley, and Strapparava (2006), the similarity in a lower dimensional semantic space is measured by the standard cosine similarity. More importantly, a vector space model allows words, synsets, and sentences to be compared with each other as well as to be represented homogeneously. In practice, the cosine of the angle between an input vector (input sentence) and an emotional vector (emotional synsets) is computed to identify which emotion the sentence connotes (Strapparava & Mihalcea, 2008). Linear combinations of emotional synonym vectors are used in order to form a vector representation of one emotion. The more closely two vectors are related semantically, the higher their cosine value is. As an example, we assume that the input sentence is "In a cottage in a large forest, I was alone for a while in the dark", and then a similarity calculation is performed. If the calculated cosine values between the input vector and "*joy*" vector, and "*fear*" vector are 0.3 and 0.7, respectively, it is concluded that the input sentence implies the *fear* emotion.

Cosine similarities can be defined in these representations, and here, as other authors have done, we use a rule that if the cosine similarity does not exceed a threshold, the input sentence is labelled as "*neutral*", the absence of emotion. Otherwise, it is

labelled with one emotion associated with the closest emotional vector having the highest similarity value. A predetermined threshold (t = 0.65) is used for the purpose of validating a strong emotional analogy between two vectors (Penumatsa et al., 2006).

If similarity is defined between a given input text, I, and an emotional class, E_j, as $\text{sim}(I, E_j)$, the categorical classification result, CCR, is more formally represented as follows:

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

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

4.1.3 Dimension Reduction Methods

Matrix dimensionality reductions derive from the following definition that there will never be a perfect TSM that accurately represents all possible term-sentence associations. These uncertainties associated with TSM largely ascribe to lingual and cultural distinctions of each person having different experiences and opinions. For a simple example, there are fundamental discrepancies in word usage between authors and readers that give rise to different understandings and interpretation of the same textual information. Because the association of terms to sentences is subject to each interpretation, the TSM A may be better represented by the matrix sum $A + E$, in which the uncertainty (or error) matrix E has values reflecting missing or incomplete information, or different opinions about documents in generating the elements of matrix A.

Determining the optimal number of dimensions to encode the original TSM is an open question and matrix factorisation can be used to produce a reduced dimensionality. This is also referred to as low-rank approximations. In particular, VSM representation can be reduced with techniques well known in Information Retrieval: Latent Semantic Analysis (LSA) (Thomas K Landauer, McNamara, Dennis, & Kintsch, 2007), Probabilistic LSA (PLSA) (Thomas Hofmann, 1999), or the Non-negative Matrix

Factorisation (NMF) (Lee & Seung, 1999) representations. Dimensionality reduction in VSM reduces the computation time and reduces the noise in the data. This enables unimportant data to dissipate and underlying semantic text to become more patent. We will review three popular statistical dimensionality reduction methods (LSA, PLSA, and NMF) that are utilised in a category-based emotion model in the following sections.

Singular Value Decomposition (SVD), which is the basis of the method of Latent Semantic Analysis (LSA), is one popular method that computes the optimal dimension reductions for the vector space representation of the data objects in VSM in order to encode m terms and n sentences in k-dimensional space.

The value k is generally expressed through the following: $k << C \min(m,n)$ where k is much smaller than the number of dimensions. Figure 4.4 shows a 3-dimensionality to the matrix from Figure 4.3 and this enables both terms and documents to be represented in three dimensions. Unlike the original vector space model, the coordination applied to dimension reductions does not explicitly reflect term frequencies in the corresponding documents. The transformed matrix contains hidden semantic information in regard to the associations of term-sentence that cannot be found in the traditional vector space coordinates. Furthermore, reducing the rank of the matrix is a means of removing extraneous information or noise from the database in terms of saving space.

The dimension reduction problem consists mathematically in finding A_k of rank k, where A_k is defined as follows:

$$A_k = \min_{X:dim(X)=k} \|A - X\|_F$$

A_k and X are both $m \times n$ matrices. In addition, $\|A - X\|_F$ is called the Frobenius matrix norm, which is defined for the real $m \times n$ matrix A by

$$\|A\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2}$$

If only k singular values are retained, the rest of the matrix is set to 0 and the parts of the matrix are not needed. That means the rank of the matrix A is equal to the number of non-zero singular values. From TSM A, we can compute the approximation A_k and

there is a row for each term and a column for each sentence in A_k. The original matrix is replaced by another matrix that is as close as possible to the original matrix but whose column space is only a subspace of the column space of the original matrix. Dimension reduction methods are used in image processing, data compression, and cryptography as well as information retrieval.

Latent Semantic Analysis (LSA) is the earliest approach successfully applied to various text manipulation areas (T.K. Landauer, Foltz, & Laham, 1998). The main idea of LSA is to map terms or documents into a vector space of reduced dimensionality that is the latent semantic space. The mapping of the given terms/document vectors to this space is based on SVD. It is known that SVD is a reliable technique for matrix decomposition. It can decompose a matrix as the product of three matrices.

$$A = U\sum V^T \approx U_k\sum_k V_k^T = A_k$$

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

Probabilistic Latent Semantic Anlaysis (PLSA) (T. Hofmann, 2001) has two characteristics distinguishing it from LSA. PLSA defines proper probability distributions and the reduced matrix does not contain negative values. Based on the combination of LSA and some probabilistic theories such as Bayes rules, the PLSA allows us to find the *latent topics*, the association of documents and topics, and the association of terms and topics. In the equation, *z* is a *latent class variable* (i.e. discrete emotion category), while *w* and *d* denote the elements of term vectors and document vectors, respectively.

$$P(d,w) = \sum_z P(z)P(w|z)P(d|z)$$

where *P(w|z)* and *P(d|z)* are topic-specific word distribution and 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 through adopting the Expectation Maximisation (EM) algorithm. For document similarities, each

row of the *P(d|z)* matrix is considered with the low-dimensional representation in the semantic topic space.

Non-negative Matrix Factorisation (NMF) (Lee & 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*.

$$A \approx WH = \sum WH$$

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

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

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 documents is performed based on the columns of matrix *H* that represent the documents.

4.2 Dimensional Emotion Estimation

In this approach, the main idea is mapping the locations of input text and each emotion into the dimensional representation, then the closest emotion to the resulting point is assigned to the input text (distance-based approach). Dimensional models provide an explicit notion of the degree of similarity between emotions and input text. Namely, adjacent emotion and text in the space are very similar while those opposite them are different from each other. The coordinate of input text is calculated through the average location of words in the text. Similarly, the location of each emotion is computed in order to obtain the resulting projection in the given dimensional space for taking each annotation of emotion. The numerical locations can be obtained by means of figures with

a setting of points representing the words and the emotions from ANEW. Note that not all the words involved in the input text and the emotion can be directly mapped into the dimensional representation by ANEW alone. The words of input text are just discarded if the words do not exist in ANEW. In contrast, we make use of the synonyms regarding each emotion given by WordNet-Affect in order to overcome this discrepancy.

4.2.1 ANEW

Dimensional models have been studied by psychologists often providing a stimulus (e.g. a photo or a text), and then asking subjects to report on the affective experience. The *Affective Norms for English Words* (ANEW) is a set of normative emotional ratings for a collection of English words (Bradley & Lang, 1999), where after reading the words that are content independent, subjects reported their emotions in a three dimensional representation. It contains a set of 1,034 English words including verbs, nouns, and adjectives and it has the values of valence, arousal, dominance and word frequency. This collection also provides mean and standard deviations of each word for the three dimensions rated by the Self Assessment Manikin (SAM). ANEW was derived by running the psycholinguistic experiments in mixed groups of 8 to 25 subjects and analysing the data by three different groups such as Male, Female, and All Participants. Tables 4.4, 4.5 and 4.6 show some example terms, their mean and standard deviations, and word frequencies from ANEW for each group, respectively.

Affective annotations in ANEW are given as three numbers (1-10) describing three dimensions of affect given by the widely used PAD model (Mehrabian, 1995), that stands for (P)leasure-displeasure, (A)rousal-nonarousal, and (D)ominance-submissiveness. The advantage of PAD is that it unifies emotional ontologies – for example, “fear” is low-pleasure, high-arousal, and low-dominance – and allows delicate emotional differences, such as between “joy” (high-arousal) and “contentment” (lower-arousal). The two-dimensional projection of ANEW words is shown in Figure 4.7. The following tables and figure represent that there are subtle emotional differences between men and women.

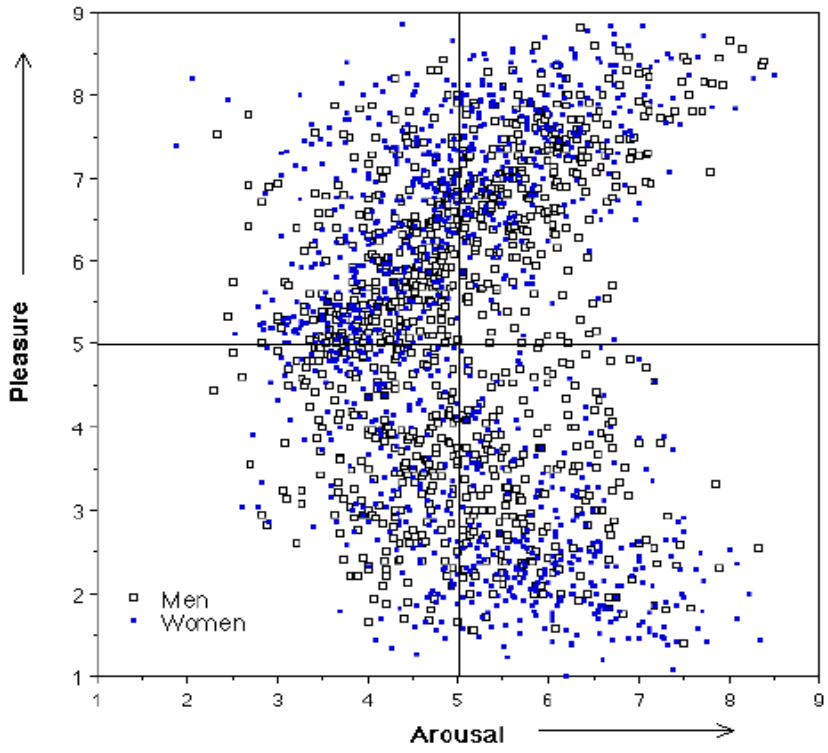


Figure 4.7: Two-dimensional affective map of ANEW terms (Bradley & Lang, 1999)

Description	Word No.	Valence Mean	Valence SD	Arousal Mean	Arousal SD	Dominance Mean	Dominance SD	Word Frequency
abduction	621	2.76	2.06	5.53	2.43	3.49	2.38	1
betray	37	1.68	1.02	7.24	2.06	4.92	2.97	4
cabinet	675	5.05	0.31	3.43	1.85	4.73	1.66	17
earth	134	7.15	1.67	4.24	2.49	5.61	2.30	150
frustrated	177	2.48	1.64	5.61	2.76	3.50	2.12	10
irritate	235	3.11	1.67	5.76	2.15	5.03	2.05	.
startled	410	4.50	1.67	6.93	2.24	4.48	1.57	21

Table 4.4: Example ANEW terms from All Subjects

Description	Word No.	Valence Mean	Valence SD	Arousal Mean	Arousal SD	Dominance Mean	Dominance SD	Word Frequency
abduction	621	3.19	1.94	4.95	1.99	4.25	2.05	1
betray	37	1.85	1.14	7.17	2.08	5.75	1.91	4
cabinet	675	5.10	0.31	3.45	1.70	4.70	1.81	17
earth	134	6.79	1.72	4.08	2.10	6.69	2.10	150
frustrated	177	2.70	1.56	5.55	2.61	3.70	1.98	10
irritate	235	2.87	1.64	5.13	2.39	5.20	2.27	.
startled	410	4.41	1.12	6.53	2.58	5.24	1.79	21

Table 4.5: Example ANEW terms from Male Subjects

Description	Word No.	Valence Mean	Valence SD	Arousal Mean	Arousal SD	Dominance Mean	Dominance SD	Word Frequency
abduction	621	2.33	2.13	6.09	2.72	2.76	2.49	1
betray	37	1.60	0.96	7.28	2.09	4.54	3.31	4
cabinet	675	5.00	0.31	3.57	1.91	4.76	1.55	17
earth	134	7.35	1.65	4.32	2.70	5.04	2.23	150
frustrated	177	2.31	1.72	5.65	2.92	3.35	2.24	10
irritate	235	3.26	1.71	6.17	1.92	4.91	1.93	.
startled	410	4.56	1.98	7.19	2.00	3.96	1.17	21

Table 4.6: Example ANEW terms from Female Subjects

In this study, we suggest that the emotional dimensions that ANEW provides have a relation with the aforementioned dimensional model. We make use of ANEW for marking up sentences with emotional dimensions by looking for the words which appear in the sentences as well as in the ANEW.

4.2.2 Three-Dimensional Estimation

We will find the answer to the question of how an input sentence is mapped into three-dimensional space and how the ANEW-based approach will perform in the dimensional estimation. As a first attempt for predicting the emotion of the input text, mapping is required onto the dimensional representation. In this environment, the *dominance* dimension has been maintained, which is different from the circumplex model only representing the valence and the activation of emotions.

For each word *w*, the normative database provides coordinates $\overline{w}$ in an affective space as:

$$\overline{w} = (valence, arousal, dominance) = ANEW(w)$$

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

As a counterpart to the categorical classification above, this approach assumes that an input sentence pertains to an emotion based on the least distance to its neighbours on the Valence-Arousal-Dominance (VAD) space. The input sentence consists of a number of words and the VAD value of this sentence is computed by averaging the VAD values of the words:

$$\overline{sentence} = \frac{\sum_{i=1}^{n} \overline{w}}{n}$$

where *n* is the total number of words in the input sentence. The above equation provides the VAD value of a sentence on the emotional dimensions.

Since not many words are available in this normative database, a series of synonyms from WordNet-Affect are used in order to calculate the position of each emotion. 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:

$$\overline{emotion} = \frac{\sum_{i=1}^{k} \overline{w}}{k}$$

where k denotes the total number of synonyms in an emotion. *Anger*, *fear*, *joy*, and *sadness* emotions are mapped on the VAD space. Let A_c, F_c, J_c, and S_c be the centroids of four emotions. Then the centroids, which are calculated by the above equation, 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, *neutral* is manually defined to be (5, 5, 5). If the centroid of an input sentence is the most approximate to that of an emotion, the sentence is tagged as the emotion (with the nearest neighbour algorithm). The centroid $\overline{sentence}$ might be close to an $\overline{emotion}$ on the VAD space, even if they do not share any terms in common. Hence, the centroid-based classifier can discern emotion classes and label each sentence in the corpus with the nearest emotion. The distance threshold (empirically set to 4) is defined to validate the appropriate proximity like the categorical classification.

Figure 4.8 shows the resulting distribution for two conflicting emotions (*joy* and *sadness*) in the affective space. The sentences that pertain to the same emotional category are scattered with the same colour all over the emotional plane. The sentences represented by red circles hold a positive emotion, *joy*. In contrast, blue 'X's mean the sentences with negative meaning, *sadness*. This dimensional estimation approach is successfully able to differentiate emotions by means of considering both data distribution and centroids.

4.3 Summary

An overview of two methods based on two emotion representations has been given in this chapter. It has been described how categorical emotion classification and dimensional emotion estimation are used to detect emotions in text. Categorical classification method utilises WordNet-Affect as a linguistic resource and vector space model for measuring the similarity between input text and emotion category. Dimension reduction methods

enable hidden semantic meaning behind the text to become more evident in the categorical emotion classification.

Dimensional estimation method also takes advantage of a lexical repository (ANEW) but this method relies on the coordinates in the VAD space to find the closest emotion to input text. The results of categorical and dimensional methods are assessed through evaluation measures, which are presented in the succeeding chapter.

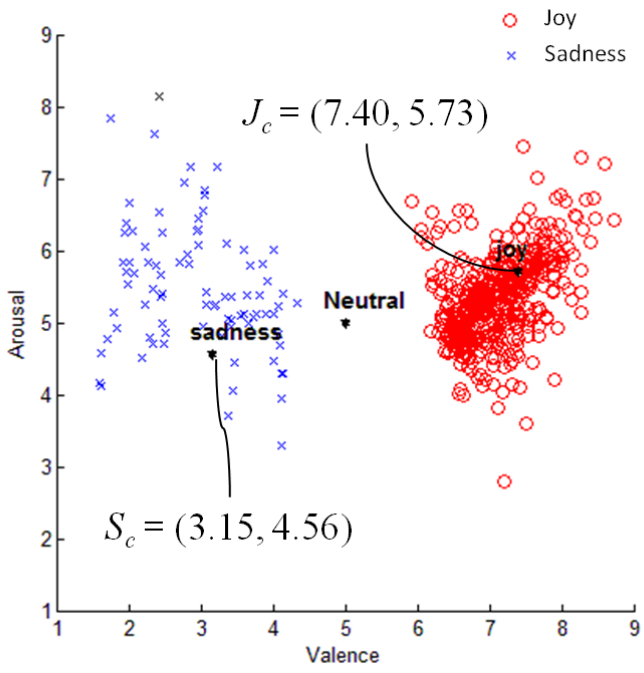


Figure 4.8: Distribution of emotions and Fairy Tales sentences in the sentiment space

CHAPTER 5
Evaluation

"*Emotions may be defined as an aroused state of the organism involving conscious, visceral, and behavioral changes. Emotions are therefore more intense than simple feelings, and involve the organism as a whole.*"

- James P. Chaplin, 1975 (Kleinginna & Kleinginna, 1981).

Identifying emotion in text is a type of text classification problem. This chapter presents some measures that are used to evaluate classifiers. There are many measures used in text classification (Van Rijsbergen, 1979). However, we introduce only the most commonly used measures. We review the precision, recall, and F-measure, followed by a look at Cohen's kappa (Cohen, 1960). The common metric of interest in evaluating classification systems is how accurate the system is. More specifically, what portion of data belonging to known classes (labelled data) is correctly assigned to those classes.

Four datasets were used for the purpose of detecting emotion or sentiment. Three of these, with sentence-level emotion annotations, are for evaluating emotion identification techniques (news headlines, reports on personal experiences and children's fairy tales). One dataset, which contains students' descriptions of their learning experiences, is utilised for sentiment classification.

5.1 Methodologies

5.1.1 Precision, Recall, and F-Measure

Classification accuracy is usually measured in terms of precision, recall, and F-measure. First of all, two basic measures (precision and recall) are explained for a given document. These are computed as follows:

$$Precision = \frac{\text{categories found and correct}}{\text{total categories found}} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$Recall = \frac{\text{categories found and correct}}{\text{total categories correct}} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

True Positive (*TP*) refers to the number of examples that are classified correctly as belonging to the class, while *False Positive* (*FP*) stands for the number of incorrectly classified examples. *False Negative* (*FN*) is the number of examples we incorrectly classify as negative.

To make it easy to understand, a binary classification problem is taken as an example. There is a positive class and a negative class for a binary classifier. A 2-by-2 confusion matrix shown in Figure 5.1 shows the number of documents predicted correctly and incorrectly into the two classes.

		Predicted	
		Positive	Negative
Actual	Positive	*a*	*b*
	Negative	*c*	*d*

Figure 5.1: 2-by-2 confusion matrix

The definition of four values is as follows for a given class:

- *a* – number of documents correctly assigned to the class (*True Positive*)

- b – number of documents incorrectly assigned to the class (*False Negative*)
- c – number of documents incorrectly rejected from the class (*False Positive*)
- d – number of documents correctly rejected from the class (*True Negative*)

Based on these values, $(a + b)$ are the number of documents which truly belong to the positive class. In contrast, (c + d) documents belong to the negative class in the confusion matrix. Precision and recall performance measures are defined and computed from these values.

$$Precision = \frac{a}{(a+b)}, \qquad \text{Recall} = \frac{a}{(a+c)}$$

These two standard measures can be represented with a Venn diagram. Figure 5.2 shows a representation of this measurement applied to the domain of classification.

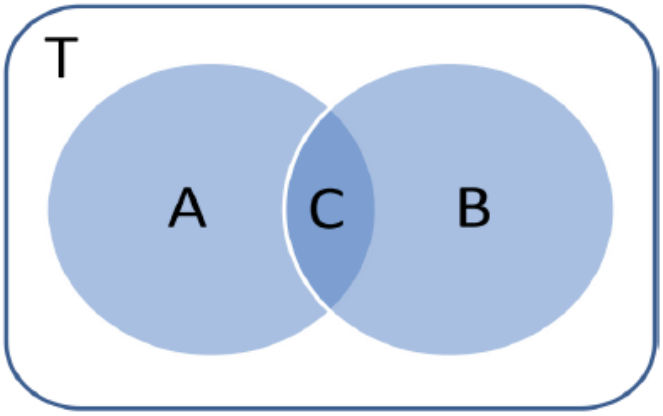


Figure 5.2: The Venn diagram representation of precision and recall

In the diagram, T is the set of all test documents and A is the set returned by the classifier. In addition, B is a set of documents in a given class and C is the set of correctly classified documents, which is the intersection of A and B. Let $|A|$ be the cardinality of set A (the number of elements in A). Precision and recall can be defined by the following equations:

$$Precision = \frac{|C|}{|A|}, \qquad \text{Recall} = \frac{|C|}{|B|}$$

These two measures indicate the tradeoff between specificity and coverage. Figure 5.3 represents this tradeoff between precision and recall. The first diagram is the case in which high precision is provided, whereas high recall is shown in the second diagram.

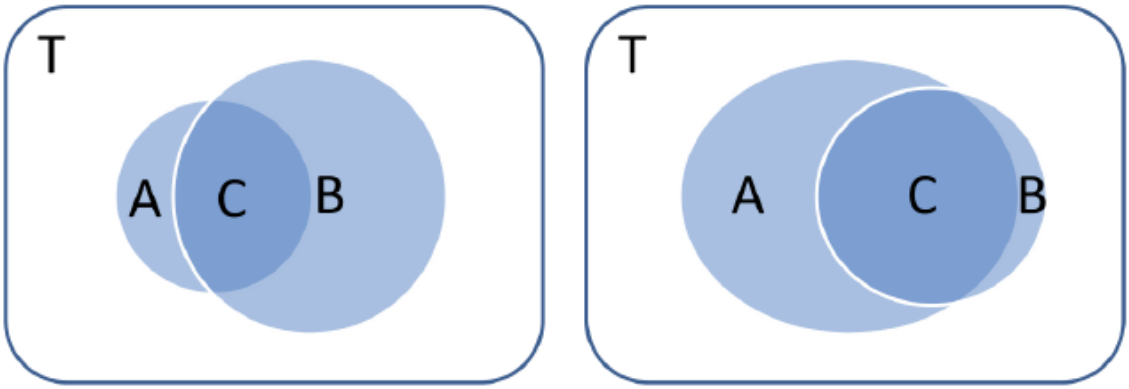


Figure 5.3: High precision and high recall

However, in general precision and recall are not taken into account alone due to this variability. The variability may mislead some of the performance measures. For this reason, two new alternatives are used: breakeven point and F-measure. Breakeven point is a value when precision and recall are the same and provides a single score that relates the two values together. For our evaluation, the F-measure is used as a metric for effectiveness of classification. The F-measure is defined as follows:

$$F\ measure = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

The F-measure was designed to balance weights of precision and recall. The F-measure values are in the interval (0,1) and larger F-measure values correspond to higher classification quality. The measure is a popular metric for evaluating classification systems and is most often used to compare the performance of classifiers.

Typical classification systems make use of a set of classes, and the results from each class need to be averaged to represent overall performance. There are two methods

of combining the results: micro-averaging and macro-averaging. The former method is simply to calculate each class performance individually, and then average those into one value. Micro-averaging is heavily influenced by the performance of highly condensed classes in the case of a skewed instance distribution. The latter method is the simple average of all values without weighting. Macro-averaging ignores the skew in the dataset distribution and treats all classes equally. The overall precision, recall and F-measure are calculated based on these averaging techniques. The following equations are the respective macro-averaging of precision, recall and F-measure.

$$P_{\text{ma}} = \frac{1}{C}\sum_{i=1}^{C} p_i\,, \qquad R_{\text{ma}} = \frac{1}{C}\sum_{i=1}^{C} r_i\,, \qquad F_{\text{ma}} = \frac{1}{C}\sum_{i=1}^{C} f_i$$

where C is total number of classes, and p_i, r_i, and f_i stand for precision, recall, and F-measure, respectively, for each category i.

Micro-averaging scores are defined as:

$$P_{\text{mi}} = \frac{\sum_{i=1}^{C} TP_i}{\sum_{i=1}^{C}(TP_i + FP_i)}, \qquad R_{\text{mi}} = \frac{\sum_{i=1}^{C} TP_i}{\sum_{i=1}^{C}(TP_i + FN_i)}, \qquad F_{\text{mi}} = \frac{2P_{\text{mi}}R_{\text{mi}}}{P_{\text{mi}} + R_{\text{mi}}}$$

For most datasets, the set of classes is not distributed uniformly (see Table 5.1 and 5.3). For instance, the USE dataset demonstrates skewing toward *Positive* class. Micro and macro-averaging produce very different results in this distribution. The experiment results are shown in Section 5.3 and 5.4.

5.1.2 Cohen's Kappa

The kappa statistic measures the proportion of agreement between two raters with correction for chance. The kappa score is used as the reliability metric to compare the performance of each classification approach (S. D'Mello & Graesser, 2007). The kappa measure is designed for categorical annotations and is defined as follows:

$$Kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

where *P(A)* is the proportion of the times the raters agreed, and *P(E)* is the proportion of the times the agreement would be made by chance. If two raters always agree, the kappa value is 1, and if they agree only at the rate given by chance, the value is 0. The negative kappa means that annotators are worse than random. So are there any criteria for interpreting kappa values? How large should kappa be to indicate good agreement? These are hard questions to answer. Kappa is not easy to interpret in terms of the precision of a single observation. In general, interpretation criteria depend on the type of assessment used. Table 5.1 gives guidelines for its interpretation. The table is from Landis and Koch (1977).

Kappa Value	Agreement Strength
< 0.20	Poor
0.21 – 0.40	Fair
0.41 – 0.60	Moderate
0.61 – 0.80	Good
0.81 – 1.00	Very good

Table 5.1: Interpretation of Kappa values (Landis & Koch, 1977)

Cohen dealt with only two raters. However, our experiments require more than two raters because we are dealing with a group of emotion categories. For this reason, we need to extend kappa to more than two raters. Fleiss (1971) extended Cohen's kappa to the study of agreement among a lot of raters. Any relationship is ignored among raters for different subjects in order to calculate Fleiss' kappa values. In addition, the Fleiss method does not take into consideration any weighting of disagreements. The Fleiss method gives the standard error of kappa not only for testing the null hypothesis of no agreement but also for many raters under the null hypothesis. If there is agreement and the distribution of kappa is not known, it means that confidence intervals and kappa comparison can only be approximate.

5.2 Emotion-Labelled Data

The following four datasets are employed in the evaluation of our categorical and dimensional methods. The first three (SemEval, ISEAR, Fairy Tales) have four emotion categories in common. The fourth (i.e. USE) does not have these categories and is discussed separately. The four datasets are also described in our publications (Sunghwan Mac Kim & Calvo, 2010; Mac Kim, Valitutti, & Calvo, 2010).

5.2.1 SemEval: News headlines

The first dataset is "Affective Text" from the SemEval 2007 task (Carlo Strapparava & Rada 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 express emotions in order to draw the reader's attention. This data-set has six emotion classes: *anger*, *disgust*, *fear*, *joy*, *sadness* and *surprise*, and is composed of 1,250 annotated headlines. The notable characteristics are that the SemEval dataset does not only allow one sentence to be tagged with multiple emotions, but the dataset also contains a *neutral* category in contrast to other datasets.

5.2.2 ISEAR

We also use the ISEAR (International Survey on Emotion Antecedents and Reactions) dataset, which consists of 7,666 sentences (Scherer & Wallbott, 1994), with regard to our experiments. For building the ISEAR, 1,096 participants who have different cultural backgrounds completed questionnaires about experiences and reactions for seven emotions including *anger*, *disgust*, *fear*, *joy*, *sadness*, *shame* and *guilt*.

5.2.3 Fairy Tales

The annotated sentences of the third dataset are culled from fairy tales (C. O. Alm, 2009). Emotions are particularly significant elements in the literary genre of fairy tales. The label set with five emotion classes is as follows: *angry-disgusted*, *fearful*, *happy*, *sad* and *surprised*. There are 176 stories by three authors: B. Potter, H.C. Andersen, and Grimm. The dataset is composed of only sentences with affective high agreements, which means that annotators highly agreed upon the sentences (four identical emotion labels).

Emotion	SemEval	ISEAR	Fairy tales	Total
Anger	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 5.2: Number of sentences for each emotion

In our study, we have taken into account four emotion classes (*Anger*, *Fear*, *Joy* and *Sadness*) which are in the intersection among three datasets (SemEval, ISEAR and Fairy tales). The number of sentences for each emotion and each dataset used in our experiment is shown in Table 5.2. In addition, sample sentences from the annotated corpus appear in Table 5.3.

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.

Table 5.3: Sample sentences labelled with sadness/sad from the datasets

5.2.4 Unit of Study Evaluation (USE)

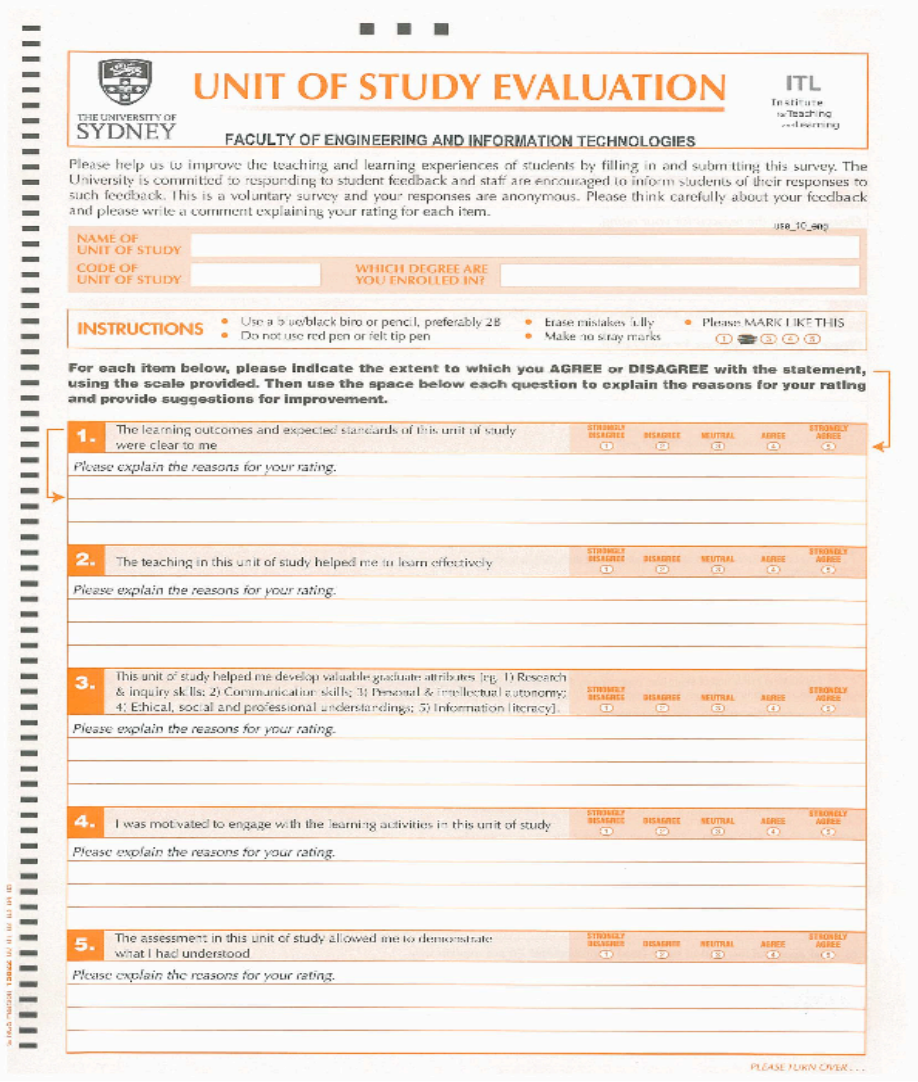
The Unit of Study Evaluations (USEs) is a survey instrument used in Australia to assess students' experience of a course, similar to the Student Evaluations of Teaching (SET) in the USA. The USE questionnaire has 12 questions, 8 of which are standardized University-wide and 4 that are selected by each 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, as monitored by the CEQ. The USE in our study contains 12 statements:

1. The learning outcomes and expected standards of this unit of study were clear to me.
2. The teaching in this unit of study helped me to learn effectively.
3. This unit of study helped me develop valuable graduate attributes.
4. The workload in this unit of study was too high.
5. The assessment in this unit of study allowed me to demonstrate what I had understood.
6. I can see the relevance of this unit of study to my degree.
7. It was clear to me that the staff in this unit of study were responsive to student feedback.
8. My prior learning adequately prepared me to do this unit of study.
9. The learning and teaching interaction helped me to learn in this unit of study.
10. My learning of this unit of study was supported by the faculty infrastructure.
11. I could understand the teaching staff clearly when they explained.
12. Overall I was satisfied with the quality of this unit of study.

Eleven items (I1-I11) focus on students' experience and one item (I12) on student satisfaction. Students indicate the extent of their agreement with each statement based on a 5 - point Likert scale: 1 - strongly disagree, 2 - disagree, 3 - neutral, 4 - agree and 5 -

strongly agree. Below each statement there is a space requesting students to explain their response. Question 4 has a different sentiment structure therefore was removed in this study. Figure 5.4 shows a screenshot of the USE survey.

The USEs of subjects taught by two academics collected over a period of six years were used to create the dataset. After removing responses to question 4, the dataset contains a total of 909 questionnaires (each with 11 ratings), and out of a 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. '?').



UNIT OF STUDY EVALUATION

THE UNIVERSITY OF SYDNEY

ITL Institute for Teaching and Learning

FACULTY OF ENGINEERING AND INFORMATION TECHNOLOGIES

Please help us to improve the teaching and learning experiences of students by filling in and submitting this survey. The University is committed to responding to student feedback and staff are encouraged to inform students of their responses to such feedback. This is a voluntary survey and your responses are anonymous. Please think carefully about your feedback and please write a comment explaining your rating for each item.

use_10_eng

NAME OF UNIT OF STUDY

CODE OF UNIT OF STUDY

WHICH DEGREE ARE YOU ENROLLED IN?

INSTRUCTIONS
- Use a blue/black biro or pencil, preferably 2B
- Do not use red pen or felt tip pen
- Erase mistakes fully
- Make no stray marks
- Please MARK LIKE THIS

For each item below, please indicate the extent to which you AGREE or DISAGREE with the statement, using the scale provided. Then use the space below each question to explain the reasons for your rating and provide suggestions for improvement.

1. The learning outcomes and expected standards of this unit of study were clear to me

Please explain the reasons for your rating.

2. The teaching in this unit of study helped me to learn effectively

Please explain the reasons for your rating.

3. This unit of study helped me develop valuable graduate attributes [eg. 1) Research & inquiry skills; 2) Communication skills; 3) Personal & intellectual autonomy; 4; Ethical, social and professional understandings; 5) Information literacy].

Please explain the reasons for your rating.

4. I was motivated to engage with the learning activities in this unit of study

Please explain the reasons for your rating.

5. The assessment in this unit of study allowed me to demonstrate what I had understood

Please explain the reasons for your rating.

PLEASE TURN OVER . . .

Figure 5.4: Unit of Study Evaluation

The textual data have two characteristics that may significantly affect the classifiers. First the sentences are hand-written in an informal style, containing spelling errors, abbreviated non-dictionary words or hard to read text. The lack of proper grammar would make it extremely challenging to use part-of-speech (POS) tagging or other computational linguistic approaches. Examples include: "Computers in labs too slowk no lecture notes" (spelling mistakes and non-grammar), "tutes were overcrowded, stopping teacher / student interaction" (non-standard words). For these reasons, the techniques used in the experiment are based on the bag-of-words assumption (so word order is not used) and we do not use POS tagging that would require relatively correct grammar.

Rating	Number	Sentiment	Number	Comments tagged with each sentiment
Strongly Agree	381	Positive	1,455	lecturer and tutor was helpful and explained concepts well.
Agree	1,074			
Neutral	611	Neutral	611	It is a bit clear about staff response but need more examples in there answer.
Disagree	571	Negative	874	Not enough computers to accommodate all the students.
Strongly Disagree	303			

Table 5.4: Number of comments and sample comments for each sentiment

Five emotion categories are utilised (*Anger*, *Fear*, *Joy*, *Sadness*, and *Surprise*) in which *Joy* and *Surprise* emotions are assigned to *positive* class while *Anger*, *Fear*, and *Sadness* are members of *negative* class. Fine-grained emotion labels, in contrast to *positive*/*negative* labels, would increase the effectiveness of sentiment classifiers (C. Strapparava & R. Mihalcea, 2007). Negative emotion, *disgust*, is removed because the emotion is similar to *anger* and leads to making sentiment classes biased. Likewise, *strongly agree* and *agree* belong to *positive*, and *strongly disagree* and *disagree* are referred to *negative*. The number of sentences for each rating and sentiment used in our experiment is shown in Table 5.4. In addition, sample comments of the annotated corpus appear in the same table.

5.3 Evaluation of Unsupervised Emotion Models to Textual Affect Recognition

The goal of affect classification is to predict a single emotional label given an input sentence. Four different approaches were implemented in Matlab: a categorical model based on a VSM with dimensionality reduction variants, (LSA, PLSA, and NMF), and a dimensional model. Two similarity measures (cosine angle and nearest neighbour) were used for the evaluation of the two models, respectively. Stopwords were removed in all approaches. A Matlab toolkit (Zeimpekis & Gallopoulos, 2006) was used to generate the term-by-sentence matrix from the text.

Both features go through pre-processing steps: stopwords listing and stemming. These steps help the significant linguistic components of a text to be focused and considered by removing unimportant features. Most languages are full of structural words that provide little meaning to the text. In the following, the explanation of each classification method proceeds in more detail.

The evaluation in Table 5.5 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* 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. The following acronyms are defined in order to identify the approaches:

- CLSA: LSA-based categorical classification
- CPLSA: PLSA-based categorical classification
- CNMF: NMF-based categorical classification
- DIM: Dimension-based estimation

Data set		SemEval			ISEAR			Fairy tales		
Emotion		Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
Anger	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

Table 5.5: Emotion identification results

The measures of accuracies used here were: Cohen's Kappa, average precision, recall, and F-measure. While kappa scores are useful in obtaining an overview of the reliability of the various classification approaches, they do not provide any insight into accuracy at the category level for which precision, recall, and F-measure are necessary.

Table 5.5 shows the values obtained by five approaches for the automatic classification of four emotions. The highest results for a given type of scoring and datasets are marked in bold for each individual class. We do not include accuracy values in our results due to the imbalanced proportions of categories (see Table 5.2). The

accuracy metric does not provide adequate information, whereas precision, recall, and F-measure can effectively evaluate the classification performance with respect to imbalanced datasets (He & Garcia, 2009).

As can be seen from the table, the performances of each approach hinge on each dataset and emotion category, respectively. In the case of the SemEval dataset, 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* category. Figure 5.5 indicates the results of 3-dimensional and 2-dimensional attribute evaluations for ISEAR. 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 fact that 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*, *fear*, and *sadness* in terms of the F-measure.

Figure 5.6 and Table 5.6 display results of different approaches obtained on the three different datasets. The classification performance is computed by macro-average, which gives equal weight to every category regardless of how many sentences are assigned to it. 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. In particular, CPLSA outperforms CLSA and CNMF in ISEAR because their performances are relatively poor. 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 non-negative factors are appropriate for dealing with these text collections.

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 5.3). In addition, the affective high agreement sentence is another plausible contributing reason for the encouraging experimental results. In summary, of those evaluated the categorical NMF model and dimensional model show the best emotion identification accuracy as a whole.

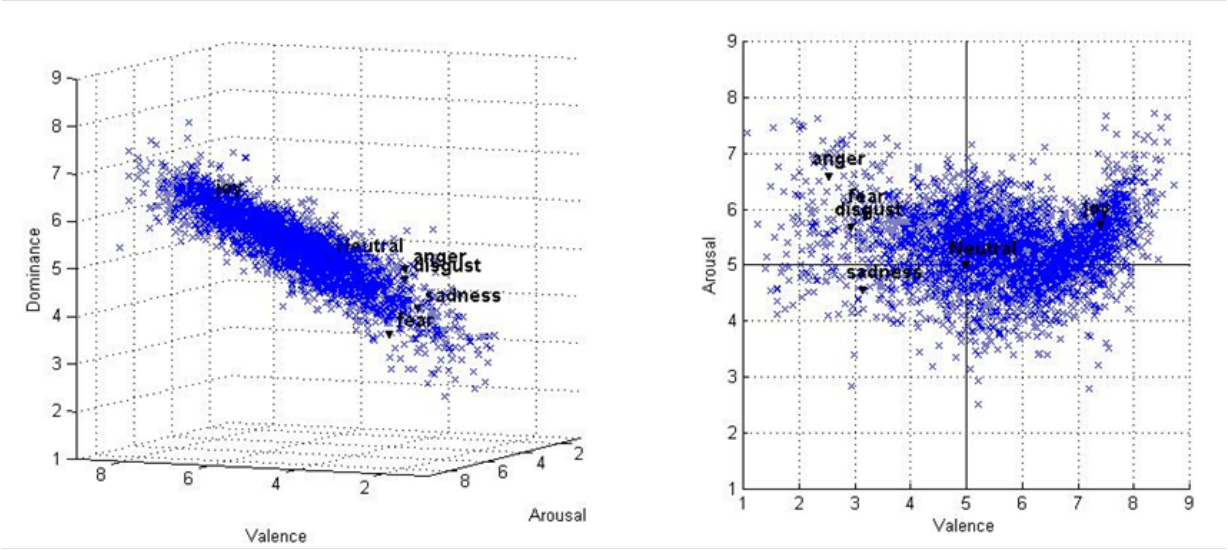
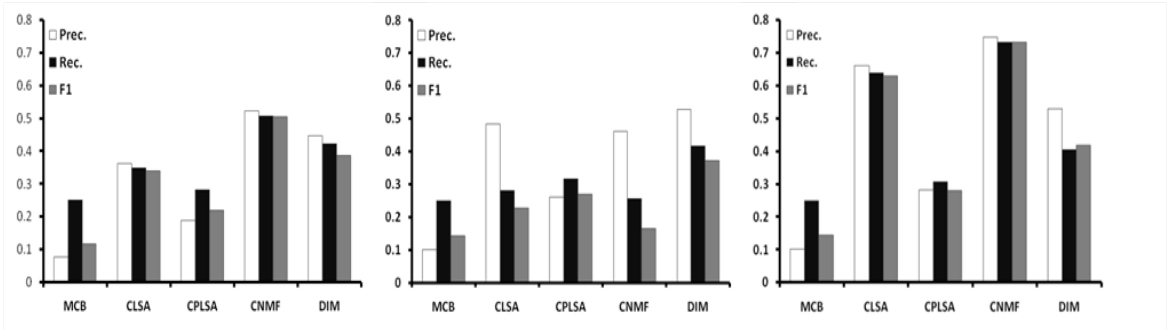


Figure 5.5: Distribution of the ISEAR dataset in the 3-dimensional and 2-dimensional sentiment space. The blue 'x' denotes the location of one sentence corresponding to valence, arousal, and dominance

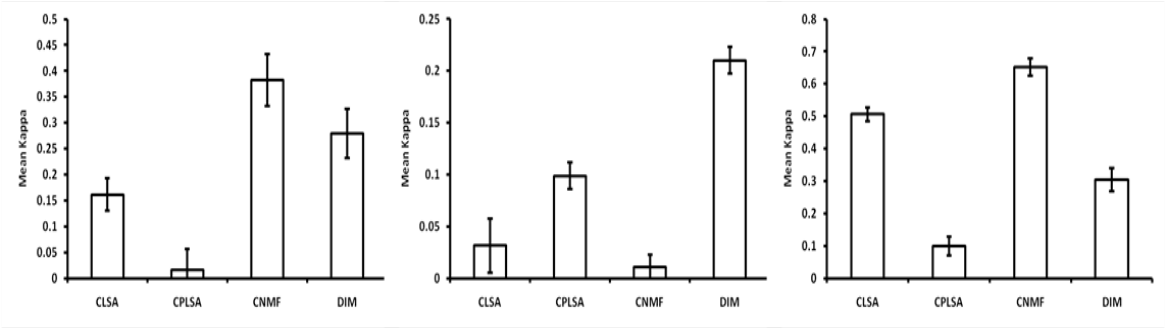


(a) (b) (c)

Figure 5.6: Comparisons of Precision, Recall, and F-measure: (a) SemEval; (b) ISEAR; (c) Fairy tales

Data set	SemEval			ISEAR			Fairy tales		
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
MCB	0.077	0.250	0.118	0.100	0.250	0.143	0.102	0.250	0.145
CLSA	0.363	0.348	0.340	0.484	0.282	0.228	0.662	0.640	0.630
CPLSA	0.189	0.282	0.219	0.260	0.317	0.270	0.282	0.307	0.280
CNMF	**0.523**	**0.506**	**0.505**	0.461	0.258	0.166	**0.747**	**0.731**	**0.733**
DIM	0.446	0.422	0.386	**0.528**	**0.417**	**0.372**	0.530	0.404	0.419

Table 5.6: Overall average results



(a) (b) (c)

Figure 5.7: Comparisons of Mean Kappa: (a) SemEval; (b) ISEAR; (c) Fairy tales

Figure 5.7 graphically depicts the mean kappa scores and its standard errors obtained from the emotion classification. Comparisons between four approaches are shown across all three datasets. MCB is excluded in the comparison because the mean kappa score of MCB is 0.

Let MK_{CLSA}, MK_{CPLSA}, MK_{CNMF}, and MK_{DIM} be the mean kappa scores of four methods. The highest score (MK_{CNMF} = 0.382) is achieved by the CNMF when the dataset is SemEval. In fairy tales, the CNMF method (MK_{CNMF} = 0.652) also displays better results than the others (MK_{CLSA} = 0.506, MK_{DIM} = 0.304). On the contrary, the achieved results are significantly different in the case of the ISEAR dataset in comparison with the aforementioned datasets. The DIM (MK_{DIM}= 0.210) clearly outperforms all methods. The kappa score of the CPLSA approach (MK_{CPLSA} = 0.099) is quantitatively and significantly higher than the CLSA (MK_{CLSA} = 0.031) and CNMF (MK_{CNMF} = 0.011). Kappa score for the NMF-based methods is remarkably lower than the other three approaches.

According to Fleiss and Cohen (1973), a kappa value higher than 0.4 means a fair to good level of agreement beyond chance alone and it is an acceptable level of agreement. On the basis of this definition, the kappa score obtained by our best classifier (MK_{CNMF} = 0.652) would be reasonable. Most of the values are too low to say that two raters (human judges and computer approaches) agreed upon the affective states. However, we have another reason to have confidence in this metric in the experiment. We make use of the kappa score as an unbiased metric of the reliability for comparing four methods. In other words, these measures are of importance in terms of relative magnitude. Hence, the kappa results are meaningful and interpretable in spite of low values. We can observe that the NMF-based categorical model and the dimensional model both experienced higher performance.

5.3.1 Comparison with other systems

The "Affective Text" task in SemEval 2007 is a kind of competition focusing on the classification of emotions and valence in news headlines. In this task, three systems took part in automatically annotating emotions. More details are described in Strapparava and Mihalcea (2007).

Table 5.7 presents the comparison results between our systems and the systems participating in the SemEval task. SWAT is a supervised learning system that used a unigram model. The system also utilised synonym expansion on the emotion-labelled words with resources (Roget's Thesaurus and additional headlines manually annotated). UA made use of statistics such as frequency and co-occurrence collected from search engines. Emotion scores are calculated by using Pointwise Mutual Information (PMI). The system is based on the hypothesis that words which tend to co-occur together across a bunch of documents with a given emotion are highly likely to express the same emotion. Lastly, UPAR7 is a linguistic rule-based system. The system evaluates emotion on all words of a news headline by using SentiWordNet (Esuli & Sebastiani, 2006) and WordNet-Affect lexical resources. UPAR7 also used a Stanford syntactic parser to find the important word.

Emotion	System	Prec.	Rec.	F1
Anger	SWAT	12.00	5.00	7.06
	UA	12.74	**21.60**	**16.03**
	UPAR7	**16.67**	1.66	3.02
	NMF	8.33	6.25	7.14
	DIM	15.79	15.00	15.38
Disgust	SWAT	0.00	0.00	-
	UA	0.00	0.00	-
	UPAR7	0.00	0.00	-
	NMF	**20.00**	**12.50**	**15.38**
	DIM	0.00	0.00	-
Fear	SWAT	25.00	14.40	18.27
	UA	16.23	26.27	20.06
	UPAR7	33.33	2.54	4.72
	NMF	**49.41**	**68.85**	**57.53**
	DIM	38.78	28.79	33.04
Joy	SWAT	35.41	9.44	14.91
	UA	40.00	2.22	4.21
	UPAR7	54.54	6.66	11.87
	NMF	**61.82**	49.28	54.84
	DIM	48.63	**78.89**	**60.17**
Sadness	SWAT	32.50	11.92	17.44
	UA	25.00	0.91	1.76
	UPAR7	48.97	22.02	30.38
	NMF	50.00	**38.71**	**43.64**
	DIM	**64.71**	12.09	20.37
Surprise	SWAT	11.86	10.93	11.78
	UA	**13.70**	**16.56**	**15.00**
	UPAR7	12.12	1.25	2.27
	NMF	3.57	5.88	4.44
	DIM	1.59	4.00	2.27

Table 5.7: Comparison results with other systems

As can be seen from the table, our systems are remarkably superior to the other systems (SWAT, UA, and UPAR7) for *Fear* (n=124), *Joy* (n=148), and *Sadness* (n=145). The three categories have a relatively large number of sentences. In contrast, UA gives the better result with respect to *Anger* (n=41) and *Surprise* (n=50), which make up a small portion of emotion categories. A notable feature is shown regarding *Disgust*. In

general, it is hard to detect *Disgust* and to distinguish *Anger* from *Disgust* since these two emotions belong in the same class. One of our systems (NMF) is able to capture Disgust emotion and this is an advantage of NMF.

The result table strongly indicates that emotion detection is a challenging and complex task in text. The participating systems in SemEval 2007 have common limitations in that they deal with only the importance of the content words of headlines (the lexical approach). They took advantage of synonyms, part-of-speech tags, and word frequency count information as their critical features. However, the emotion expression of text does not rely on each word in the sentence. Positive sentences can have negative meaning words, and vice versa. Moreover, text is able to represent emotions even though all words composing the text have no-emotion (*Neutral*). Likewise, a text might not express any emotion even though each word has its own emotion. For example, in "Surprise: China cuts price of gas", newspaper readers might not feel surprised although there is a 'surprise' in the headline. This headline means that the fact is surprising. As mentioned earlier, NMF and DIM are excellent approaches that can make an inference of hidden emotions from text. They do not depend on the presence of each word, whereas they capture the mutual implications of words.

5.3.2 Frequently occurring words

The most frequent words used in fairy tales for each emotion are listed in Table 5.8. This dataset is chosen since there are varying lexical items and affective high agreement sentences, as mentioned before. Stemming is not used because it might hide important differences as between '*loving*' and '*loved*'. CNMF and DIM were selected for the comparison with the Gold Standard because they were the two methods with a better performance than the others. Gold Standard is the annotated dataset by human raters for the evaluation of algorithm performance. The words most frequently used to describe anger across all methods include: *cried*, *great*, *tears*, *king*, *thought*, and *eyes*. Those used to describe fear include: *heart*, *cried*, *mother*, *thought*, *man*, and *good*. Joy contains *happy*, *good*, and *cried* whereas sadness has only *cried* for three methods.

There is something unexpected about the word frequencies. We can observe that the association between frequently used words and emotion categories is unusual and even opposite. For instance, '*joy*' is one of the most frequent words referred to for *sadness* in the Gold Standard (GS). In CNMF and DIM, '*good*' is employed frequently with regard to *fear*. Moreover, some words occur with the same frequency in various categories. For example, the word '*cried*' is utilised to express *anger*, *fear*, and *joy* in the Gold Standard, CNMF, and DIM. In order to find a possible explanation in the complexity of language used in emotional expression, some sentences extracted from fairy tales are listed 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*

From these examples, we can observe that the affective meaning is not simply propagated from the lexicon, but is the effect of the linguistic structure at a higher level.

Model	Emotion	Top 10 words
GS	Anger	king, thought, eyes, great, cried, looked, joy, mother, wife, tears
	Fear	great, cried, good, happy, thought, man, heart, poor, child, mother
	Joy	thought, mother, good, cried, man, day, wept, beautiful, back, happy
	Sadness	cried, fell, father, mother, back, joy, dead, danced, wife, tears
CNMF	Anger	great, cried, eyes, mother, poor, joy, king, heart, thought, tears
	Fear	cried, king, happy, good, man, heart, thought, father, boy, mother
	Joy	mother, thought, cried, king, day, great, home, joy, good, child
	Sadness	thought, cried, good, great, looked, mother, man, time, king, heart
DIM	Anger	eyes, fell, heart, tears, cried, good, stood, great, king, thought
	Fear	king, cried, heart, mother, good, thought, looked, man, child, time
	Joy	eyes, man, children, danced, cried, good, time, happy, great, wedding
	Sadness	cried, thought, great, king, good, happy, sat, home, joy, found

Table 5.8: Most frequent 10 words from fairy tales

5.4 Sentiment Analysis in Student Experiences of Learning

The first goal of this study is to evaluate the feasibility of using sentiment analysis to study textual responses in USE, an aspect of the data normally sidelined by the ratings. The second goal is to evaluate the merits of using two conceptualizations of emotions (*categorical model* and *dimensional model*) on this data.

The following five different approaches are implemented on Matlab, namely one categorical model that has two variants according to two corresponding methods of dimension reduction, one dimensional method, and two similarity comparison methods for each model which are implemented similarly to the first experiment. However, Majority Class Baseline (MCB) and Keyword Spotting Baseline (KSB) are employed as our two baselines in this experiment for the purpose of evaluation. We get rid of stop words and make use of stemming. Text to Matrix Generator (TMG), a Matlab toolkit, is used to generate term-by-sentence Matrix.

- Majority Class Baseline (MCB): classification that always predicts the majority class, which in this dataset is *Positive* across all sentiment classifications.
- Keyword Spotting Baseline (KSB): a naïve approach that counts the presence of obvious affect words like “frustrating” and “satisfaction”, which are extracted from WordNet-Affect for five emotion categories.
- CLSA: LSA-based categorical classification
- CNMF: NMF-based categorical classification
- DIM: Dimension-based estimation

Table 5.9 shows the precision, recall, and F-measure values obtained by the five approaches for the automatic classification of three sentiments. The highest results are marked in bold for each individual class. Accuracy measures are not included in our results because of the imbalanced categories (see Table 5.4).

Sentiment	Positive			Negative			Neutral		
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
MCB	0.495	**1.000**	**0.662**	0.000	0.000	-	0.000	0.000	-
KSB	0.527	0.220	0.310	0.270	0.061	0.099	0.212	**0.743**	**0.330**
CLSA	0.575	0.362	0.445	0.388	0.203	0.266	0.218	0.560	0.314
CNMF	0.505	0.897	0.646	0.378	0.120	0.182	**0.421**	0.052	0.093
DIM	**0.591**	0.329	0.423	**0.398**	**0.317**	**0.353**	0.223	0.522	0.312

Table 5.9: Sentiment identification results

As can be seen from the table, the performances of each approach depend on each sentiment category. In the case of the *positive* class, which has the largest number of sentences, MCB and CNMF get the best sentiment detection performance in terms of recall and F-measure. DIM achieves a rather high precision score in comparison with all other classifications. We can see that the DIM approach gives the best results for *negative* class. When it comes to *neutral*, KSB shows the best performance with respect to recall and F-measure. On the other hand, CNMF particularly outperforms the others for precision. Figure 5.8 indicates a result of the 3-dimensional and 2-dimensional attribute evaluation for USEs.

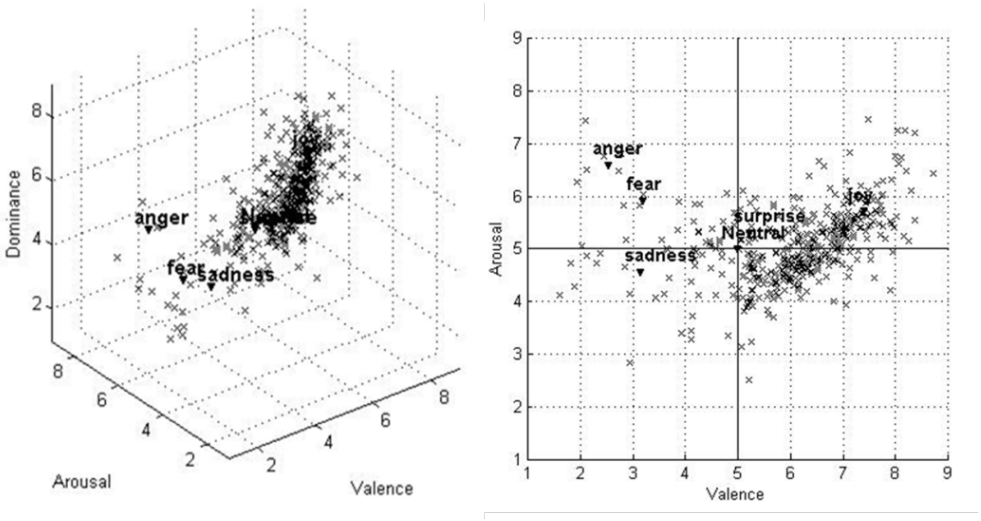


Figure 5.8: Distribution of the USEs dataset in the 3-dimensional (left) and 2-dimensional (right) sentiment space. The 'x' denotes the location of one comment corresponding to valence, arousal, and dominance.

A notable aspect observed in the USE data is that there are some inconsistencies between students' ratings and written responses, illustrated with examples in Table 5.10. For instance, the third row is unambiguously negative but the student graded this sentence as neutral. Therefore, all approaches have a weakness in recognizing sentiments due to the peculiarity of these data. Another factor, which makes the automatic classification difficult, is that all classifiers are not specific to education domains. For this reason, we speculate that this mediocre performance of the methods is owing to poor coverage of the features found in education domains.

Student's feedback	Student rating	System rating
It should be core to software gingerbeering	Positive (5)	Neutral (LSA)
The labs were not long enough with too few tutorials. 4 labs were too few. How about one for FETS/MOSFETS? Given the instruction was for AC/DC components (i.e. lower/uppercase) it was difficult to follow the hadn written notes on the overhead. Maybe print it all up?	Positive (4)	Negative (NMF)
We never got personal feedback.	Neutral (3)	Negative (DIM)
Hi my name is ABC, I like this LECTURER_NAME, I mean this course!!	Negative (2)	Positive (NMF)

Table 5.10: Sample feedbacks from misclassified results. (Positive values are those rates 4 as 5, neutral as 3 and negative 1 or 2)

Table 5.11 shows overall precision, recall, and F-measure comparison with respect to MCB, KSB, CLSA, CNMF, and DIM in two averaging perspectives (micro-averaging and macro-averaging). If micro-averaging is used, a weight is placed on the performance on the larger categories. On the other hand, a weight is placed on the smaller ones if macro-averaging is used. The notable difference between calculating these is that micro-averaging gives equal weight to every sentence whereas macro-averaging weights equally all the categories. Therefore, micro-averaging results in a loss of information. The results of macro-averaging are usually broken into groupings based on the size of categories, so that overall results are not skewed. From this summarized table, we can see that MCB, KSB, and CLSA perform less effectively with a small low number of evaluation scores compared with CNMF and DIM. In case of macro-averaging, CNMF is

superior to other classifications in precision, while DIM surpasses the others in recall and F-measure. On the other hand, DIM has the best precision and CNMF performs better for F-measure in micro averaging. Overall, CNMF and DIM vie with each other in precision, recall and F-measure and the best F-measure is obtained with the approach based on CNMF or DIM for each average. Our KSB conducted in all experiments is inferior to CNMF, DIM as well as CLSA. The result implies that keyword spotting techniques cannot handle sentences which evoke strong emotions through underlying meaning rather than affect keywords. In addition, we can infer that the models (CNMF and DIM) with non-negative factors are appropriate for dealing with text collections. In summary, the NMF-based categorical model and the dimensional model show better sentiment recognition performance as a whole.

The most frequent words used by students to describe aspects of their experience, include terms such as *labs*, *lecturer*, *lectures*, *students*, *tutors*, *subject*, and *work*. When these terms are removed, the words most frequently used to describe positive experiences include: *good* (*n*=263), *helpful* and *helped* (*n*=183), *online* (*n*=79), *understand* (*n*=49). Those used to describe negative experiences include: *hard* (*n*=72), *understand* (*n*=67), *time* (*n*=47). Neutral experiences contain a combination of both. These words lists are obtained from CNMF and DIM because these two classifications have better overall performance as aforementioned. Stemming was not used for this analysis since in this particular corpus it might hide important differences as between ‘lecturer’ and ‘lecture’.

Mean	Micro			Macro		
	Prec.	Rec.	F1	Prec.	Rec.	F1
MCB	0.245	**0.495**	0.328	0.165	0.333	0.221
KSB	0.385	0.281	0.325	0.337	0.341	0.247
CLSA	0.445	0.356	0.396	0.394	0.375	0.342
CNMF	0.450	0.490	**0.469**	**0.434**	0.356	0.307
DIM	**0.457**	0.366	0.406	0.404	**0.389**	**0.363**

Table 5.11: Overall average results

5.5 Summary

In this chapter we presented performance measurements of two experiments: one for emotion detection, evaluated with three datasets (SemEval, ISEAR, Fairy Tales), and another for sentiment detection, evaluated with USE.

In the first experiment we compared the performance of three techniques based on the categorical representation of emotion, and one based on the dimensional representation. Our results show that the NMF-based categorical classification performs the best among categorical approaches to classification. When comparing categorical classification with dimensional estimation, the categorical NMF model and the dimensional model have better performances. Nevertheless, we cannot generalise inferences on which of these techniques is the best performer since results vary among datasets.

The second experiment utilised a dataset of ratings and textual responses of student evaluations of teaching. Sentiment analysis techniques for automatically rating textual responses as *positive*, *negative* or *neutral* using the students' ratings were evaluated. In particular, the performances of the categorical model and the dimensional model were compared, each of which makes use of different linguistic resources. The results highlighted that NMF-based categorical and dimensional models have a better performance than the others. Moreover, despite not having an appropriate set of emotional categories to use, the efficacy of two emotion lexicons (WordNet-Affect and ANEW) promises to be useful in these sentiment classification tasks.

CHAPTER 6
Conclusion

"*Emotions are phases of an individual's intuitive appraisals either of his own organismic states and urges to act or of the succession of environmental situations in which he finds himself.*"

- John Bowlby, 1969 (Kleinginna & Kleinginna, 1981).

We contribute a new computational modelling of emotions approach 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 a way of visualising 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's assumption used in much of the literature, yet they provide good performances 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.

While two approaches (categorical and dimensional) and two lexicons (WordNet-Affect and ANEW) are promising for identifying emotions or sentiments, there are still challenges to overcome. We believe that affective expressivity of text rests on the basis of more complex linguistic features. The common bag-of-words assumption used in this thesis and 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. Therefore, our approaches based on emotion models have the lexical limitations. In the future, it would be worth studying techniques that combine NLP techniques with the use of normative databases.

Another limitation of the first experiment is that results depend on datasets. As a future work, we aim at performing a further investigation of this connection in order to identify more effective strategies applicable to a generic dataset. For the second experiment future work will include extending the corpora with more student evaluations and this should provide more reliable results. The categorical model should be evaluated with a set of emotion categories better grounded in the educational research literature and we suspect that the literature on motivation would be particularly useful. With regards to the use of normative databases to study the dimensional model, we are aware that the terms in ANEW are not the best suited for the vocabulary that students use to describe their experiences, but we are not aware of other more appropriate databases.

As mentioned earlier, emotions in text can be analysed in different ways. Text can evoke or trigger emotions in those who read it and text can also reflect or express the emotional state of the person writing it. Our approaches do not distinguish amongst them in this thesis. In future research it would be important to discriminate these two perspectives more explicitly.

Appendix A
Detailed Results

Following tables show the detailed results of our experiments. This is not a whole list and the four tables correspond to the four datasets. Each row retrieved from the results by our experiment depicts the emotion categories of gold standard, CNMF, and DIM with respect to the input sentence. Besides, two similarity measures (cosine and distance) are shown in the table.

Sentence	GS	CNMF	Cosine	DIM	Distance
Mortar assault leaves at least 18 dead	fear/ sadness	sadness	0.90940114	fear	1.4192303
Goal delight for Sheva	joy	fear	0.964934899	joy	0.996839717
Nigeria hostage feared dead is freed	fear/joy	sadness	0.941142697	fear	1.237688735
PM: Havana deal a good experiment	Neutral	joy	0.829772368	joy	0.368111053
Happy birthday, iPod	joy	Neutral	0.602429244	joy	1.065104495
United Finds Good Connection in Win	joy	joy	0.803427669	joy	0.286583735
We were 'arrogant and stupid' over Iraq, says US diplomat	sadness	fear	0.907557433	sadness	0.644272712
Moderate drinking reduces men's heart attack risk	joy	joy	0.965055882	joy	0.940879762
'House of Cards' actor Ian Richardson dead	sadness	sadness	0.670393105	Neutral	0.687477272
Salmonella outbreak traced to peanut butter	fear	sadness	0.892490406	Neutral	1.888570888
UPDATE 1-Plane crashes at Moscow airport, no passengers	sadness	sadness	0.690151338	fear	1.416728094
Venezuela, Iran fight U.S. dominance	anger	fear	0.919604849	anger	1.348270245
Hurricane Paul Weakens To Tropical Storm	Neutral	Neutral	0.34945105	fear	1.023394254
UK announces immigration restrictions	Neutral	anger	0.660302049	Neutral	6.835994457

Table A.1: Detailed results from SemEval 2007

Sentence	GS	CNMF	Cosine	DIM	Distance
When I was involved in a traffic accident.	fear	fear	0.935825075	fear	1.213506216
When I was driving home after several days of hard work, there was a motorist ahead of me who was driving at 50 km/hour and refused, despite his low speed to let me overtake.	anger	fear	0.950205376	joy	1.424353463
When I lost the person who meant the most to me.	sadness	joy	0.968344885	Neutral	0.99295015
When I got a letter offering me the Summer job that I had applied for.	joy	joy	0.94925549	joy	1.235524194
When I was going home alone one night in Paris and a man came up behind me and asked me if I was not afraid to be out alone so late at night.	fear	fear	0.935865721	Neutral	0.675943785
When I was talking to HIM at a party for the first time in a long while and a friend came and interrupted us and HE left.	anger	joy	0.791997159	joy	0.636399194
On days when I feel close to my partner and other friends. When I feel at peace with myself and also experience a close contact with people whom I regard greatly.	joy	Neutral	0.618102613	joy	0.876540865
Every time I imagine that someone I love or I could contact a serious illness, even death.	fear	sadness	0.99264235	Neutral	0.302211846
When I had been obviously unjustly treated and had no possibility of elucidating this.	anger	anger	0.942894537	joy	0.437130757
When, for the first time I realized the meaning of death.	fear	sadness	0.936912321	sadness	0.32166992
When a car is overtaking another and I am forced to drive off the road.	anger	fear	0.935363025	joy	0.9895101
When I recently thought about the hard work it takes to study, and how one wants to try something else. When I read a theoretical book in English that I did not understand.	sadness	anger	0.771770109	Neutral	1.046231332

Table A.2: Detailed results from ISEAR

Sentence	GS	CNMF	Cosine	DIM	Distance
Then he got up and clambered out of the cave, went into the forest, and thought, Here I am quite alone and deserted, how shall I obtain a horse now?	sadness	fear	0.971013962	Neutral	0.328785644
Alas, thou canst not help me.	sadness	sadness	0.955495577	Neutral	6.835994457
They leapt nimbly upstairs and downstairs, and were merry and happy.	joy	Neutral	0.567194612	joy	0.915652512
Then she opened the door of the small house, and when she had opened it, there stood twelve horses, such horses, so bright and shining, that his heart rejoiced at the sight of them.	joy	fear	0.958813011	joy	1.415987473
Then they laughed and said, Indeed, stupid Hans, where wilt thou get a horse?	anger	fear	0.948093133	sadness	1.003416118
In the morning when he awoke, the three days had passed, and a coach came with six horses and they shone so bright that it was delightful to see them! and a servant brought a seventh as well, which was for the poor millers boy.	joy	fear	0.779268447	joy	1.19029596
The little tailor did not let himself be frightened away, but was quite delighted, and said, Boldly ventured is half won.	joy	fear	0.970616746	joy	0.996839717
Eh! thought he, what a stupid blockhead I am!	anger	fear	0.948408579	Neutral	0.850367568

When the bear heard the music, he could not help beginning to dance, and when he had danced a while, the thing pleased him so well that he said to the little tailor, Hark you, is the fiddle heavy?	joy	fear	0.959049622	joy	0.857089065
Then a vise was brought, and the bear put his claws in it, and the little tailor screwed it tight, and said, Now wait until I come with the scissors, and he let the bear growl as he liked, and lay down in the corner on a bundle of straw, and fell asleep.	anger	sadness	0.873043381	Neutral	0.934197517
The bear in great fury ran after the carriage.	anger	fear	0.972834328	Neutral	6.835994457
The princess heard him snorting and growling; she was terrified, and she cried, Ah, the bear is behind us and wants to get thee!	fear	sadness	0.962078503	anger	2.482067933
The tailor drove quietly to church, and the princess was married to him at once, and he lived with her as happy as a woodlark.	joy	Neutral	0.527745572	joy	0.435871256
In the evening they came to a large forest, and they were so weary with sorrow and hunger and the long walk, that they lay down in a hollow tree and fell asleep.	sadness	sadness	0.958897717	Neutral	1.460428019
And now the sister wept over her poor bewitched brother, and the little roe wept also, and sat sorrowfully near to her.	sadness	sadness	0.710313287	joy	1.512028517
The little sister, however, was dreadfully frightened when she saw that her fawn was hurt.	fear	fear	0.967919342	fear	1.333631981

Table A.3: Detailed results from Fairy tales

Sentence	GS	CNMF	Cosine	DIM	Distance
Good teaching and tutorials	Positive	Positive	0.781315	Positive	0.368111
Yes it is.	Positive	Neutral	0	Neutral	6.835994
Good	Positive	Positive	0.773172	Positive	0.368111
helpful tutor but bad lab set-tup	Positive	Negative	0.695358	Neutral	6.835994
Very bad. SVN didn't work, small home dir space practically impossible to do project on lab computers	Negative	Negative	0.820045	Positive	1.332654
A difficulty was in the labs where we had to communicate with the internet across a proxy	Positive	Negative	0.475548	Neutral	6.835994
I could not complete my projet and was therefore unable to demonstrate it	Negative	Negative	0.919477	Neutral	6.835994
emails were answered within 2 days always	Positive	Negative	0.857074	Positive	0.902479
the tutor/lecturer was great but unfortunately the difficulty with the lab computers made the tutorials difficulties	Neutral	Negative	0.657685	Negative	1.297767
can't attend lectures	Neutral	Negative	0.615177	Neutral	6.835994

It's hard to set up working environment in lab	Negative	Negative	0.92094	Negative	0.641015
Spring, Hibernate, Maven are all useful stuff.	Positive	Positive	0.891914	Positive	0.370168
I do not know what the standards is.	Negative	Negative	0.968247	Neutral	6.835994
It's good to thihk how to write my idea in a formal report.	Positive	Positive	0.892435	Positive	0.225405
I think it's okay.	Neutral	Neutral	0	Neutral	6.835994
no enough computer	Negative	Negative	0.447198	Negative	1.297767
I understand the material more on my self-reading	Negative	Negative	0.920327	Negative	0.99222
Cannot always catch a tutor and ask everything, somehow, it's not clear enough of what to do in tutorial. And there are not enough computer to use.	Negative	Negative	0.903884	Negative	1.297767
I don't know what I've done in assignment are what I should have understood.	Neutral	Negative	0.896086	Neutral	6.835994
Lecture is not interactive and not interested. The stuff of lecture was too abstract, example needed to help me understand the course. Overall, I learned lots of stuff in report writing and teamwork, but not the content of this unit	Negative	Negative	0.883063	Negative	1.291482

Table A.4: Detailed results from USE

Appendix B

TMG: A Matlab Toolbox for Generating Term-Document Matrices from Text Collections

Various tasks in text mining are performed with the Matlab toolbox called Text to Matrix Generator (TMG). This can be found at http://scgroup20.ceid.upatras.gr:8000/tmg/. TMG is particularly appropriate for text mining applications where data is high-dimensional but extremely sparse. The original version of TMG was built as a pre-processing tool for creating term-document matrices from unstructured text. The current version provides a wide range of tools such as Dimensionality Reduction, Clustering, and Classification.

In our experiment, TMG is used for the purpose of the pre-processing step (indexing) and we will only mention this step-related tool. The pre-processing constructs new term-document matrices from documents. This toolbox enables us to automatically perform the text parsing, stopwords removal, stemming, term weighting and normalisation.

The followings are brief descriptions with respect to optional parameters. There are two types of frequency restrictions defined as local frequencies and global frequencies. The former means the frequency of a given word inside a single document whereas the latter represents the frequency of a given word inside every document. The local and global term weighting schema evaluates the statistical importance of a given word to a document and also related to a collection of documents.

- delimiter: The delimiter for textual parsing, particularly, between sentences within the same file in our experiment.
- line_delimiter: The delimiter for a whole line of text.

- stoplist: The filename containing stop words which are common and unnecessary words.
- stemming: The indicator for stemming or not.
- min_length: The minimum length of a term.
- max_length: The maximum length of a term.
- min_local_freq: The minimum local frequency for a term.
- max_local_freq: The maximum local frequency for a term.
- min_global_freq: The minimum global frequency for a term.
- max_global_freq: The maximum global frequency for a term.
- local_weight: The local term weighting schema.
- global_weight: The global term weighting schema.
- normalisation: The indicator for normalising the document vectors or not.
- dsp: The indicator for displaying results or not on the command window.

The optional parameters and values used for the pre-processing in our experiment are tabulated in Table B.1.

Option parameter	Value
delimiter	@ (default 'emptyline')
line_delimiter	0 (default 1)
stoplist	stopwords.txt (default no stoplist used)
stemming	1 (default 0)
min_length	1 (default 3)
max_length	inf (default 30)
min_local_freq	1 (default)
max_local_freq	inf (default)
min_global_freq	1 (default)
max_global_freq	inf (default)
local_weight	l (default 't')
global_weight	e (default 'x')
normalisation	c (default 'x')
dsp	1 (default)

Table B.1: Parameters and values used in TMG toolbox for pre-processing

The possible values of local_weight are listed as follows:

- ‘t’: Term frequency
- ‘b’: Binary
- ‘l’: Logarithm
- ‘a’: Alternate log
- ‘n’: Augmented normalised term frequency

The following lists show the possible values of global_weight.

- ‘x’: None
- ‘e’: Entropy
- ‘f’: Inverse document frequency (IDF)
- ‘g’: GfIdf
- ‘n’: Normal
- ‘p’: Probabilistic inverse

The value (‘c’) of normalisation stands for *Cosine*.

Figure B.1 shows the sequential steps within TMG for pre-processing textual data. The four (TDM, *Dictionary*, *Titles*, and *Files*) objects are utilised in our experiment among the full list of outputs. *Dictionary* is a collection of words that are row names in TDM. *Titles* are the titles of each document and *Files* are processed filenames with set title and document’s first line.

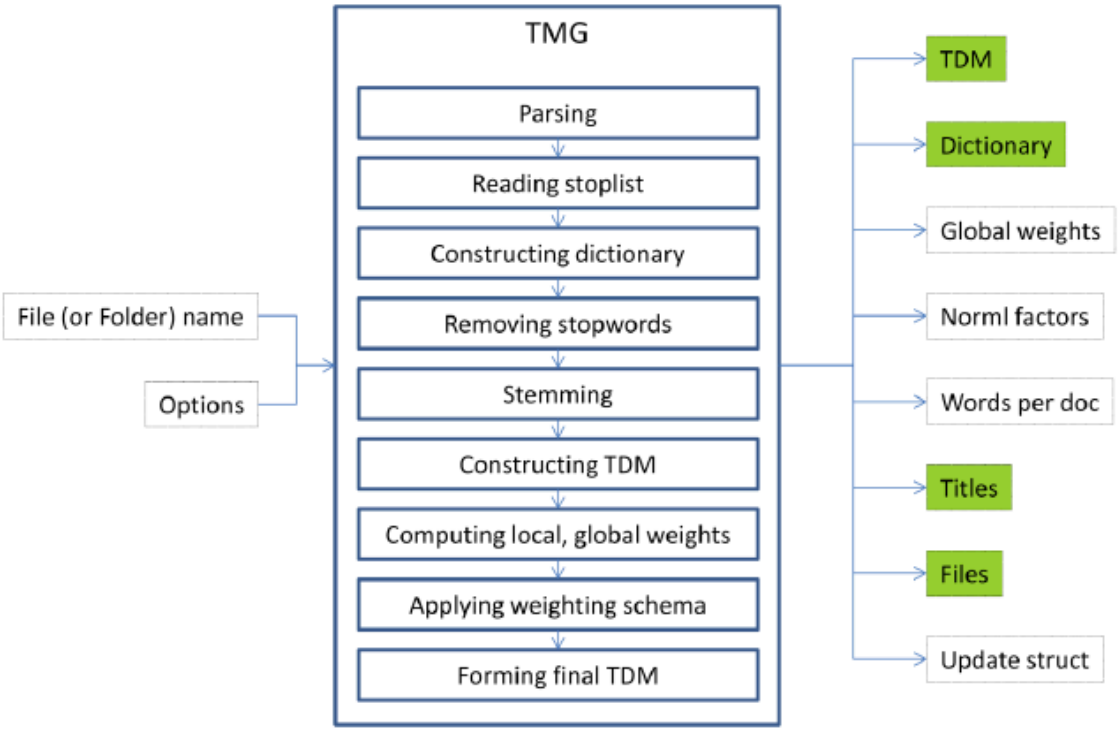


Figure B.1: Processing steps in TMG

After the pre-processing step, a dataset is, as a result, presented as an $m \times n$ matrix, which is the form of a sparse *term-document matrix* of n documents over an index of m terms.

Appendix C
Emotion ML

Through the years, several studies and theories on emotions have emerged. As a consequence, languages are needed to express emotions and (Marc Schröder) made use of their own XML language called EDML (Emotion Dimension Markup Language) for the representation of emotional data. EDML was created in order to annotate text by using a single <emotion> tag in which the positions on emotion dimensions are specified as the values of the activation, evaluation and power attributes. However, this is not a common interchange language for the purpose of dealing with emotion and affect. Namely, a general-purpose emotion markup language is required to overcome the variety of emotion representations. There has been an effort to create a new markup language by W3C Emotion Incubator Group and the group defined an Emotional Markup Language (EmotionML) based on the prioritisation on requirements: mandatory and optional requirements. The latest version is an EmotionML 1.0 (Baggia et al., 2008) and the language is used as a "plug-in" for three different areas: manual annotation of emotion, automatic recognition of emotion, and generation of emotion. EmotionML has structural elements and a hierarchy of children elements and attributes.

In addition to the classification result, we have developed the system that is able to annotate emotions using EmotionML in our experiment. The system produces an XML file, the EmotionML and this file contains the sentences tagged by categorical and dimensional model representations.

We describe the syntax of the main elements of EmotionML which is particularly related to our project. For example, the specification includes *Appraisal Model*, which is another different emotion representation. However, the model is not explained since categorical and dimensional models are exploited in our experiment.

```
<emotionml xmlns="http://www.w3.org/2008/11/emotionml">
    <emotion date="2009-11-23T14:36Z">
        <modality set="basicModalities" mode="text"/>
        <category set="basicEmotions" name="fear"/>
        <link uri="'C:\Users\MATLAB\workspace\dataset\test_AlmFairy/AlmFairy.txt.20
        The little sister, however, was dreadfully frightened when she saw that her fawn was hurt.'"/>
    </emotion>
    <emotion date="2009-11-23T14:36Z">
        <modality set="basicModalities" mode="text"/>
        <dimensions set="valenceArousal">
            <arousal value="0.2"/>
            <valence value="0.7"/>
            <dominance value="0.3"/>
        </dimensions>
        <link uri="'C:\Users\MATLAB\workspace\dataset\test_AlmFairy/AlmFairy.txt.20
        The little sister, however, was dreadfully frightened when she saw that her fawn was hurt.'"/>
    </emotion>
</emotionml>
```

Figure C.1: Emotion-labelled sentence and emotional state annotation tags

An EmotionML document starts with the <emotionml> element, which is the document root. Each single emotion annotation on a sentence is marked with the <emotion> element with respect to each emotion representation. The *date* attribute indicates the absolute date when the annotation occurred. The children elements are as follows:

- The <modality> element: The element used for the annotation of modality such as “voice” or “text”. Only one element is allowed in each <emotion>.
- The <category> element: The element is the emotion representation of categorical model. A *set* attribute means the name of a label set and a *name* attribute is the name of the label in the label set identified by the *set* attribute.
- The <dimensions> element: The element describes an emotion in terms of dimensional emotion representation. The element has three children elements such as *arousal*, *valence*, and *dominance* and the value of each sub-element denotes the scale value of each dimension.
- The <link> element: The element provides reference information about the annotated sentence. The information contains the location, the filename, and the content of the sentence.

Appendix D
SAM and Feeltrace

The dimensional model approach implies a very different rating method, which is based on the assumption that emotions can be described by their degree of each dimension. SAM called for Self Assessment Manikin (see Figure D.1) was devised by (Lang, 1980). SAM consists of pictures of manikins which are rating scales for the dimensions. A non-verbal pictorial assessment enables the emotion annotation to become cross-cultural and language-independent. The emotional states consist of *Pleasure*, *Arousal*, and *Dominance* (PAD, the initials of each emotional state). Synonyms are used for the expression of the PAD dimensions. *Pleasure* can correspond to *Valence* and *Evaluation*, *Arousal* to *Activation* and *Activity*, and *Dominance* to *Power* and *Potency*. Figures representing *Pleasure* range from a widely smiling manikin (pleasant) to a frowning one (unpleasant). The *Arousal* rating is illustrated by figures which range from a very relaxed, eyes-closed manikin to a highly energetic figure, displaying wide-open eyes. Finally, the dominance dimension is depicted by figures of increasing size. A middle stance on three figures shows a *neutral* state. Ratings of each dimension are given in a 9-point format, meaning that participants are free to choose any of the five figures or any of the four intervals in-between them. In the experiment participants are prompted with each of the ratings separately and then indicate their respective point that is written underneath either each manikin or in between two manikins. Several of the widely used affective stimuli database (IAPS, IADS, ANEW), which is expressed as means and standard deviation scores, have been normatively rated regarding these dimensions through SAM.

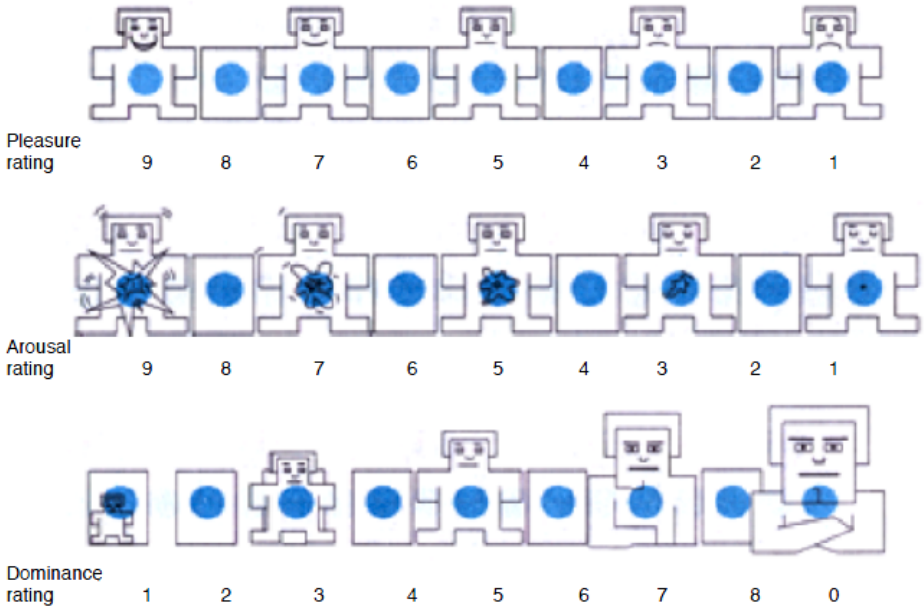


Figure D.1: The SAM tool on three emotion dimensions.

Feeltrace tool (Cowie, Douglas-Cowie, Savvidou*, et al., 2000) provides a new, dynamic method for the assessment of emotion dimension ratings. In contrast to static scale tools like SAM, it allows the tracking of emotional information continuously over time. Feeltrace is able to capture emotional variations in affective states within the time intervals. Figure D.2 presents a graphical representation of two emotion dimensions: evaluation (from *negative* to *positive*) and activation (from *passive* to *active*). Emotions are located by means of coloured cursors and emotion words are displayed at their coordinates in the two dimensional space. The cursor leaves a trail of shrinking circles behind and this gives us the information about recent cursor movement in the evaluation-activation space. This indicates the indirect representation of time dimension. The track of Figure D.2 shows a recent cursor movement from the active-negative to the passive-positive place. In this space, a neutral state is in the centre of the circle, whereas emotions on the circle radius have the maximum intensity values of affective states. Feeltrace is suitable for describing the emotional information of speech or video since it keeps the time stamp. Researchers exploited the Feeltrace for the purpose of annotating speech data such dialogue (Craggs & McGee Wood, 2004; Marc Schröder, 2004). In particular, (M. Schröder, Cowie, Douglas-Cowie, Westerdijk, & Gielen) made use of the Feeltrace tool

to annotate the Belfast Naturalistic Emotion Database with respect to emotional content. The database contains audio and visual recordings of 100 English speakers expressing relatively spontaneous emotion such as TV recordings of chat shows, religious programs, and interviews recorded in a studio. According to (Cowie, Douglas-Cowie, Savvidou, et al., 2000), audio and visual emotional data taken from the database covered the centre and all four quadrants of the activation-evaluation space in the Freeltrace.

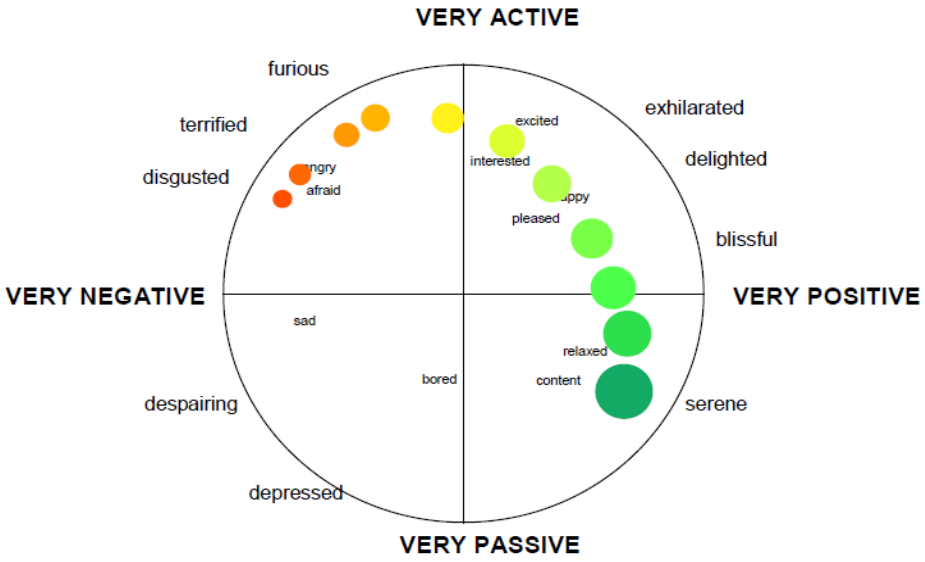


Figure D.2: The Feeltrace tool on two emotion dimensions.

Appendix E

A Categorical Annotation Scheme for Emotions

The corpora annotated for emotion would be a fundamental and valuable resource as a training data set for the purpose of building smart sentiment classifiers. However, annotation is a tedious and time-consuming task, and identifying emotions in text is more difficult, error prone and a highly subjective task. Therefore, it is first necessary to develop an annotation scheme in order to prevent annotators from making these mistakes. There are two different emotion models on the assessment of emotions: one is the categorical model and the other is the dimensional model. A categorical annotation scheme for emotion is a collection of predefined labels for emotions which annotators apply to each individual annotation unit such as a sentence or an utterance. The reason why we choose the former is that this model is more intuitive and easier to implement than the latter model. The goal of this annotation scheme is to provide annotators with some criteria and strategies to use in the annotation process.

Task Overview

What annotators will do is to read the documents of students and to annotate them with information about the "mental state" of the students. This means that you keep trying to answer this question based on what you read "what goes on in this student's mind?" The annotation scheme is designed for the education domain. For this reason, target documents are the essays and feedbacks of students. Emotion labels are selected by

taking into account a pedagogical point of view. Choosing a specific emotion label for a sentence is a difficult task. In brief, the annotation procedure proceeds as follows:

- Annotators read a sentence in provided documents and try to imagine what the emotion states of the student are in each sentence.
- Annotators have to choose one relevant label from a predefined emotion set that best fits the mental state of the student in the sentence best.
- Annotators resume reading the next sentence and the annotation process reiterates.

Annotation Instructions

1. There are no formal and fixed rules about how a particular sentence should be annotated. This document provides the annotations of specific examples, but do not state that a sentence which contains specific words should always be annotated a certain emotion. The annotators are encouraged to follow their first intuition and use their judgment based on the tagging criteria and on the examples of tagged sentences provided in this document.
2. Sentences should be interpreted with respect to the contexts in which they appear. Emotion is tagged from the point of view of the text, that is, the feeler in the sentence. The annotators should not take sentences from context and think what they could mean, but rather should judge them as they are being used in that particular sentence and document.
3. When applying to this scheme, annotators are restricted to applying one emotion label for each sentence. Although it is obviously possible to express more than one emotion per sentence, it is relatively rare. Furthermore, allowing annotators to select more than one emotion would increase the scheme's complexity, incurring greater disagreement and lengthening annotation time. A focus of this work is identifying one emotion state in each context.
4. The annotators should be consistent as they can be with reference to their own annotations and the sample annotations given to them.

Annotation Guidelines

Creating an annotation for each sentence consists of the following elements.

1. Emotion Category (name)

   This is the name of the emotional label or tag and these are defined in detail in the following Section 3. There are five emotion labels that were defined especially for the collaborative process as follows: *neutral*, *anger*, *boredom*, *anxiety* and *excitement*. Generally, the choice of these labels is tailored to the domain.

2. Annotator Certainty (confidence)

   This attribute presents whether the annotator is sure about a given emotion tag, which means the annotator's confidence that the annotation is correct. The annotators use this attribute if they are not sure that a given emotion is present, or if, given the context, there are multiple possible interpretations of the sentence and the annotator is not sure which interpretation is correct. This is an optional value and there is a three-level likert item as follows:

   ++ (high confidence) – "Oh That is wonderful news"
   + (medium confidence) – "The report is full of absurdities"
   0 (low confidence) – "I think people are happy because Chavez has fallen."

3. Emotion Intensity (intensity)

   This attribute indicates the strength of the emotion expressed in subjective sentences and intensity will be a discrete value which is rated on a three-point scale from 1 to 3. A neutral state will be marked as the intensity level 0.

   1 (low intensity) – "I'm not feeling too great"
   2 (medium intensity) – "Oh she's annoying that girl"
   3 (high intensity) – "I can't bear to talk about it"

Annotation Strategy

Two annotators work in pairs on the same documents and they independently read the essays and feedbacks in order to avoid any annotation bias. Each annotator marks the sentence level with one of the five emotions. (Table E.1) Sentences with ambiguous labels will be judged by a third independent annotator in order to decide the final label.

Abbreviation	Emotion Class
NT	Neutral
AG	Anger
BD	Boredom
AX	Anxiety
EX	Excitement

Table E.1: Five emotions used in annotation

Essays and feedbacks are segmented into sentences and nominal labels are applied to every sentence for its emotional expression. The benefit of employing sentences as a basic annotation unit is that labelling at this granularity makes the result of analysis easier to apply because a sentence is a convenient unit for the essays or related feedbacks of students.

Emotion Category Definitions

- *Neutral* – nothing remarkable is happening. (Neutral is used in the place where there is no emotion present in the sentence or where there is no emotion discernable in the sentence.)
  "He's moved to Blackgurn"

- *Anger* (choler, ire, wrath) – a strong feeling of annoyance, displeasure, or hostility. (The students express that a certain situation or person has upset them such that they feel passionately about it.)
  "A close person lied to me"
  "A colleague asked me for some advice and as he did not have enough confidence in me he asked a third person"

- *Boredom* (ennui, tedium) – the state of feeling bored. (The students are clearly bored with something.)

- *Anxiety* (anxiousness) – a feeling of worry, nervousness, or unease, typically about an imminent event or something with an uncertain outcome.
  "I'm sorry, I'm having trouble understanding you. Please try again."

- *Excitement* (exhilaration, inflammation, fervor) – a feeling of great enthusiasm and eagerness.
  "This week has been the best week I've had since I can't remember when! I have been so hyper all week, it's been awesome!!!"

The above synonyms come from WordNet that is a large lexical database of English.

Training

The annotators focus on learning the annotation scheme and then they should know how to create the annotations using the annotation tool, which is UAM Corpus Tool [http://www.wagsoft.com/CorpusTool/index.html]. This chapter will provide you with the opportunity to be familiar with this text annotation tool.

Get the Project Files

The first thing to do is to obtain the predefined project files that include coding schemes, corpus documents, and so forth before you actually start annotating. This task enables annotators to start annotating immediately without annoying prerequisite setup. You have to create the folder where your project files are to be stored and then you may have to copy these files to this folder on your computer.

Start the Annotation Tool

You can start the annotation tool simply by double-clicking the .cptr file in the project folder because you have already obtained the project files above. This file has an icon as below:

Mac OSX

Windows

Begin Annotating

Once you have double-clicked the above icon (the .cptr file) in order to launch your project directly, the CorpusTool Main Window will open, showing the Project Management pane. You can see this in Figure E.1.

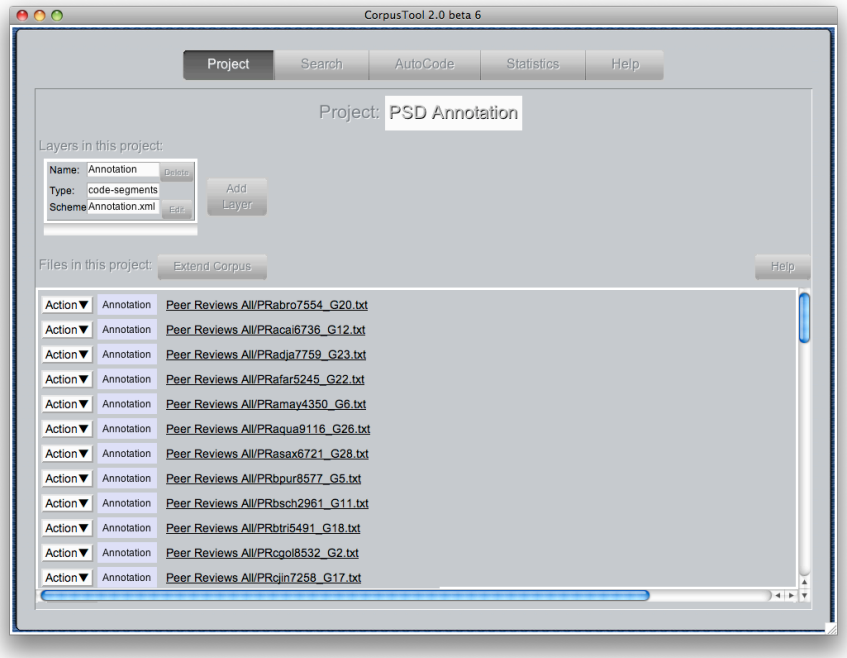


Figure E.1: The Project Management pane

If you click on the "Annotation" button for one of your text files, then a window like that in Figure E.2 will appear. This is an annotation window for the document.

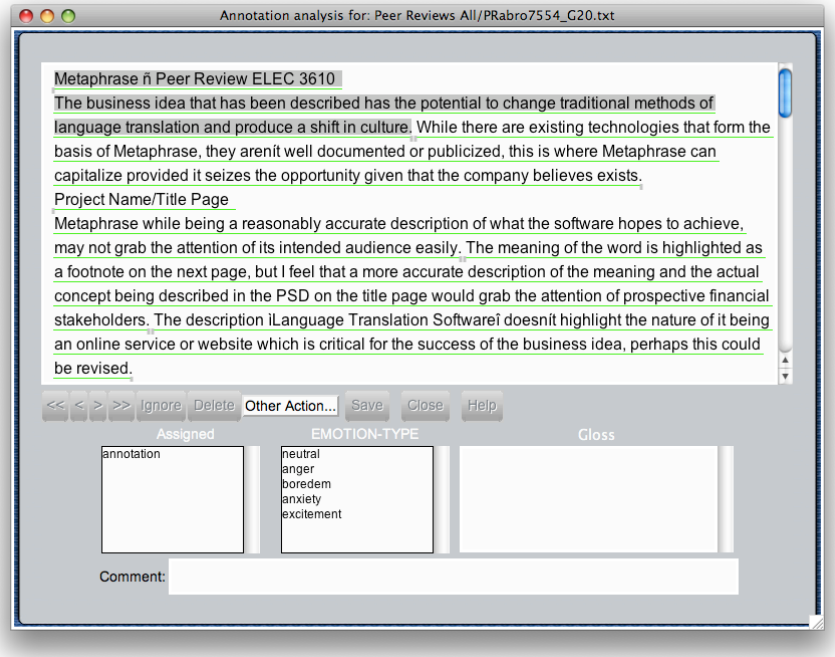


Figure E.2: Annotation Window

The Annotation Window has 4 parts:

1. Text Frame: you can see the whole text in this frame. You can also select a sentence by clicking on the sentence line which runs under each sentence. You can tell which sentence the mouse is over, as the line of the sentence is highlighted.
2. Tool Bar: there are a few buttons such as <<, <, >, >>, Save, Close and Help that you can use.
    a. Select next/previous sentence: you can use the < and > buttons in the toolbar to move around between sentences.
    b. Select next/previous incomplete sentence: you can use the << and >> buttons in the toolbar to move to the next or previous sentence which is not tagged yet.
3. Coding Frame contains three boxes:
    a. Selected Features (labelled 'Assigned'): the features (emotion, intensity and confidence) already assigned to the sentence. Initially, this will contain one feature (annotation), the leftmost ('root') feature of the coding scheme. As other features are assigned, they will appear here. You can delete the features by double-clicking on the features in the Selected Feature box. The root feature cannot be deleted, since it applies by default to all documents.
    b. Current Choice: the middle box is a choice which needs to be made for this sentence. Double-click on one of the options. That choice will be moved to the Selected Feature box. If there are more choices in the coding scheme, the next choice will then be displayed.
    c. Gloss Box: if you click on a feature in the Current box, the gloss will be displayed in this space. This is useful when you have forgotten what exactly the annotating criteria are for this feature or you need to see some examples related to this feature.
4. Comment Frame: you can type comment about the current sentence in this box. This frame records any additional annotator comments concerning their judgment,

the annotation, etc. For instance, you can use this frame so as to jot your comment down in case of ambiguous sentences for a third annotator.

In summary, to annotate a sentence:

1. Select from the options shown in the Current Choice box until no options remain.
2. If you make a mistake, double click on features in the Selected Features box to undo the selection.
3. You should save you annotation before closing the window.

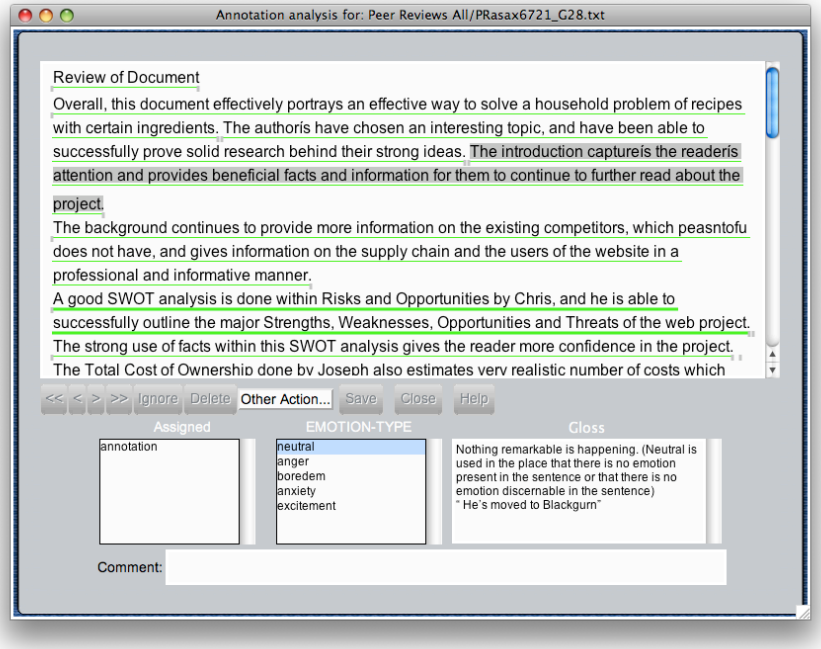


Figure E.3: Annotation Window (The choice of emotion type)

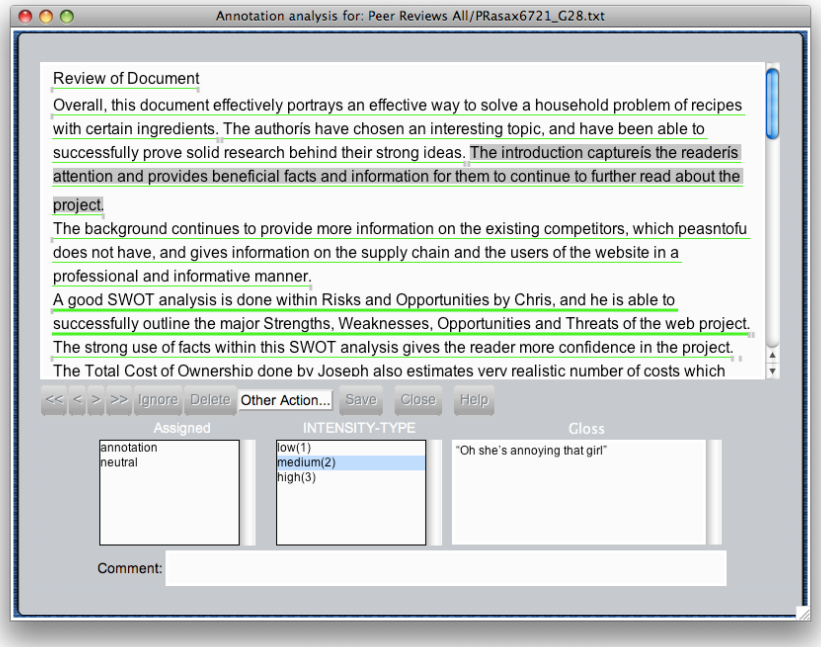


Figure E.4: Annotation Window (The choice of intensity type)

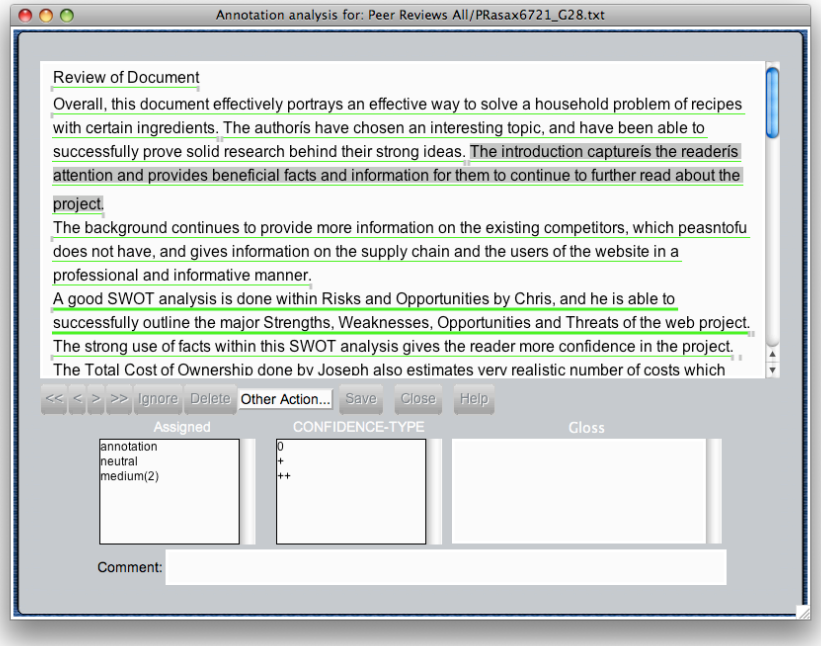


Figure E.5: Annotation Window (The choice of confidence type)

text	uri	start	end	name	confidence	intensity
Please find my review here split into the appropriate section.	PRafar5245.txt	16	78	Anxiety	Low (1)	+
I hope my review will help you improve on your PSD and you will receive an excellent result.	PRafar5245.txt	80	175	Excitement	Low (1)	+
A good idea for national tourism is present and kudos for originality.	PRafar5245.txt	195	265	Neutral	Medium (2)	++
Thorough explanation of exactly who the system users are is required.	PRafar5245.txt	267	336	Neutral	Medium (2)	++
The term ?All Australians?	PRafar5245.txt	340	366	Excitement	Low (1)	+

Figure E.6: Annotation Examples (based on EmotionML 1.0)

Bibliography

Abbasi, A., Chen, H., & Salem, A. (2008). Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums. *ACM Transactions on Information Systems, 26*(3), 1-34.

Agrawal, R., Rajagopalan, S., Srikant, R., & Xu, Y. (2003). *Mining newsgroups using networks arising from social behavior.* Paper presented at the Proceedings of the 12th international conference on World Wide Web.

Akkaya, C., Wiebe, J., & Mihalcea, R. (2009). *Subjectivity word sense disambiguation.* Paper presented at the the 2009 Conference on Empirical Methods in Natural Language Processing, Singapore.

Alm, C. O. (2009). Affect in Text and Speech. *VDM Verlag Dr. Müller*.

Alm, C. O., Roth, D., & Sproat, R. (2005). *Emotions from text: Machine learning for text-based emotion prediction.* Paper presented at the Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Vancouver.

Aman, S., & Szpakowicz, S. (2007). Identifying expressions of emotion in text *Text, Speech and Dialogue* (Vol. 4629, pp. 196-205): Springer Berlin / Heidelberg.

Anttonen, J., & Surakka, V. (2005). *Emotions and heart rate while sitting on a chair.* Paper presented at the Proceedings of the SIGCHI conference on Human factors in computing systems.

Baeza-Yates, R., & Ribeiro-neto, B. (1999). *Modern Information Retrieval*. New York: Addison Wesley.

Baggia, P., Burkhardt, F., Martin, J., Pelachaud, C., Peter, C., Schuller, B., et al. (2008). Elements of an EmotionML 1.0. *W3C Incubator Group Report, M. Schröder, Ed. W3C*.

Bänziger, T., & Scherer, K. R. (2007). Using actor portrayals to systematically study multimodal emotion expression: The GEMEP corpus. *Affective computing and intelligent interaction*, 476-487.

Barrett, L. F. (1998). Discrete emotions or dimensions? The role of valence focus and arousal focus. *Cognition and Emotion, 12*(4), 579-599.

Beineke, P., Hastie, T., & Vaithyanathan, S. (2004). *The Sentimental Factor: Improving Review Classification via Human-Provided Information.* Paper presented at the Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics.

Berry, M. W., & Browne, M. (2005). *Understanding Search Engines: Mathematical Modeling and Text Retrieval*: Society for Industrial and Applied Mathematics.

Bingham, E., & Mannila, H. (2001). *Random projection in dimensionality reduction: applications to image and text data.* Paper presented at the Proceedings of the

seventh ACM SIGKDD international conference on Knowledge discovery and data mining.

Bloom, K., Garg, N., & Argamon, S. (2007). *Extracting appraisal expressions.* Paper presented at the In Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT), Rochester, NY.

Boiy, E. a. H., P. and Deschacht, K. and Moens, M.F. (2007). *Automatic sentiment analysis in on-line text.* Paper presented at the Proceedings of the 11th International Conference on Electronic Publishing (ELPUB 2007), Vienna, Austria.

Bradley, M. M., & Lang, P. J. (1999). Affective norms for English words (ANEW): Instruction manual and affective ratings. [C1]. *University of Florida: The Center for Research in Psychophysiology, Technical*.

Breck, E., Choi, Y., & Cardie, C. (2007). *Identifying expressions of opinion in context.* Paper presented at the the 20th International Joint Conference on Artificial Intelligence Hyderabad, India.

Calvo, R. A., & D'Mello, S. (to appear). Affect Detection: An Interdisciplinary Review of Models,Methods, and their Applications. *IEEE Trans. on Affective Computing*.

Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., et al. (2006). The AMI meeting corpus: A pre-announcement. *LECTURE NOTES IN COMPUTER SCIENCE, 3869*, 28-39.

Chesley, P., Vincent, B., Xu, L., & Srihari, R. (2006). *Using Verbs and Adjectives to Automatically Classify Blog Sentiment.* Paper presented at the Proceedings of AAAICAAW06 the Spring Symposia on Computational Approaches to Analyzing Weblogs.

Cho, Y. H., & Lee, K. J. (2006). Automatic affect recognition using natural language processing techniques and manually built affect lexicon. *IEICE Transactions on Information and Systems, 89*(12), 2964-2971.

Chuang, Z. J., & Wu, C. H. (2004). *Emotion recognition using acoustic features and textual content.* Paper presented at the 2004 IEEE International Conference on Multimedia and Expo.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement, 20*(1), 37-46.

Cowie, R., Douglas-Cowie, E., Apolloni, B., Taylor, J., Romano, A., & Fellenz, W. (1999). What a neural net needs to know about emotion words. *Computational intelligence and applications*, 109-114.

Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., & Schröder, M. (2000). *'FEELTRACE': An instrument for recording perceived emotion in real time.* Paper presented at the ISCA Tutorial and Research Workshop ITRW on Speech and Emotion.

Cowie, R., Douglas-Cowie, E., Savvidou*, S., McMahon, E., Sawey, M., & Schrder, M. (2000). *'FEELTRACE': An instrument for recording perceived emotion in real time.* Paper presented at the ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., et al. (2001). Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine, 18*(1), 32-80.

Craggs, R., & McGee Wood, M. (2004). A categorical annotation scheme for emotion in the linguistic content of dialogue. *Affective Dialogue Systems*, 89-100.

Cui, H., Mittal, V., & Datar, M. (2006). *Comparative experiments on sentiment classification for online product reviews.* Paper presented at the Proceedings of the 21st National Conference on Artificial Intelligence, Boston, Massachusetts.

D'Mello, S., & Graesser, A. (2007). Mind and Body: Dialogue and posture for affect detection in learning environments. *Artificial intelligence in education building technology rich learning contexts that work*, 161.

D'Mello, S., Picard, R. W., & Graesser, A. (2007). Toward an Affect-Sensitive AutoTutor. *IEEE Intelligent Systems, 22*(4), 53-61.

D'Mello, S. K., Craig, S. D., Witherspoon, A., McDaniel, B., & Graesser, A. (2008). Automatic detection of learner's affect from conversational cues. *User Modeling and User-Adapted Interaction, 18*(1), 45-80.

Danisman, T., & Alpkocak, A. (2008). *Feeler : Emotion Classification of Text Using Vector Space Model.* Paper presented at the Proceedings of the Symposium on Affective Language in Human and Machine.

Das, S. R., & Chen, M. Y. (2001). *Yahoo! for Amazon: Extracting market sentiment from stock message boards.* Paper presented at the In Proceedings of the Asia Pacific Finance Association Annual Conference (APFA).

Dave, K., Lawrence, S., & Pennock, D. M. (2003). *Mining the peanut gallery: Opinion extraction and semantic classification of product reviews.* Paper presented at the Proceedings of the 12th international conference on World Wide Web.

Efron, M. (2004). *Cultural orientation: Classifying subjective documents by cocitation analysis.* Paper presented at the AAAI Fall Symposium on Style and Meaning in Language, Art, and Music.

Ekman, P. (1992). An argument for basic emotions. *Cognition Emotion, 6*(3), 169-200.

Esuli, A., & Sebastiani, F. (2006). *SentiWordNet: A publicly available lexical resource for opinion mining.* Paper presented at the Proceedings of the 5th Conference on Language Resources and Evaluation.

Fei, Z., Liu, J., & Wu, G. (2004). *Sentiment Classification Using Phrase Patterns.* Paper presented at the Proceedings of the 4th International Conference on Computer and Information Technology.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*(5), 378-382.

Fleiss, J. L., & Cohen, J. (1973). The Equivalence of Weighted Kappa and the Intraclass Correlation Coefficient as Measures of Reliability. *Educational and psychological measurement, 33*(3), 613-619.

Francisco, V., & Gervás, P. (2006). *Automated mark up of affective information in english texts*. Paper presented at the Proceedings of the International Conference on Text, Speech and Dialogue.

Gamon, M. (2004). *Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis.* Paper presented at the

Proceedings of the 20th International Conference on Computational Linguistics COLING.

Généreux, M., & Evans, R. (2006). *Towards a validated model for affective classification of texts.* Paper presented at the Proceedings of the Workshop on Sentiment and Subjectivity in Text.

Grings, W. W., & Dawson, M. E. (1978). *Emotions and bodily responses: A psychophysiological approach*: Academic Press.

Hancock, J. T., Landrigan, C., & Silver, C. (2007). *Expressing emotion in text-based communication.* Paper presented at the In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, San Jose, CA.

Hatzivassiloglou, V., & McKeown, K. R. (1997). *Predicting the semantic orientation of adjectives.* Paper presented at the Proceedings of the 35th annual meeting on Association for Computational Linguistics.

He, H., & Garcia, E. A. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering, 21*(9), 1263-1284.

Hevner, K. (1936). Experimental Studies of the Elements of Expression in Music. *Americal Journal of Psychology, 48*(2), 246-268.

Hofmann, T. (1999). *Probabilistic latent semantic indexing.* Paper presented at the the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, Berkeley, California, United States.

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning, 42*(1), 177-196.

Holzman, L. E., & Pottenger, W. M. (2003). Classification Of Emotions in Internet Chat: An Application of Machine Learning Using Speech Phonemes: Oxford University Press.

Hummel, R. A., & Zucker, S. W. (1983). On the Foundations of Relaxation Labeling Processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 5*(3), 267-287.

Kanayama, H., Tetsuya, N., & Hideo, W. (2004). *Deeper sentiment analysis using machine translation technology.* Paper presented at the Proceedings of the 20th International Conference on Computational Linguistics, Geneva, Switzerland.

Kim, S.-M., & Hovy, E. (2004). *Determining the sentiment of opinions.* Paper presented at the Proceedings of the 20th international conference on Computational Linguistics, Geneva, Switzerland.

Kim, S. M., & Calvo, R. A. (2010). *Sentiment Analysis in Student Experiences of Learning.* Paper presented at the the 3rd International Conference on Educational Data Mining, Pittsburgh, USA.

Kim, S. M., & Hovy, E. (2006). *Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Text.* Paper presented at the Proceedings of the Workshop on Sentiment and Subjectivity in Text.

Kleinginna, P. R., & Kleinginna, A. M. (1981). A categorized list of emotion definitions, with suggestions for a consensual definition. *Motivation and Emotion, 5*(4), 345-379.

Kort, B., Reilly, R., & Picard, R. W. (1990). *An affective model of interplay between emotions and learning: reengineering educational pedagogy-building a learning*

companion. Paper presented at the Proceedings IEEE International Conference on Advanced Learning Technologies.

Krenn, B. (2003). *The NECA project: Net environments for embodied emotional conversational agents.* Paper presented at the Proceedings of Workshop on Emotionally Rich Virtual Worlds with Emotion Synthesis at the 8th International Conference on 3D Web Technology, St. Malo, France.

Krenn, B., Pirker, H., Grice, M., Baumann, S., Piwek, P., Van Deemter, K., et al. (2002). *Generation of multimodal dialogue for net environments.* Paper presented at the Konvens, Saarbrücken, Germany.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes,* pp. 259-284.

Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (2007). *Handbook of Latent Semantic Analysis*: Laurence Erlbaum and Associates.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174.

Lang, P. J. (1980). Behavioral treatment and bio-behavioral assessment: Computer applications. *Technology in mental health care delivery systems*, 119-137.

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature, 401*(6755), 788-791.

Li, Y., Bontcheva, K., & Cunningham, H. (2007). *Experiments of Opinion Analysis on the Corpora MPQA and NTCIR-6.* Paper presented at the Proceedings of the 6th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access, Tokyo, Japan.

Lichtenstein, A., Oehme, A., Kupschick, S., & Jürgensohn, T. (2008). Comparing Two Emotion Models for Deriving Affective States from Physiological Data. *Affect and Emotion in HumanComputer Interaction*, 35-50.

Lin, D. (1998, 24–27 July). *An information-theoretic definition of similarity.* Paper presented at the Proceedings of the 15th International Conference on Machine Learning, Madison, WI USA.

Litman, D. J., & Forbes-Riley, K. (2004). *Predicting student emotions in computer-human tutoring dialogues.* Paper presented at the Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, Barcelona, Spain.

Litman, D. J., & Silliman, S. (2004). *ITSPOKE: An intelligent tutoring spoken dialogue system.* Paper presented at the Proceedings of the Human Language Technology Conference and North American Chapter of the Association for Computational Linguistics, Boston, Massachusetts.

Liu, B., Hu, M., & Cheng, J. (2005). *Opinion observer: Analyzing and comparing opinions on the web.* Paper presented at the Proceedings of the 14th international conference on World Wide Web, Chiba, Japan.

Liu, H., Lieberman, H., & Selker, T. (2003a). *A model of textual affect sensing using real-world knowledge.* Paper presented at the Proceedings of the 8th International Conference on Intelligent User Interfaces.

Liu, H., Lieberman, H., & Selker, T. (2003b). *A model of textual affect sensing using real-world knowledge.* Paper presented at the the 8th international conference on Intelligent user interfaces, Miami, Florida, USA.

Ma, C., Prendinger, H., & Ishizuka, M. (2005). Emotion Estimation and Reasoning Based on Affective Textual Interaction. *Affective computing and intelligent interaction*, 622-628.

Mac Kim, S., Valitutti, A., & Calvo, R. A. (2010, June 2010). *Evaluation of Unsupervised Emotion Models to Textual Affect Recognition.* Paper presented at the the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, Los Angeles, California.

Manning, C. D., Raghavan, P., & Schütze, H. (2009). *An Introduction to Information Retrieval*: Cambridge University Press.

Manning, C. D., & Schütze, H. (2000). *Foundations of Statistical Natural Language Processing*. Cambridge, MA USA: MIT Press.

Mehrabian, A. (1995). Framework for a comprehensive description and measurement of emotional states. *Genetic, social, and general psychology monographs, 121*(3), 339-361.

Mehrabian, A. (1996). Pleasure-arousal-dominance: A general framework for describing and measuring individual. *Current Psychology, 15*(4), 505-525.

Mihalcea, R., Corley, C., & Strapparava, C. (2006). *Corpus-based and Knowledge-based Measures of Text Semantic Similarity.* Paper presented at the Proceedings of the National Conference on Artificial Intelligence.

Mihalcea, R., & Liu, H. (2006). *A corpus-based approach to finding happiness.* Paper presented at the Proceedings of the AAAI Spring Symposium on Computational Approaches to Weblogs.

Morinaga, S., Yamanishi, K., Tateishi, K., & Fukushima, T. (2002). *Mining Product Reputations on the Web.* Paper presented at the Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining.

Mullen, T., & Collier, N. (2004). *Sentiment analysis using support vector machines with diverse information sources.* Paper presented at the Proceedings of the Empirical Methods in Natural Language Processing.

Nasukawa, T., & Yi, J. (2003). *Sentiment analysis: capturing favorability using natural language processing.* Paper presented at the Proceedings of the 2nd International Conference on Knowledge Capture.

Neviarouskaya, A., Prendinger, H., & Ishizuka, M. (2007). Textual Affect Sensing for Sociable and Expressive Online Communication. *LECTURE NOTES IN COMPUTER SCIENCE, 4738*, 218-229.

Ng, V., Dasgupta, S., & Arifin, S. M. N. (2006). *Examining the Role of Linguistic Knowledge Sources in the Automatic Identification and Classification of Reviews.* Paper presented at the Proceedings of the COLINGACL 2006 Main Conference Poster Sessions.

Nigam, K., & Hurst, M. (2004). *Towards a robust metric of opinion.* Paper presented at the AAAI Spring Symposium on Exploring Attitude and Affect in Text.

Ortony, A., Clore, G. L., & Collins, A. (1988). *The Cognitive Structure of Emotion*: Cambridge, UK.: Cambridge University Press.

Osgood, C. E., May, W. H., & Miron, M. S. (1975). *Cross-cultural universals of affective meaning*. Urbana: University of Illinois Press.

Pang, B., & Lee, L. (2004). *A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts.* Paper presented at the

Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, Barcelona, Spain.

Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends® in Information Retrieval, 2*(1-2), 1-135.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval, 2*(1-2), 1-135.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment Classification using Machine Learning Techniques. *Proceedings of the ACL02 conference on Empirical methods in natural language processing EMNLP 02, 10*(July), 79-86.

Parlade, M. V., Messinger, D. S., Delgado, C. E. F., Kaiser, M. Y., Van Hecke, A. V., & Mundy, P. C. (2009). Anticipatory smiling: Linking early affective communication and social outcome. *Infant Behavior and Development, 32*(1), 33-43.

Penumatsa, P., Ventura, M., Graesser, A. C., Franceschetti, D. R., Louwerse, M., Hu, X., et al. (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.

Pereira, F., Tishby, N., & Lee, L. (1993, 22–26 June). *Distributional clustering of English words.* Paper presented at the Proceedings of the 31st annual meeting on Association for Computational Linguistics, Columbus, Ohio USA.

Picard, R. W. (1997). *Affective Computing*: MIT Press.

Plutchik, R. (1980). *Emotion: A Psychoevolutionary Synthesis*: Harpercollins College Div.

Popescu, A.-M., & Etzioni, O. (2005). *Extracting product features and opinions from reviews.* Paper presented at the Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, Vancouver, British Columbia, Canada.

Rainville, P., Bechara, A., Naqvi, N., & Damasio, A. R. (2006). Basic emotions are associated with distinct patterns of cardiorespiratory activity. *International Journal of Psychophysiology, 61*(1), 5-18.

Riloff, E., & Wiebe, J. (2003). *Learning Extraction Patterns for Subjective Expressions.* Paper presented at the Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing.

Ritz, T., Thöns, M., Fahrenkrug, S., & Dahme, B. (2005). Airways, respiration, and respiratory sinus arrhythmia during picture viewing. *Psychophysiology, 42*(5), 568-578.

Russell, J. A. (1979). Affective space is bipolar. *Journal of Personality and Social Psychology, 37*(3), 345-356.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology, 39*(6), 1161-1178.

Russell, J. A., Lewicka, M., & Niit, T. (1989). A Cross-Cultural Study of a Circumplex Model of Affect. *Journal of Personality and Social Psychology, 57*(5), 848-856.

Schapire, R. E., & Singer, Y. (2000). BoosTexter: A Boosting-based System for Text Categorization. *Machine Learning, 39*(2), 135-168.

Scherer, K. R., & Wallbott, H. G. (1994). Evidence for universality and cultural variation of differential emotion response patterning. *Journal of Personality and Social Psychology, 66*(2), 310-328.

Schröder, M. (2004). Dimensional emotion representation as a basis for speech synthesis with non-extreme emotions. *Affective Dialogue Systems*, 209-220.

Schröder, M., Cowie, R., Douglas-Cowie, E., Westerdijk, M., & Gielen, S. (2001). *Acoustic correlates of emotion dimensions in view of speech synthesis.* Paper presented at the 7th European Conference on Speech Communication and Technology.

Scott, S., & Matwin, S. (1998). *Text classification using WordNet hypernyms.* Paper presented at the Use of WordNet in Natural Language Processing Systems Proceedings of the Conference.

Seol, Y. S., Kim, D. J., & Kim, H. W. (2008). *Emotion Recognition from Text Using Knowledge-based ANN.* Paper presented at the Proceedings of 23rd International Technical Conference on Circuits/Systems, Computers and Communications.

Shaikh, M., Prendinger, H., & Ishizuka, M. (2008). Sentiment assessment of text by analyzing linguistic features and contextual valence assignment. *Applied Artificial Intelligence, 22*(6), 558-601.

Sindhwani, V., & Melville, P. (2008). *Document-Word Co-Regularization for Semi-supervised Sentiment Analysis.* Paper presented at the In Proceedings of the Eighth IEEE International Conference on Data Mining (ICDM).

Somasundaran, S., Ruppenhofer, J., & Wiebe, J. (2007). *Detecting arguing and sentiment in meetings.* Paper presented at the Proceedings of the SIGdial Workshop on Discourse and Dialogue.

Strapparava, C., & Mihalcea, R. (2007). *Semeval-2007 task 14: Affective text.* Paper presented at the Proceedings of the Fourth International Workshop on Semantic Evaluations.

Strapparava, C., & Mihalcea, R. (2007). *SemEval-2007 task 14: affective text.* Paper presented at the the 4th International Workshop on Semantic Evaluations, Prague, Czech Republic.

Strapparava, C., & Mihalcea, R. (2008). *Learning to identify emotions in text.* Paper presented at the Proceedings of the 2008 ACM symposium on Applied computing, Fortaleza, Ceara, Brazil.

Strapparava, C., & Valitutti, A. (2004). *WordNet-Affect: an Affective Extension of WordNet.* Paper presented at the Proceedings of the 4th International Conference on Languages Resources and Evaluation.

Strapparava, C., Valitutti, A., & Stock, O. (2006). *The affective weight of lexicon.* Paper presented at the Proceedings of the 5th International Conference on Language Resources and Evaluation.

Strapparava, C., Valitutti, A., & Stock, O. (2007). *Dances with words.* Paper presented at the Proceedings of the International Joint Conference on Artificial Intelligence.

Subasic, P., & Huettner, A. (2001). Affect analysis of text using fuzzy semantic typing. *Fuzzy Systems, IEEE Transactions on, 9*(4), 483-496.

Suzuki, Y., Takamura, H., & Okumura, M. (2006). *Application of semi-supervised learning to evaluative expression classification.* Paper presented at the In Proceedings of the 7th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2006), Mexico City, Mexico.

Thayer, R. E. (1989). *The biopsychology of mood and arousal*: Oxford University Press.

Thomas, M., Pang, B., & Lee, L. (2006). *Get out the vote: Determining support or opposition from Congressional floor-debate transcripts.* Paper presented at the Proceedings of Empirical Methods in Natural Language Processing (EMNLP).

Turney, P. D. (2002). *Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews.* Paper presented at the Proceedings of the 40th annual meeting of the Association for Computational Linguistics.

Van Rijsbergen, C. J. (1979). *Information Retrieval, 2nd edition*: Dept. of Computer Science, University of Glasgow.

Whissell, C. M. (1989). The dictionary of affect in language *Emotion: Theory, Research, and Experience* (pp. 113–131). New York: Academic Press.

Whitelaw, C., Garg, N., & Argamon, S. (2005a). *Using appraisal groups for sentiment analysis.* Paper presented at the Proceedings of the 14th ACM international conference on Information and knowledge management.

Whitelaw, C., Garg, N., & Argamon, S. (2005b). *Using appraisal taxonomies for sentiment analysis.* Paper presented at the Proceedings of the 14th ACM International Conference on Information and Knowledge Management.

Wiebe, J., Wilson, T., Bruce, R., Bell, M., & Martin, M. (2004). Learning Subjective Language. *Computational linguistics, 30*(3), 277-308.

Wiebe, J. M., Bruce, R. F., & O'Hara, T. P. (1999). *Development and use of a gold-standard data set for subjectivity classifications.* Paper presented at the Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics.

Wilson, T., Wiebe, J., & Hoffmann, P. (2005). *Recognizing contextual polarity in phrase-level sentiment analysis.* Paper presented at the Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing.

Yi, J., Nasukawa, T., Bunescu, R., & Niblack, W. (2003). *Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques.* Paper presented at the Third IEEE International Conference on Data Mining.

Yu, H., & Hatzivassiloglou, V. (2003). *Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences.* Paper presented at the In Proceedings of the Conference on Empirical Methods in Natural Language Processing.

Zara, A., Maffiolo, V., Martin, J., & Devillers, L. (2007). Collection and annotation of a corpus of human-human multimodal interactions: Emotion and others anthropomorphic characteristics. *Affective Computing and Intelligent Interaction, 4738*, 464-475.

Zeimpekis, D., & Gallopoulos, E. (2006). TMG: A MATLAB toolbox for generating term-document matrices from text collections *Grouping multidimensional data Recent advances in Clustering* (pp. 187-210): Springer.

Zhang, L., Barnden, J. A., Hendley, R. J., & Wallington, A. M. (2006). *Exploitation in affect detection in open-ended improvisational text.* Paper presented at the In Proceedings of Workshop on Sentiment and Subjectivity in Text, Sydney, Australia.