uOttawa

L'Université canadienne
Canada's university

FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTOCTORALES

uOttawa

L'Université canadienne
Canada's university

FACULTY OF GRADUATE AND POSDOCTORAL STUDIES

Saima Aman

AUTEUR DE LA THÈSE / AUTHOR OF THESIS

M.C.S.

GRADE / DEGREE

School of Information Technology and Engineering

FACULTÉ, ÉCOLE, DÉPARTEMENT / FACULTY, SCHOOL, DEPARTMENT

Recognizing Emotions in Text

TITRE DE LA THÈSE / TITLE OF THESIS

Stan Szpakowicz

DIRECTEUR (DIRECTRICE) DE LA THÈSE / THESIS SUPERVISOR

CO-DIRECTEUR (CO-DIRECTRICE) DE LA THÈSE / THESIS CO-SUPERVISOR

EXAMINATEURS (EXAMINATRICES) DE LA THÈSE / THESIS EXAMINERS

F. Oppacher

P. Turney

Gary W. Slater

Le Doyen de la Faculté des études supérieures et postdoctorales / Dean of the Faculty of Graduate and Postdoctoral Studies

Recognizing Emotions in Text

by

Saima Aman

Thesis submitted to the
Faculty of Graduate and Postdoctoral Studies
In partial fulfillment of the requirements
For the degree of Master of Computer Science (MCS)

Ottawa-Carleton Institute for Computer Science
School of Information Technology and Engineering
University of Ottawa

Your file Votre référence
ISBN: 978-0-494-34054-7
Our file Notre référence
ISBN: 978-0-494-34054-7

Abstract

People express emotions as part of everyday communication. Emotions can be judged by a combination of cues such as facial expressions, prosodies, gestures, and actions. Emotions are also articulated by written texts. Inspired by works in sentiment analysis, this thesis explores approaches to automatic detection of emotions in text. I draw from emotion theories in the fields of psychology and linguistics, and use natural language processing and machine learning techniques for automatic emotion detection.

In this thesis, I describe studies and experiments in manual and automatic recognition of expressions of the six basic emotions (Ekman, 1992) – *happiness*, *sadness*, *anger*, *disgust*, *surprise*, and *fear* – in text form. The text under study comprises data collected from blogs, representing texts rich in emotion content and therefore suitable for this study. The first task I consider is to prepare a corpus annotated with emotion-related information. The annotations include emotion category out of the aforementioned six; emotion intensity at one of the four levels – *high*, *medium*, *low*, and *neutral*; and spans of text indicating emotion expressions within sentences. To measure consensus amongst human judges on annotation and to gauge the complexity of the task under investigation, an inter-annotator agreement study is also performed.

In my work, I investigate features that can help differentiate emotion from non-emotion. A combination of lexical and semantic features is used to train classifiers for emotion/non-emotion classification. I describe the techniques implemented for discerning the six basic emotion categories. I use a knowledge-based approach for this task based on semantic resources – WordNet-Affect and Roget's Thesaurus. This approach provides results that significantly surpass the baseline term-counting method. For emotion intensity recognition, I use corpus-based syntactic bigrams to improve the performance of a baseline system that only uses unigrams as features.

Acknowledgements

It is by the blessings of God that I was able to produce this work. If emotions can truly be expressed in text, I would like to use this medium to express my gratitude to all those who have made this endeavor possible.

No words can express my gratitude to my husband, Suhail Jalil, for his unwavering support and encouragement during my Master's. His patience and support provided me the necessary strength to work on my thesis. I could always count on him to discuss my ideas and get feedback on them.

I thank my parents for having faith in me and for their support in my academic pursuits. I also thank Manal, Aydah, Iram, Mummy and Papa for their support. Their love and support was vital for enabling me to complete this project.

I would like to thank my supervisor, Dr. Stan Szpakowicz, for his guidance and support during my Master's program. The high standards that he sets for his students ultimately help in bringing out the best in them. He was always quick in his response to my questions, and encouraged me all the way. His attention to detail while still being able to see the forest for the trees was a unique asset in this work.

I would also like to thank members of the NLP group at University of Ottawa for many educational and inspiring sessions. In particular, I would like to thank Dr. Diana Inkpen, Dr. Marina Sokolova, and Oana Frunza for their help whenever I needed. I take this opportunity to thank all my teachers at Aligarh and Ottawa who taught me so many things both inside and outside the classroom.

I am grateful to the annotators who gave their time to label the data used in this work. Without their efforts, this work would not have been possible.

My friends from Aligarh, Ottawa, and San Diego have encouraged me through this endeavor. I will cherish the special bond we have for life.

Table of Contents

Abstract iii
Acknowledgements v
Table of Contents vii
List of Figures ix
List of Tables x
Chapter 1 Introduction 1
1.1 Problem Definition and Scope 2
1.2 Motivation and Context 3
1.3 Thesis Roadmap 4
Chapter 2 Background and Related Work 6
2.1 Emotion Theories 6
2.2 Sentiment Analysis 9
2.3 Emotion Recognition 14
2.4 Sentiment Classification Methods and Resources 20
2.5 Conclusion 25
Chapter 3 Data Acquisition and Annotation 26
3.1 Data Selection 26
3.2 Annotation Scheme 28
3.3 Data Description 33
3.4 Annotation Agreement Measurement 36
3.5 Conclusion 42
Chapter 4 Differentiating Emotion Sentences from Non-emotion Sentences 43
4.1 Defining the Feature Set 43
4.2 Experiments and Results 47
4.3 Discussion 49
4.4 Conclusion 50
Chapter 5 Fine-grained Emotion Classification 51
5.1 The Data 52
5.2 Developing a Baseline 53

5.3 Corpus-based features 55
5.4 Features derived from *Roget's* Thesaurus 56
5.5 Features derived from WordNet-Affect 60
5.6 Experiments and Results 61
5.7 Discussion 63
5.8 Conclusion 65
Chapter 6 Emotion Intensity Recognition 66
6.1 The Data 66
6.2 Emotion Intensity Expressions 67
6.3 Experiments and Results 71
6.4 Discussion 72
6.5 Conclusion 73
Chapter 7 Conclusion 74
7.1 Contributions 76
7.2 Future Work 77
Appendix A Lists of emotion-related seed words used to build blog corpus 79
Appendix B Annotation Instructions 80
Bibliography 85

List of Figures

Figure 2.1 Watson and Tellegen's Circumplex Theory of Affect (Watson and Tellegen, 1985) 8
Figure 3.1 A sample of the blog data used in this research 29
Figure 3.2 An example of dynamic progression of emotion across sentences 29
Figure 3.3 An example of mixed emotions 30
Figure 3.4 Another example of more than one emotion in a sentence 30
Figure 3.5 An example of a sentence with no clear emotion 31
Figure 3.6 Examples of emotion indicator spans 33
Figure 3.7 Example of emotion modifiers in emotion indicator spans 33
Figure 3.8 Disagreement in marking emotion indicator spans 33
Figure 3.9 Sample annotated data from the corpus 34
Figure 3.10 Calculation of MASI on sets A and B 40
Figure 3.11 Sample sentence to illustrate agreement measurement using MASI 41
Figure 3.12 Sample sentence to illustrate the IO method of agreement measurement 42
Figure 4.1 Sample GI entries for some emotion-related words 45
Figure 4.2 Sample emotive sentence with no emotion-bearing word 50
Figure 5.1 Sample emotion sentence illustrating negation 55
Figure 5.2 Top-level Classes in the Roget's Thesaurus 57
Figure 6.1Sample output from Link Parser 68

List of Tables

Table 2.1 Basic Emotion Categories Identified by Researchers 7
Table 3.1 Emotion Categories used in Annotation 31
Table 3.2 Emotion Intensities used in Annotation 32
Table 3.3 Details of the datasets in the corpus 34
Table 3.4 Distribution of Emotion Categories in Annotation Sets A1 and A2 36
Table 3.5 Distribution of Emotion Intensity in Annotation Sets A1 and A2 36
Table 3.6 Most frequent Emotion Indicators in the data 37
Table 3.7 Pairwise agreement in emotion/non-emotion labeling 38
Table 3.8 Pairwise agreement in emotion categories 38
Table 3.9 Pairwise agreement in emotion intensities 39
Table 3.10 Pairwise agreement in emotion indicators (using MASI) 41
Table 3.11 Pairwise agreement in emotion indicators (using kappa) 42
Table 4.1Description of emotion word lists extracted from WordNet-Affect 46
Table 4.2 Summary of features used in emotion/non-emotion classification 47
Table 4.3 Class distribution in the dataset used in emotion/non-emotion classification 48
Table 4.4 Results of emotion/non-emotion classification 49
Table 5.1 Distribution of emotion classes in the dataset 52
Table 5.2 Performance metrics of the baseline system 54
Table 5.3 Sample words from the emotion lexicon built using Roget's Thesaurus 59
Table 5.4 Results of fine-grained classification using SVM 62
Table 5.5 Results of fine-grained classification using Naive Bayes 64
Table 6.1 Distribution of emotion intensity levels in the data 67
Table 6.2 Adverb-adjective links and examples 69
Table 6.3 Adverb-adverb links and examples 69
Table 6.4 Other adverb related links and examples 69
Table 6.5 Other adjective related links and examples 70
Table 6.6 Some other emotion-related links and examples 70
Table 6.7 Results of emotion intensity classification using SVM 72

Chapter 1
Introduction

The advantage of the emotions is that they lead us astray,
and the advantage of science is that it is not emotional.
***Oscar Wilde** (1854 - 1900)*

Language is a powerful tool to communicate and convey information. It is also a means to express emotion. Natural Language Processing (NLP) techniques have long been applied to automatically identify the information content in text. Applications such as topic-based text categorization, summarization, question-answering systems, and information retrieval systems typically focus on the information contained in text. This work is an endeavor to apply NLP techniques to identify emotions expressed in text.

In recent years, research inspired by Artificial Intelligence (AI) has focused increasing efforts on developing systems that incorporate emotion. Emotions are crucial to several natural processes that are modeled in AI systems. These include perception, reasoning, learning, and natural language processing. Emotion research is significant for developing affective[1] interfaces – ones that can make sense of emotional inputs, provide appropriate emotional responses, and facilitate online communication through animated affective agents. Such interfaces can greatly help improve user experience in Computer-Mediated Communication (CMC) and Human-Computer Interaction (HCI). Emotion research is also vital for text-to-speech (TTS) synthesis systems. Emotion-aware TTS systems can identify

[1] The word "*affect*" has been used interchangeably with the word "*emotion*" at many places in this thesis.

emotional nuances in written text and accordingly provide more natural rendering of text in spoken form.

Automatic emotion detection and analysis methods are also useful in many applications with psychological basis. For example, they can be successfully applied to learn user preferences and interests from users' personal writings and speeches. These methods are often studied in the scope of the domain of personality modeling and consumer feedback analysis. Similarly, e-learning systems can benefit from affective tutoring approaches.

Emotion research has recently attracted renewed attention of the scientific community, as evident from the increased number of events related to it. The *International Conference on Affective Computing and Intelligent Interaction (ACII-2007)*[2] seeks to bring together researchers from various disciplines that study emotion and related phenomena. Within the NLP domain, emotion recognition is one of the tasks at the *International Workshop on Semantic Evaluations (Semeval-2007)*[3]. The *International Conference on Language Resources and Evaluation (LREC-2006)*[4] also had a workshop on Emotional Corpora.

1.1 Problem Definition and Scope

Before embarking on the task of emotion recognition, it is imperative to define precisely the goals of this work. I address the task of determining emotions expressed in text at the sentence level. More specifically, the goal is to assign automatically an emotion label to each sentence in the given dataset, indicating the predominant emotion type expressed in the sentence. The possible labels are *happiness*, *sadness*, *anger*, *disgust*, *surprise*, *fear* and *no-emotion*. Those are the six basic emotion categories identified by Ekman (1992), and an additional label to account for the absence of a clearly discernible emotion. I also address the

[2] http://gaips.inesc-id.pt/acii2007/index.html

[3] Affective Text: Semeval Task at the 4th International Workshop on Semantic Evaluations, 2007, Prague. (http://nlp.cs.swarthmore.edu/semeval/tasks/task14/summary.shtml)

[4] http://www.lrec-conf.org/lrec2006/IMG/pdf/programWSemotion-LREC2006-last1.pdf

task of determining the intensity of the emotion expressed in a sentence. I consider four levels – *high, medium, low,* and *neutral*. The data for this work is drawn from blogs. Blogs mostly comprise unedited, first person narratives related to a variety of interpersonal and public issues, which makes them potentially rich in emotive content.

Emotional states have cognitive bases and are shaped by several factors. Emotions manifest themselves in the form of facial expressions as well as linguistic expressions – both verbal and written. The scope of this work is limited to determining the emotional orientation of sentences, as much as is evident from the written text. The criterion for determining what constitutes emotional text and what is the type of emotion expressed is human judgment. It is common experience that just as emotion can be expressed in many ways, the same expression may be interpreted differently by different readers. The techniques introduced in this work were evaluated against data whose emotional orientation was judged the same by at least two humans.

1.2 Motivation and Context

The motivation for this work has come from the recent growing interest in the sentiment analysis field. The rapid growth of the World Wide Web has facilitated increased online communication and opened up newer avenues for the general public to post their opinions online. This has led to generation of large amounts of online content rich in user opinions, sentiments, emotions, and evaluations. We need computational approaches to successfully analyze this online content, recognize and aggregate relevant information, and draw useful conclusions.

Much of the current work in this direction has typically focused on recognizing the polarity of sentiment (*positive/negative*). Among the less explored sentiment areas is the recognition of types of emotions and their intensity – the focus of this work. Recognizing emotions conveyed by a text can provide an insight into the author's intent and sentiment, and can lead to better understanding of the text's content.

The inspiration for this work has also come from studies in psychology, which focus on analyzing emotive texts to gain a deeper understanding of the way people express different kinds of emotions. These studies are typically carried out in controlled laboratory settings, or drawn from academic writings or medical domain (Pennebaker et al., 2003). Acquiring this kind of data from the Web – as done in this work – brings attention to a hitherto unexplored source of emotive text.

1.3 Thesis Roadmap

This thesis presents a wide spectrum of work investigating emotion expression in text – ranging from manual identification of emotive content in text to computational approaches to automatic emotion identification.

The thesis is organized as follows:

- **Chapter 2** builds a background for the work presented in subsequent chapters of the thesis. It informs the reader about the popular theories of emotion in the scientific domain as well as about the various knowledge resources used in automatic emotion analysis. A survey of related works – utilizing computational approaches in sentiment analysis and textual emotion detection – puts into perspective the work reported in this thesis.
- **Chapter 3*** describes how data used in this work was acquired and labeled with various emotion annotations. The description of manual labeling work is accompanied by an assessment of annotation agreement among human judges. Sample sentences from the annotated data illustrate various aspects of emotion expression and identification.
- **Chapter 4*** presents techniques used for distinguishing emotion sentences from non-emotion ones. A description of the features chosen for characterizing emotional content in sentences as well as of ML techniques adopted for emotion/non-emotion classification is provided. The usefulness of external knowledge resources in automatic emotion detection task is demonstrated.

---

* The work presented in Chapters 3 and 4 is slated to appear in (Aman and Szpakowicz, 2007).

- **Chapter 5** presents experiments in fine-grained emotion classification. First, a naïve rule-based baseline is developed. Later, approaches using ML techniques are presented. These approaches use corpus-based and emotion-related features. A novel method of automatically acquiring words relevant to emotions is also described.
- **Chapter 6** describes experiments in emotion intensity recognition. A combination of corpus-base unigram and syntactic bigram features are used in the ML experiments. The usefulness of the latter in intensity recognition is demonstrated.
- **Chapter 7** summarizes the work presented in this thesis and distills the main contributions. An outline of future work is also presented.

Chapter 2
Background and Related Work

If I have seen further, it is by standing on the shoulders of giants.
***Isaac Newton** (1643-1727)*

This chapter presents a review of the work done in the different domains that this research draws upon. It begins, in Section 2.1, with a survey of the various theories and models of emotion as well as communication of emotion. The broad area of sentiment analysis is introduced in Section 2.2. In Section 2.3, the discussion is narrowed down to the particular area of sentiment analysis that is the focus of this work, namely, emotion recognition. The chapter concludes with a survey of the different methods and approaches taken by the researchers in sentiment analysis.

2.1 Emotion Theories

Emotions have fascinated researchers for long, as is evident in the vast body of research work related to emotion in fields of psychology, linguistics, social sciences, and communication. Human emotion manifests itself in the form of facial expressions, speech utterances, writings, and in gestures and actions. Consequently, scientific research in emotion has been pursued along several dimensions and has drawn upon research from various fields. This thesis addresses the task of emotion recognition by attempting to automatically learn emotions from text.

The French philosopher René Descartes' treatise, *Les passions de l'âme* (Passions of the Soul), published in 1649, is considered to be among the earliest works to theorize emotions (Anscombe and Geach, 1970; Cowie, 2000). The basic hypothesis presented in the treatise categorizes emotions into primary emotions and secondary emotions.

More recently, researchers have investigated several aspects of human emotion in order to arrive at a set of emotion categories that are universally acceptable (Picard, 1997). Several works in this direction have been reported in the literature (Tomkins, 1962; Izard, 1977; Plutchik, 1980; Ortony et. al., 1988; and Ekman, 1992). Table 2.1 lists the basic emotion categories identified by the different researchers.

Table 2.1 Basic Emotion Categories Identified by Researchers

Tomkins (1962)	Izard (1977)	Plutchik (1980)	Ortony et.al. (1988)	Ekman (1992)
joy	enjoyment	joy	joy	happiness
anguish	sadness	sorrow	sadness	sadness
fear	fear	fear	fear	fear
anger	anger	anger	anger	anger
disgust	disgust	disgust	disgust	disgust
surprise	surprise	surprise	surprise	surprise
interest	interest	acceptance		
shame	shame	anticipation		
	shyness			
	guilt			

Some psychologists have investigated facial expressions of emotion to identify the basic discriminable expressions among them, and mapped them to basic human emotions. Ekman (1992) has defined basic emotions as those that have universally accepted distinctive facial expressions. The six basic emotions defined on this basis are *happiness*, *sadness*, *fear*, *anger*, *disgust*, and *surprise*. This work uses Ekman's emotion categories since these emotions have been most widely accepted by the different researchers (see Table 2.1). Ekman's emotion categories have also been previously used in other computational approaches to emotion recognition (Liu et al., 2003; Alm et al., 2005; and Neviarouskaya et al., 2007a,b).

Some researchers such as Schlosberg (1954) have referred to continuous dimensions of emotion instead of distinct emotion categories. There is agreement among researchers on at least two of these dimensions: *valence* (positive/negative) and *arousal* (calm/excited) (Barrett, 1998). Plutchik (1980) and Frijda et al. (1992) have highlighted the role of the intensity component in the study of emotion.

The Circumplex Theory of Affect (Watson and Tellegen, 1985) identifies two main dimensions of positive and negative affect, which range from high to low. It also recognizes an alternative set of dimensions in the model, which consist of *pleasantness-unpleasantness* and *engagement-disengagement*. These two systems of axes define eight emotion octants in the model (Figure 2.1). Each emotion octant is marked with three to six characteristic affect words. These words were chosen on the basis of self-reported experiences of emotions by human subjects (Rubin et al., 2004).

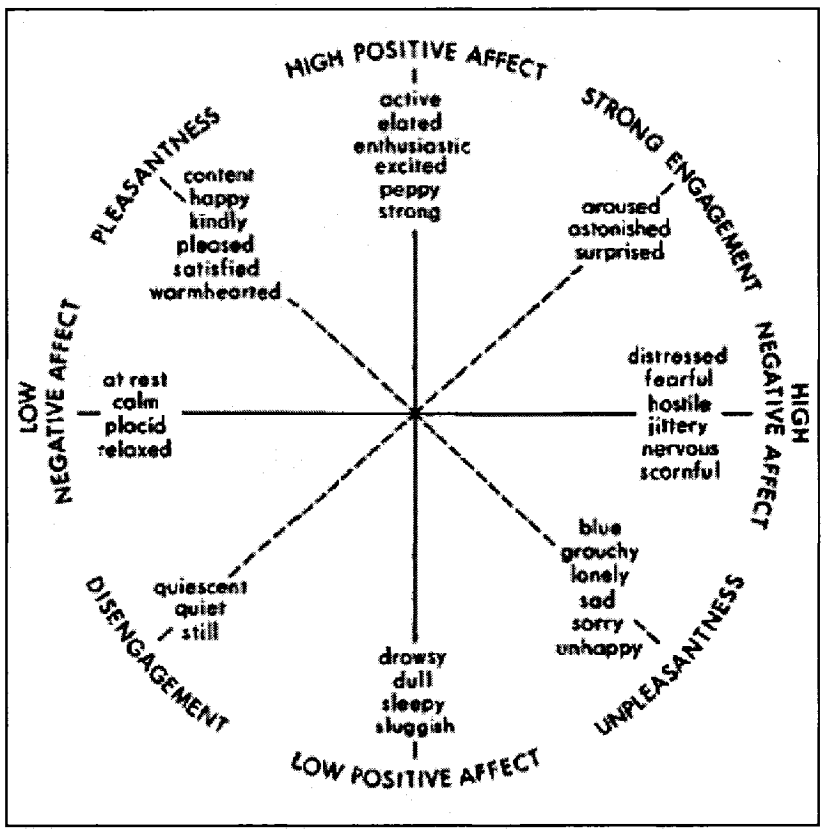

Figure 2.1 Watson and Tellegen's Circumplex Theory of Affect (Watson and Tellegen, 1985)

Several works have been reported in the literature on the study of emotion expression in texts. The communicative function model of language introduced by the Russian-American linguist, Roman Jakobson (1960) identifies *emotive* function as one the six functions of language. The written expression of emotion lacks gestures, tones, and facial expressions, and instead relies on creative use of words for communicating emotion. Johnson-Laird and Oatley (1989) have deduced basic emotions by analyzing 590 English words, which describe emotion. Osgood's theory of Semantic Differentiation (Osgood et al., 1957) deals with assigning emotive meanings to words along three dimensions. Osgood et al. performed factor analysis of texts to identify three main factors on which the affective words can be rated. The three factors are: *evaluative factor* (good or bad), *potency factor* (strong or weak), and *activity factor* (active or passive). The evaluative factor carries the strongest relative weight, and several works have focused on it (Kamps and Marx, 2002; Turney and Littman, 2003; and Mullen and Collier, 2004).

Some words convey emotion explicitly, while some other words can be used to convey emotion implicitly depending on the context (Clore et al., 1987). Strapparava and Valitutti (2006) have classified words into '*direct affective words*' (explicit) and '*indirect affective words*' (implicit) categories. My work has utilized both these types of words. The experiments reported in this thesis show that it is important to take into account a variety of emotion-related words for automatic recognition of emotion, including direct and indirect affective words.

2.2 Sentiment Analysis

Sentiment Analysis is a rapidly growing area of research. It focuses on developing automatic systems that can analyze natural language texts to determine the sentiment expressed in them. The word "sentiment" is often used in a wide sense to refer to expressions of subjectivity, opinion, affect, attitude, orientation, feelings, emotions, and tone in the text.

Much of the current work in sentiment analysis has focused on the task of determining the presence of sentiment in the given text, and on determining its valence, that is, the classification of sentiment according to positive or negative orientation. The sentiment classification task is quite often contrasted with that of topic-based categorization (Pang et al., 2002; Whitelaw et al., 2005). Pang et al. (2002) have empirically investigated if sentiment classification can be regarded as a special case of topic-based categorization, with positive and negative sentiment being taken as the two topics. They found that accuracy achieved in the sentiment classification problems is not as high as achieved in the conventional topic-based categorization, despite the different types of features tried for representing text. The results indicate that sentiment classification is comparably more challenging, because while topics are frequently expressed by relevant keywords, sentiment information is often embedded in the text in more subtle ways.

The automatic recognition of sentiment, particularly in large volumes of documents, can have a variety of applications, notably in summarizing popular sentiment about any product or issue. Such information may particularly be of interest to policy-makers, economists and market researchers, political analysts, and social scientists.

One of the earliest areas to attract the attention of researchers in sentiment analysis is that of movie reviews (Pang et al., 2002; Turney, 2002; Turney and Littman, 2003; Pang and Lee, 2004; Whitelaw et al., 2005; Kennedy and Inkpen, 2006). There are many popular movie review sites on the Web. Addition of sentiment information to movie reviews can provide additional information to the readers and increase the popularity of the reviews. It can also help in automatic analysis of movie reviews to determine the viewers' opinions.

Pang et al. (2002) have attempted to classify the movie reviews drawn from the Internet Movie Database (IMDB)[5] into positive and negative categories using machine-learning techniques. Turney (2002) has proposed an unsupervised method of classifying movie reviews into *recommended* and *not recommended* categories based on the semantic orientation of adjectives and adverbs in the review. Kennedy and Inkpen (2006) have

[5] http://reviews.imdb.com/Reviews/

investigated the role of valence shifters in movie review classification. Contextual valence shifters can change the intensity of sentiment or even the sentiment itself according to the context (Polanyi and Zaenen, 2004). The authors have investigated the role of three types of valence shifters – intensifiers, diminishers, and negations – in predicting the sentiment of movie reviews and found that taking valence shifters into account helps improve the classification results.

Pang and Lee (2005) have gone beyond the binary classification of movie reviews to address the task of fine-grained classification of the reviews on a multi-point scale. Their objective is to capture the reviewers' numerical ratings similar to the five-star ratings provided by the review authors. The main challenges cited for this task are the presence of certain degrees of similarities between the labels, which makes categorization into distinct classes difficult; and the misclassification of the borderline cases. This thesis also deals with multiple emotion classes to which the sentences have to be assigned. Here also I found some similarity in classes, as all of them are basically emotion classes. Particularly notable in this regard are two classes – *anger* and *disgust*, which the human annotators found hard to distinguish in many cases.

In another example of multi-class sentiment assignment, Nadeau et al. (2006) address the task of automatic dream sentiment classification on a 4-level scale from neutral to highly negative. In their study, the dreams gathered from a dream bank were manually annotated for use in different classification approaches. The best accuracy achieved was 50%, indicating the difficulty of the task. The baseline for the task was 33% using a majority class guessing approach.

Besides movie reviews, the Web also abounds in reviews for all kinds of products and services ranging from consumer goods to restaurants and vacation spots. In an article in Forbes.com, Hoffman (2005) acknowledges the fact that "successful applications (based on sentiment analysis) could help automate market and product research". Market research needs feedback from the consumers, and the Web provides a convenient and popular public forum from where the feedback can be collected. However, the feedback information is often present in the form of unstructured free-form text on review sites, blogs, and other

discussion forums, and can be useful only if automatic techniques for consumer sentiment analysis are available. Many researchers are involved in development of automatic systems that can aggregate the consumer sentiment from this vast amount of consumer-generated media, classify them into positive or negative, and also track them over time. The work reported in thesis also draws its data from the Web.

Hu and Liu (2004) process customer reviews collected from C|Net[6] and Amazon[7] to detect customers' opinion on different product features. They first find the product features mentioned in the reviews, and then identify the related opinion sentences and their orientation. Finally, results from all the customer reviews are summarized. This summarization differs from conventional topic-based summarization, as it only extracts those sentences in which the customers have commented on the product features. Such product feature summaries can help potential customers to know about the product and make an informed decision. It can also help the manufacturers to know the customers' opinions about various product features.

Mishne and Glance (2006) analyze blog data to show correlation between popular blogger sentiment about movies and corresponding movie sales. They collect relevant blog posts discussing movies from the posts appearing in the BlogPulse[8] Index (Glance et al., 2004), and estimate the blogger sentiment prior to and after the movies' release. Experiments revealed that positive sentiment about movies correlates heavily with the movies' success at the box office.

Recognizing sentiment from blogs is a more challenging task, as the data in blogs tends to be more informal and less focused compared to the data drawn from dedicated review sites (Mishne and Glance, 2006). This task is further complicated by use of informal language and lack of proper structure in blogs (Java et al., 2006; and Mishne, 2005). For a more detailed description of the nature of information present in blogs and of the challenges associated with mining information from the blogosphere, see (Mishne, 2006).

---

[6] http://www.cnet.com/
[7] http://www.amazon.com/
[8] http://www.blogpulse.com/

Zhang et al. (2006) draw attention to the variety of challenges posed by the kind of language used in non-standard text. These include the presence of misspellings, slang, ungrammaticality, abbreviations, and onomatopoeic elements (such as "grrr", "hmm") as well as use of upper case, special punctuation (such as "!!!"), and repetitions of letters or words (such as "sweeeet") for affective emphasis. They apply several preprocessing steps to their corpus to address these challenges. These steps include a look-up table to deal with abbreviations and a small dictionary, containing base forms of certain special words. These resources are used to provide the appropriate replacements for non-standard usage in text. In addition, they have also used two spelling-correction algorithms. Another work, which has discussed the various kinds of noise present in the online text-based communication, is by Sokolova et al. (2005b). They discuss the challenges posed by the presence of noise in the context of the electronic negotiations data. To address the problem, they have used spell-checking methods based on the frequency counts to use the most appropriate replacement for misspelled words.

One particular area of sentiment analysis, which focuses on identifying expression of opinion in the text as well as of their orientation, is called Opinion Mining or Opinion Analysis. Hurst and Nigam (2004) combine shallow NLP techniques and statistical machine learning to find the opinion sentences about the given topics from a corpus of data containing both relevant and irrelevant sentences. Opinion classification has also been studied by Liu et al., (2005), Wilson et al. (2006) and Riloff et al. (2006).

Hiroshima et al. (2006) propose a web search engine that can search for the sentences expressing opinion about a given topic. Search engines can improve the usefulness of results provided by them if these results present a comparison of the different opinions on any topic.

While in movie reviews, product reviews, and blogs, the opinions are most likely to be that of the review author, news articles may contain the opinions of various people. Kim and Hovy (2006) have investigated the task of finding the opinion-holder and the topic of opinion from the online news articles. Mullen and Malouf (2006) have used the posts from a political discussion forum to classify them on the basis of political sentiments. All posts are self-

labeled with the writer's stated political affiliations, and the authors have mapped those affiliations into two broad political classes: right and left.

Some researchers focus on identifying the instances of subjective language in the text as a precursor to finding the sentiment information. Subjective language is used to express opinions, evaluations, and other private states (Wiebe et al, 2004). The motivation behind this approach is that sentiment is often expressed using subjective language, and once the subjective content is sifted out of the text, further techniques can be applied to find the overall orientation, as well as the specific instances of opinion or emotion from it (Hatzivassiloglou and Wiebe, 2000; and Wiebe et al., 2004).

Pang and Lee (2004) demonstrate the effectiveness of extracting subjective portions of text before performing polarity classification of movie reviews. Chesley et al. (2006) have also investigated the problem of subjectivity and polarity classification of blog posts.

2.3 Emotion Recognition

Emotion recognition in text is just one the several dimensions of the task of making the computers make sense of and respond to emotions. The word "affect" is often used interchangeably with "emotion" in the literature. Human emotion can be sensed from such cues as facial expression, gestures, speech and writings. Research in emotion has focused on all these aspects (Cowie et al., 2001). This area of research is collectively also referred to as Affective Computing, based on the work of Picard (1995).

Computational approaches to emotion analysis have focused on various emotion modalities, resulting in a large number of multi-modal emotion-annotated data[9]. However, only limited work has been done in the direction of automatic recognition of emotion in text. In this section, the discussion will be confined to a description of the work done in emotion recognition in written text. Recognition and classification of emotion in text can be regarded as a sub-field of sentiment analysis. It can find useful application in several areas, such as personality analysis and modeling (Liu and Maes, 2004), text-to-speech synthesis (Alm et al., 2005), consumer feedback analysis, Human-Computer Interaction and Affective Interfaces

(Liu et al., 2003), affective tutoring in e-learning applications (Zhang et al., 2006), affective communication systems (Neviarouskaya et al., 2007a), virtual counseling, and in the design of agents that learn user preferences.

Some researchers have studied emotion in a wider framework of *private states* (Quirk et al., 1985). Wiebe et al. (2005) worked on the manual annotation of emotions, opinions, and sentiment in a 10,000-sentence corpus (MPQA corpus[10]) of news articles. They developed a detailed annotation scheme for labeling the expressions and properties of a variety of private states that indicate a person's internal state in the text. They did fine-grained annotation at the word and phrase-level that also included marking the source and target of the private sate expressed. In addition, they marked other related properties, such as intensity, significance, and type of attitude. However, no distinction is made in identifying the type of private state, which means that they have not identified if a particular expression expresses emotion or opinion or other kind of sentiment, such as speculation or evaluation. In labeling the private states, the judges were asked to interpret words in context, but were not asked to look for any particular category of words or part-of-speech. This labeled corpus has been used in a variety of experiments for subjectivity and opinion classification (Riloff et al., 2003; Riloff and Wiebe, 2003), in polarity and intensity classification of individual clauses (Wilson et al., 2005; Wilson et al., 2006), and in Question Answering applications (Stoyanov et al., 2005, Somasundaran et al., 2007). However, I did not find the MPQA corpus suitable for my research as the corpus marks several types of sentiment annotations besides emotions (which are the focus of my work), and that emotions are in no way distinguished from the other types of sentiments.

Expressions of emotions in text have also been studied within the *Appraisal Framework* (Martin and White, 2005), which is a functional theory of the language used for conveying attitudes, judgments and emotions (Taboada and Grieve, 2004; Whitelaw et al., 2005a,b; and Read et al., 2007). Appraisal, in general, defines fine-grained semantic distinctions between different types of sentiments. The Appraisal Framework defines a taxonomy dealing with the

---

[9] http://www.emotion-research.net/wiki/Databases
[10] http://www.cs.pitt.edu/mpqa

meanings in interpersonal communication that consists of three subsystems – *attitude*, *engagement* and *graduation*. The feelings of emotion are covered under the "attitude" subsystem of this framework. Generally, the unit of study in this framework is not individual words, but appraisal groups. The appraisal groups consist of a head adjective defining an attitude type optionally preceded by a list of appraisal modifiers, such as "not really happy", "awfully bad", and "very good". Whitelaw et al., (2005a,b) have used features based on the appraisal groups extracted from text in machine learning algorithms for classification of movie reviews. The results indicate that the features based on the appraisal groups lead to improvement in the classification performance. In their experiments, the authors have built a lexicon of appraisal words and their attributes using semi-automatic methods. This thesis also presents an automatic method of building an emotion lexicon. Use of semantic resources such as specialized lexicons is important to many NLP tasks. Manual creation of such resources is expensive and also lacks consistency and coverage, especially in case of continuously evolving language of online communication. Automatic methods of building lexicons are typically based on similarity measures and can prove useful in many cases.

Read et al. (2007) have carried out appraisal annotation of a corpus of book reviews, a genre that provide ample instances of the various kinds of appraisal classes. They found that out of the three subsystems of the Appraisal Framework, the attitude subsystem's instances (which include emotion expressions) were the easiest to identify. Two human annotators were involved in this work. They were asked to identify spans of text that express appraisal, and also to specify the particular type of appraisal being expressed. The authors also performed an inter-annotator agreement study to look into the reliability of the annotations as well as to assess the difficulty of the task. The metrics chosen for calculating agreement are those that have been earlier used for evaluating Message Understanding Conference (MUC) tasks (Chinchor, 1998). One major point of disagreement between the judges was the choice of lengths in selection of the same appraisal unit. This problem has previously been pointed out by Wiebe et al. (2005) in their annotation project. They have emphasized the need to relax the constraints in measuring overlap, such that intersecting text spans are considered to be matches. The same approach was adopted by Read et al. (2007).

The aforementioned works study emotion under a wider framework along with a variety of other types of sentiments. The focus of this thesis is on learning specific emotions from text. In this context, I will now present a review of the research work done to exclusively recognize expressions of emotions in text.

Alm et al. (2005) have explored the task of automatically classifying sentences in the children's fairy tales according to the basic emotions identified by (Ekman, 1992). They however distinguish between *positively surprised* and *negatively surprised* emotions resulting in two classes instead of one *surprise* class in the original set identified by Ekman. In the preliminary work reported in (Alm et al., 2005), the authors have conducted experiments to classify sentences into emotional versus non-emotional, as well as according to valence into – positive emotion, negative emotion, and no emotion. In the former case, all emotion classes, that is, *happy*, *sad*, *angry*, *disgusted*, *fearful*, *positively surprised* and *negatively surprised* are coalesced into one emotion class. In the latter case, *happy* and *positively surprised* were coalesced into the *positive emotion* class, while *sad*, *angry*, *disgusted*, *fearful*, and *negatively surprised* were coalesced into the *negative emotion* class. The data used in the experiments consists of 22 fairy tales, which were annotated with emotion information at the sentence level. Such an affectively annotated corpus of children's fiction can be later utilized in a text-to-speech synthesis system for the expressive rendering of stories.

In another related work dealing with fiction, Read (2004) has explored the task of classifying the sentences from short stories according to their affective content. The sentences in their fiction corpus were manually annotated with a sentiment tag – *positive*, *negative* or *unclassifiable*, and an affect tag – *high positive affect*, *low positive affect*, *high negative affect*, *low negative affect*, *high pleasantness*, *low pleasantness*, *high engagement*, *low engagement* or *unclassifiable*. The affect tags are based on the Watson and Tellegen's Affect model (Watson and Tellegen, 1985).

Liu et al. (2003) have explored the task of classifying sentences into Ekman's (1992) basic emotion categories. Neviarouskaya et al. (2007b) have also reported on their approach for determining Ekman's basic emotions in the sentences in blog posts.

Mihalcea and Liu (2006) have focused on two particular emotions – *happiness* and *sadness*. They work on blog posts drawn from LiveJournal[11], which are self-annotated by the blog writers with *happy* and *sad* mood labels. They perform linguistic analysis of the text in these blog posts to identify happy and sad words, phrases, and topics. They have derived happiness-factor scores for words in the corpus based on their relative frequency in happy posts. They also apply various semantic analysis techniques to identify the causes of happiness from the text. They perform temporal analysis by investigating the levels of happiness, as indicated by the linguistic indicators in text, at various times in a day and days in a week. Further, the authors utilized the categorization scheme of WordNet (Fellbaum, 1998) to find the *human-centeredness* of words. All words that were subsumed by either the *act, human action, human activity* or the *psychological feature* top-level categories of WordNet through closure over hypernym relations were considered human-centered. It was found that the "sad words" were 50% more human-centered than the "happy words". The latter were found to have more *socialness*, which was judged on the basis of their statistical co-occurrence with socializing phrases such as "with friends", "with family" and "together".

Holzman and Pottenger (2003) have addressed the problem of classifying emotions in online chat conversations. In their approach, they first automatically convert textual chat messages into speech using the Microsoft Speech SDK[12], and then use frequency counts of the phonemes extracted from the speech version of the text messages for ML-based emotion classification. This approach is advantageous in case of chat data, as it is immune to the presence of such noise as misspellings, grammatical errors and abbreviated form of words used in chats. However, its main limitation is the need for voice reconstruction, which may not always be feasible or accurate.

Zhang et al. (2006) have reported on the development of a module that can detect affect in the characters' speeches, which would be used in a program for virtual dramatic improvisation. They use an e-drama corpus for their research work, and aim at identifying a

[11] http://www.livejournal.com/
[12] http://www.microsoft.com/speech

broad spectrum of affect, including the basic and complex emotions, moods, as well as value judgments. The authors have also investigated the use of metaphor for expressing affect.

Rubin et al. (2004) have performed a study of the manual classification of texts drawn from blogs and product reviews on the basis of Circumplex Theory of Affect (Watson and Tellegen, 1985). Their study has also identified the commonly agreed upon linguistic clues related to each affect category. The affect categories studied in their experiments reflect the eight octants identified in the Circumplex Theory. Their work thus provides empirical data in support of the correlation between emotion and linguistic clues. Owsley et al. (2006) have proposed automatic techniques for affective classification of the blog posts belonging to specific domains (e.g, movies, politics, etc.) into positive and negative affect categories. In their work, they have also identified the domain-specific emotional adjectives that can be used as features in the classification experiments.

Mihalcea and Strapparava (2006) present results in favor of automatic recognition of humor in texts. They perform experiments to identify humorous one-liners, which are one-line sentences generally characterized by simple syntax and use of rhetoric which gives them a humorous connotation. Mishne (2005) has attempted to classify LiveJournal[13] blogposts according to the mood indicated by the writers while writing the blog posts. The possible number of mood annotations is unlimited as blog writers can choose from one of the given 132 moods or enter their own. In the experiments reported in (Mishne, 2005), the study was limited to the top 40 moods in the corpus.

Linguistic analysis of personal texts can reveal a variety of information about their authors. In (Liu and Maes, 2004), an automatic system is introduced that can generate models of people's attitudes and opinions by automated analysis of their personal texts. These texts include blogs, emails, editorial papers, and transcribed speeches. Attitude is a result of a person's affective reaction to concepts, topics, and situations. This model analyzes people's personal texts to assess the affect of each sentence, and uses this information in determining their attitude. The model has been further utilized with other approaches in (Liu, 2006) for

13 http://www.livejournal.com/

computing people's *point-of-view*, based on the models of their attitudes, judgment, and cultural tastes.

Some researchers have investigated the techniques for visualizing the affective content in the text. Such techniques can be successfully used in developing intelligent interfaces. For example, (Liu et al., 2003) map the affective content of a document to a color bar, such that the changing colors in the bar represent the progression of affect through the document. An affect-to-color correspondence scheme is used to encode the six basic emotion categories defined by Ekman (1992). Furthermore, the color bar is hyperlinked to allow affect-based navigation of the document.

Neviarouskaya et al. (2007a) propose a system for augmenting online conversations with a graphical representation (*avatar*) of the user, which displays emotions and social behavior in accordance with the text. This system performs automatic estimation of affect in text on the basis of symbolic cues such as emoticons, popularly used IM (Instant Messaging) abbreviations, as well as word-, phrase-, and sentence-level analysis of text. Such a system can help improve the experience of online social interactivity by allowing expression of emotion in real-time online conversations.

2.4 Sentiment Classification Methods and Resources

A variety of approaches have been taken to sentiment classification. Most researchers have approached the problem of sentiment classification as a kind of text classification (Pang et al, 2002; Mihalcea and Strapparava, 2005).

The prevalent approach to sentiment classification is based on the premise that the overall sentiment of a document is the aggregate of the sentiment of the words comprising it. These techniques therefore look for the presence of appropriate affect words in text. Some words are quite unambiguously affect words, while others convey affect to some degree. This method either uses a corpus-driven approach to assign affective orientation or scores to words, or it relies on some existing affect lexicons.

Turney (2002) and Turney and Littman (2003) use unsupervised methods to classify movie reviews based on the similarity of the phrases in the review to the words "excellent" and

"poor". Turney (2002) has classified movie reviews by first extracting all the phrases in the review that contain adjectives or adverbs. Then, the semantic orientation of each extracted phrase is estimated by comparing the similarity of the phrase to a positive reference word ("excellent") with its similarity to a negative reference word ("poor"). This method is based on the PMI-IR algorithm (Turney, 2001), which uses the AltaVista Advanced Search Engine to study the statistical associations (based on co-occurrences) between words using the Web as a corpus. Finally, the review is classified on the basis of the average semantic orientation of all phrases in the review.

Read (2004) has also utilized the PMI-IR algorithm for detecting the affective orientation of sentences. The method called as AO-PMI-IR (Affective Orientation from Pointwise Mutual Information using Information Retrieval) first selects the bigram patterns from the input sentences that convey affective information, as is done in (Turney, 2002). The affective orientation of these patterns is then determined based on their statistical co-occurrence with the words representing the opposite extremes of each dimension in the Watson and Tellegen's Theory of Affect (Watson and Tellegen, 1985). The corpus used for calculating the co-occurrences in this case is the Waterloo MultiText System (WMTS) (Turney, 2004).

Another corpus-driven method of determining the emotional affinity of words is to learn their probabilistic affective scores from large corpora. Mihalcea and Liu (2006) have used this method to assign a *happiness factor* to words depending on the frequency of their occurrences in happy-labeled blogposts compared to their total frequency in a corpus containing blogposts labeled with "happy" and "sad" mood annotations. They also compare the happiness factor scores of words with the scores in the ANEW list (Bradley and Lang, 1999). The ANEW list, prepared on the basis of psychological experiments, assigns scores to words along the three dimensions of affect in the PAD model (Mehrabian, 1995). These dimensions are: *Pleasure/displeasure*, *Arousal/non-arousal*, and *Dominance/submissiveness*. A weak correlation was found between the *happiness factors* learned from the corpus and pleasure scores of the ANEW list. The authors interpret this as a result of the considerable difference between the ideal experimental setups (as in the ANEW list) and the real-world description of emotions (as in blogs). Furthermore, many words from the corpus with a high

happiness factor did not even appear in the ANEW list. This was taken as a reference to the difference in the private feelings of happiness (as in blogs) from what is commonly perceived as the public displays of happiness in the social circles (as in the ANEW list). In their binary classification experiments, the authors achieve 79% accuracy, using a Naïve Bayes Classifier trained with the unigram features. Bi-gram and tri-gram features gave similar accuracy results.

Other approaches have used lexical resources such as WordNet (Fellbaum, 1998) to automatically acquire emotion-related words for emotion classification experiments. Starting from a set of primary emotion adjectives, Alm et al. (2005) retrieve similar words from WordNet utilizing all senses of all words in the synsets that contain the emotion adjectives. They also exploit the synonym and hyponym relations in WordNet to manually find words similar to nominal emotion words.

Kamps and Marx (2002) and Hu and Liu (2004) have used the synset relations in WordNet to derive sets of affective adjectives. (Kamps and Marx, 2002) assign scores to adjectives based on their synonym depth in WordNet from two defining polar synonyms, such as *good* and *bad.*

Kennedy and Inkpen (2006) have used another lexical resource, namely, the General Inquirer (Stone et al., 1966) to find positive and negative terms in a review. These terms are then used in combination with contextual valence shifters for unsupervised review classification.

Strapparava et al. (2007) have used WordNet-Affect (Strapparava and Valitutti, 2004), an affective extension of WordNet in their experiments to automatically detect emotion in text. They use five of the six basic emotional categories described by Ekman (1992). For directly affective words, they use weights from WordNet-Affect. However, for indirectly affective words, their approach is to assign them affective weights based on their semantic similarity to an emotional category. The affective weights are automatically acquired from a very large text corpus in an unsupervised fashion. The semantic similarity is measured using a variant of Latent Semantic Analysis (LSA) technique (Deerwester et al., 1990).

The approach of using sentiment orientation of constituting words to determine the overall sentiment of the document suffers from drawbacks, as it relies only on superficial features, whereas sentiment is often communicated through the composite meaning of the text, rather than exclusively through the use of affect words.

An alternative approach adopted by many researchers takes into account a variety of features for training machine learning algorithms. Machine learning algorithms can automatically learn various rules to characterize the different classes. Though these methods also lack a semantic underpinning, they work reasonably well when trained with sufficient amounts of sentiment-annotated data (Liu et al., 2003).

Pang et al. (2002) have used n-gram and part-of-speech information in their feature set. They have tested three machine-learning techniques, namely, Naive Bayes, Maximum Entropy classification, and SVM learning algorithms, and found SVM to give the best performance.

Owsley et al. (2006) use adjectives as training features in Naïve Bayes classifiers. In their machine learning experiments, they characterize each document with a vector representing the number of times each adjective feature occurred in the document. They also compare adjectives from their corpus to words in the ANEW list (Bradley and Lang, 1999). A weak correlation between the two indicates that a general-purpose affective lexicon (such as ANEW) may not be sufficient for domain-specific classification.

In their machine learning classification experiments using Linear Regression and Naïve Bayes classifier, Nadeau et al. (2006) have used frequency-based features calculated using resources such as the General Inquirer (Stone et al., 1966), the Linguistic Inquiry and Word Count (Pennebaker et al., 2001), and the adjective lexicon used in (Hatzivassiloglou and McKeown, 1997).

Liu et al. (2003) use a novel approach of utilizing real-world knowledge about affect drawn from the common-sense knowledge base, OMCS (Singh et al., 2002). Their approach aims at understanding the semantics of text to identify emotions at the sentence level. They begin with the extraction of those sentences from OMCS which contain some affective information. An affective model of text is then built that characterizes each sentence with a

six-tuple corresponding to Ekman's six basic emotions (Ekman, 1992). They also apply various smoothing techniques to ensure continuity in transition of emotions from one sentence to next.

Machine learning approaches require affect-annotated data for training purpose. As often the corpus under analysis lacks this kind of sentiment information, most of the research work has to make use of manual labeling. The agreement between judges is generally measured using Kappa statistic (Cohen, 1960; Fleiss, 1981). Wilson and Wiebe (2003) have developed a detailed annotation scheme for labeling sentences as objective and subjective, and also for positive and negative labeling of subjective sentences.

Rubin et al. (2004) involve human judges in the classification of online product reviews and blogs into eight categories of emotion. The unit of text to be classified is a segment ranging from 2 to 20 sentences. The judges were also required to identify textual clues within segments related to the chosen emotion category and intensity. The length of these clues ranged from a few words to a sentence. The motivation behind identifying these clues was that they could be used in future experiments for automatic emotion classification. The clues chosen by different respondents were found to have similar keywords and length.

Chesley et al. (2006) have used subjective texts manually labeled into positive and negative. In their experiments, subjective texts were drawn from newspaper columns, letters-to-editors, reviews, and blogs, while texts drawn from news, health, business, and technology sites were considered as objective.

Hiroshima et al. (2006) use a corpus, which was manually annotated for opinion sentences and subjective clues, for training an SVM-based machine learning classifier. Hu and Liu (2004) manually annotate descriptions of product features in customer reviews for training classifiers.

Read (2005) has performed sentiment analysis experiments on a dataset drawn from newsgroup messages, and labeled with smileys or emoticons. No manual labeling of affect information is required in this case as these labels are provided by the writers of the message themselves.

Mihalcea and Strapparava (2006) have collected positive and negative examples from the Web to train their humor classifier. Positive examples consist of humorous one-liners, which were collected using automatic bootstrapping process, beginning with a short seed list of manually identified one-liners. The negative examples consist of news titles, proverbs, and sentences from the BNC. A main consideration in selecting negative examples was to make them structurally and syntactically similar to positive examples.

Besides document-level analysis, sentiment in text has been evaluated at lower levels of granularity as well. Turney and Littman (2003), Hatzivassiloglou and Wiebe (2000), and Wiebe et al. (2004) have investigated sentiment orientation at word and phrase level. Studying smaller text units can provide insight into the kinds of linguistic clues and features that would be relevant for higher-level classification experiments.

Hurst and Nigam (2004) emphasize that document-level sentiment analysis is more suitable for documents that address a single topic, such as movie or product reviews. The assumption here is that all expressed sentiment in the document is about the topic. But this approach may not generalize well to other domains. So, sometimes, it may be more useful to do a fine-grain sentiment analysis at sentence and expression level.

2.5 Conclusion

A great body of work exists in the field of sentiment analysis. The work done in this area includes distinguishing subjective portions in text, identifying the presence of sentiment, finding sentiment orientation and, in few cases, determining fine-grained distinctions in sentiment, such as emotion and appraisal types. Work exclusively on emotion detection is comparatively rare and lacks empirical evaluation.

Previous work has investigated the application of sentiment detection methods to a variety of text genres, including product and movie reviews, news stories, editorials and opinion articles and, more recently, blogs.

Chapter 3
Data Acquisition and Annotation

Data is a precious thing and will last longer than the systems themselves.
***Tim Berners-Lee** (1955-)*

In this chapter, I describe the whole process of data selection and annotation adopted in this study, beginning with an introduction to the data in Section 3.1. The annotation scheme is presented in Section 3.2, while details of the data prepared using this scheme appear in Section 3.3. The measures of annotation agreement employed in this study and the agreements found among judges are discussed in Section 3.4. The chapter ends with concluding remarks in Section 3.5.

3.1 Data Selection

The objective of this research is to be able to automatically recognize emotions from text. This calls for an appropriate corpus of text that can be used in emotion recognition experiments. For training machine learning systems and for the evaluation of any automatic learning system, it is pre-requisite to have an annotated data. Research in automatic text-based emotion analysis is hampered by the unavailability in the public domain of emotion-annotated data for written text. A survey of the existing corpora revealed that none was appropriate for the type of emotion learning addressed in my research. I considered the MPQA[14] Opinion Corpus (Wiebe et al., 2005) consisting of 530 news articles with elaborate

[14] http://www.cs.pitt.edu/mpqa/

opinion and emotion annotations. But I found it unsuitable for my work as it marks several types of sentiment and private state annotations besides emotions, and that emotions are in no way distinguished from the other types of sentiments. The Fifty Word Fiction[15] Corpus used by Read (2004) labels sentences according to *valence*, *pleasantness*, and *engagement* as defined in the affect theory of Watson and Tellegen (1985). This was not suitable for my research, which is aimed at identifying the six basic emotion categories defined by Ekman (1992). Alm et al. (2005) have reported on the development of a corpus of children's fairy tales, annotated at the sentence level with Ekman's six basic emotions; but the corpus is not yet available in the public domain. Thus, in light of this survey of the existing corpora, the initial focus of my work was on selecting the appropriate data for my research and on labeling the data with emotion annotations at several levels. An evident benefit of the latter task was that it provided an opportunity for keen investigation of the expressions of emotions in language. The annotated corpus will be eventually made publicly available.

The primary consideration in the selection of data for my research was that the data should be rich in emotion expressions so as to permit numerous learning instances. Another essential consideration was that the data should comprise ample instances of all the emotion categories considered in this research. Such data can be found in personal texts such as diaries, emails, and blogs, and in narrative texts such as fiction. While much of the personal texts are proprietary and not available publicly, blogs are an exception. Blogs are online personal journals containing owner's reflections and comments. The narratives present in blogs cover everyday topics ranging from home to school to work, as well as intimate musings on personal relationships and global issues. They make good candidate for emotion study, as they are likely to be rich in emotion content. Furthermore, with blogs proliferating on the Web, there is a potential to find examples of all the types of emotion categories for investigation in this research.

A popular genre for emotion study in previous works is fiction (Read, 2004; and Alm et al., 2005), as emotions are integral to most of the story plots. However, blogs were chosen as

[15] http://www.informatics.sussex.ac.uk/users/jlr24/data/fwf-corpus.zip

the data source for this research as blogs offer real-world examples of emotion expression in text, unlike fiction. And most applications of automatic emotion recognition deal with real-world text – often containing noise, such as misspellings and slang, for instance, in affective interfaces and communication systems (Liu and Maes, 2004; Neviarouskaya et al., 2007a). Another consideration in selecting blog text was that such text does not conform to the style of any particular genre *per se*, thus offering a variety in writing styles, choice and combination of words, as well as topics. Thus, the methods learned for discerning emotions using the blog data would be general and widely applicable rather than genre-specific.

The blog data was collected from the Web in the following manner. First, a set of seed words was identified for each of the six emotion categories. (Appendix A contains the lists of these seed words). In preparing the set, I took words that are commonly used in the context of a particular emotion. Thus, I chose words such as "happy", "enjoy", "pleased" as seed words for the *happiness* category; "afraid", "scared", "panic" for the *fear* category, and so on. Next, the seed words for each category were fed to the blog search engine, BlogPulse[16] and blog posts containing one or more of those words were retrieved. A total of 173 blog posts were collected in this manner. A sample excerpt from this data is shown in Figure 3.1.

3.2 Annotation Scheme

The primary goal behind the annotation task was to identify the emotional affinity of sentences. While most work in sentiment analysis focuses on document-level analysis, this research has focused on the sentence-level analysis for learning emotions in text. The main consideration behind this decision was that there is often a dynamic progression of emotions in the narrative texts found in fiction, as well as in the conversation texts and blogs. It is

[16] http://blogpulse.com/

Today was just kind of stressful.

My night last night turned out amazing towards 8:00-9:00.

I had a really nice conversation with Jack*, another amazing one with Pat*, and then another one with Sid*. It was definitely a good night of conversation, although I did stay up until 12:30, and I payed for that this morning in 1st period (Physics)

But all of a sudden today it's hit me that I have all this work due. But it's not, "Oh this is due tomorrow.", but more like, "Okay, this is due next friday." I have so much long-term work that I'm not sure if I'll get it all finished with band practice every night, a football game friday, and a competition all day Saturday.

I'm a little stressed out.

But, tonight I'm attempting to get a good portion of it finished. I'm starting my argument on illegal immigration (which I may say is turning out fantastic with the first two praragraphs I have), and I finished one of my US History charts, and I plan on outlining a bit once I start watching America's Next Top Model.

Figure 3.1 A sample of the blog data used in this research

Today was just kind of stressful. (*sadness*)

My night last night turned out amazing towards 8:00-9:00. (*happiness*)

I had a really nice conversation with Jack*, another amazing one with Pat*, and then another one with Sid*. (*happiness*)

It was definitely a good night of conversation, although I did stay up until 12:30, and I payed for that this morning in 1st period (Physics) (*mixed emotion*)

But all of a sudden today it's hit me that I have all this work due. (*surprise*)

Figure 3.2 An example of dynamic progression of emotion across sentences

* Names changed

quite likely that differing emotions are expressed in consecutive sentences of a passage or a conversation. See, for example, the sentences shown in Figure 3.2.

For each sentence in the corpus, the annotators were required to label the predominant emotion category in the sentence, the intensity level of the emotion, as well as the emotion indicators in the sentence. A detailed description of the annotation scheme adopted in this research is given below.

Emotion Category

The annotation scheme required each sentence to be labeled with the appropriate emotion category, which best described its affective content. The possible emotion categories included Ekman's six basic emotions: *happiness*, *sadness*, *anger*, *disgust*, *surprise*, and *fear*. An additional category, *no emotion*, was added to account for those sentences that had no emotion content in them.

The initial annotation effort suggested that in many cases, a sentence was found to exhibit more than one emotion – consider the sentence in Figure 3.3, for example, which carries the emotions of both *happiness* and *surprise*.

Everything from trying to order a baguette in the morning to asking directions or talking to cabbies, we were always pleasantly surprised at how open and welcoming they were.

Figure 3.3 An example of mixed emotions

Similarly, the sentence in Figure 3.4 shows how more than one type of emotion can be present in a sentence in which the emotional states of more than one person are referred to.

I felt bored and wanted to leave at intermission, but my wife was really enjoying it, so we stayed.

Figure 3.4 Another example of more than one emotion in a sentence

Furthermore, it was found that the emotion conveyed in some sentences could not be attributed to any of the basic categories; see for example the sentence in Figure 3.5.

It's like everything everywhere is going crazy, so we don't go out any more.

Figure 3.5 An example of a sentence with no clear emotion

In view of these cases, it was decided to have an additional category called *mixed emotion* to account for all such instances. Thus, in the final annotation scheme, there were eight emotion categories to select from for each sentence. Table 3.1 lists all the emotion categories and their labels.

Table 3.1 Emotion Categories used in Annotation

Emotion Category	Label
Happiness	hp
Sadness	sd
Anger	ag
Disgust	dg
Surprise	sp
Fear	fr
Mixed emotion	me
No emotion	ne

Emotion Intensity

The second part of the annotation scheme involved assigning an emotion intensity label to all the emotion sentences in the corpus, regardless of the emotion category assigned to them. The intensity label was assigned to only those sentences which expressed some type of

emotion; no intensity label was assigned to the *no emotion* sentences. Table 3.2 shows the emotion intensities and their corresponding labels used in the corpus.

Table 3.2 Emotion Intensities used in Annotation

Emotion Intensity	Label
High	h
Medium	m
Low	l

A study of emotion intensity is significant as it can help recognize the linguistic choices writers make to modify the strength of their expressions of emotion. The knowledge of emotion intensity can help locate the highly emotional snippets in text, which can then be further analyzed to identify emotional topics in the text. The intensity values can also help distinguish the borderline cases of emotion from the clear cases (Wiebe et al., 2005), as the latter will generally have higher intensity.

Emotion Indicators

Besides labeling the emotion category and the emotion intensity at the sentence level, the secondary objective of the annotation scheme was to identify the spans of text (individual words or strings of consecutive words) in each sentence that convey emotional content. They are called emotion indicators. Knowing them could help identify a broad range of affect-bearing lexical tokens and possibly, syntactic phrases. The annotation scheme allows any number of emotion indicators of any length in a sentence.

Several annotation schemes were considered for labeling the emotion indicators in a sentence. Initially, I thought to identify only individual words for this purpose. That would have simplified calculating the agreement between the annotation sets by the different human annotators. However, I soon realized that individual words might not be sufficient for this purpose. Emotion is often conveyed by longer units of text or by phrases; consider for

example, the expressions "can't believe" and "blissfully unaware" in the sentence shown in Figure 3.6.

I can't believe this went on for so long, and we were blissfully unaware of it.

Figure 3.6 Examples of emotion indicator spans

The identification of longer units of text as emotion indicators in a sentence would also allow the study of the various linguistic features that serve to emphasize or modify emotion in language, as the use of word "blissfully" in the sentence in Figure 3.6 and "little" in the sentence shown in Figure 3.7.

The news brought them little happiness.

Figure 3.7 Example of emotion modifiers in emotion indicator spans

The drawback of allowing longer units of text as emotion indicators, however, is that makes the task of agreement calculation more complex. In this case, the human annotators are more likely to mark different spans of text even while identifying essentially the same indicators, with the disagreement in the placement of the boundaries for the spans of text considered as emotion indicators. Figure 3.8 illustrates this kind of disagreement.

Today was just kind of stressful.
Today was just kind of stressful.

Figure 3.8 Disagreement in marking emotion indicator spans

3.3 Data Description

The data used in this research comprises 173 blog posts containing a total of 5205 sentences. As a separate list of seed words was used for each category, the resulting corpus

consists of six datasets, one for each emotion category. Table 3.3 gives the details of these datasets. It is important to note that blog posts collected using the seed words for a category do not necessarily contain emotions of that particular category only. Rather, that emotion category is likely to be the dominant emotion type in that dataset. The objective was to have ample emotion instances of all the types available in the corpus. A sample of the annotated blog data appears in Figure 3.9.

Table 3.3 Details of the datasets in the corpus

Dataset	Number of posts	Number of sentences	collected using seed words for
Ec-hp	34	848	*happiness*
Ec-sd	30	884	*sadness*
Ec-ag	26	883	*anger*
Ec-dg	21	882	*disgust*
Ec-sp	31	847	*surprise*
Ec-fr	31	861	*Fear*
Total	173	5205	

Today was just kind of stressful. (*sd, m*)

My night last night turned out amazing towards 8:00-9:00. (*hp, m*)

I had a really nice conversation with Jack*, another amazing one with Pat*, and then another one with Sid*. (*hp, h*)

It was definitely a good night of conversation, although I did stay up until 12:30, and I payed for that this morning in 1st period (Physics) (*me, m*)

But all of a sudden today it's hit me that I have all this work due. (*sp, h*)

But it's not, "Oh this is due tomorrow.", but more like, "Okay, this is due next friday." (*sp, l*)

Figure 3.9 Sample annotated data from the corpus

The blog data retrieved from the web was stripped of HTML markup and the blog posts were split into sentences[17]. The data was then manually labeled with emotion information. As identifying emotion in text is a subjective task, for any emotion labeling to be considered reliable, it is imperative to have more than one judgment for each label. To this end, four human judges were involved in the manual annotation process, and two sets of independently performed annotations were produced for the corpus. I produced the first set of annotations, while three other judges shared the task of producing the second set. The annotators received no training, though they were given instructions and samples of annotated sentences to illustrate the kind of annotations required. (See Appendix B for the Annotator Instructions.)

The annotations in these two sets are described below:

Emotion Categories

Table 3.4 provides a description of the distribution of emotion categories in the two sets of annotations. As shown in the table, the *no emotion* category was found to be most frequent. It is important to have no-emotion sentences, as both positive and negative examples are required to train any automatic emotion recognition system. It should also be noted that in both sets, a significant number of sentences were assigned to the *mixed emotion* category, justifying its addition in the first place.

Emotion Intensity

The second kind of annotations involved assigning emotion intensity (high, medium, or low) to all emotion sentences in the corpus. Table 3.5 shows the distribution of emotion intensities in the two sets of annotations. (No intensity label was assigned to the *no emotion* category sentences.)

[17] I used the module Lingua::EN::Sentence for this purpose. It s available at CSPAN (http://search.cpan.org/)

Table 3.4 Distribution of Emotion Categories in Annotation Sets A1 and A2

Emotion Category	A1	A2
Happiness	613	718
Sadness	214	280
Anger	219	291
Disgust	257	232
Surprise	177	191
Fear	139	151
Mixed emotion	341	387
No emotion	3245	2955
TOTAL	5205	5205

Table 3.5 Distribution of Emotion Intensity in Annotation Sets A1 and A2

Emotion Intensity	A1	A2
High	872	799
Medium	719	889
Low	369	562
TOTAL	1960	2250

Emotion Indicators

Table 3.6 shows some the most frequent emotion indicators identified by the annotators.

3.4 Annotation Agreement Measurement

The interpretation of sentiment information in text is highly subjective. As a result, annotations performed by different human judges have both commonalities and disparities. Difference in skills and focus of the judges, and ambiguity in the annotation guidelines and in the annotation task itself also contribute to disagreement between the judges (Passonneau, 2006). The inter-annotator agreement study seeks to find how much the judges agree in

assigning a particular annotation by using relevant metrics that quantify these agreements. In the following sections, I describe the inter-annotator agreement measurements for the emotion annotation scheme applied in this study. In each case, the agreement was measured for the three pairs, A↔B, A↔C, and A↔D, representing the pairs of annotators who worked on the same sections of the data.

Table 3.6 Most frequent Emotion Indicators in the data

Happiness	Sadness	Anger	Disgust	Surprise	Fear
love	hurt	fucking	hate	surprised	afraid
lol	miss	angry	dislike	amazing	scared
fun	sorry	bitch	sucks	incredible	nervous
good	bad	furious	shit	wonder	worry
happy	sad	annoyed	stupid	unexpected	security
nice	lost	pissed	fucking	can't believe	fear
awesome	cry	yelling	disgusting	weird	what if
funny	stress	upset	crap	suddenly	threat
great	wept	mad	bitch	odd	freak
excited	longing	shut up	sick	strange	dangerous

Emotion Categories

First I measured how much the annotators agree on distinguishing emotion sentences from non-emotion sentences. For this calculation, all emotion classes, namely, *happiness*, *sadness*, *anger*, *disgust*, *surprise*, *fear*, and *mixed-emotion* were coalesced into one class – *emotion*. The other class was the *no-emotion* class. I chose Cohen's kappa (1960) for calculating the agreement. Kappa is popularly used to compare the extent of consensus between judges in classifying items into known mutually exclusive categories. Table 3.7 shows the pair-wise agreement between the annotators on emotion/non-emotion labeling of the sentences in the corpus. The average inter-annotator agreement was kappa = 0.76.

Table 3.7 Pairwise agreement in emotion/non-emotion labeling

	A↔B	A↔C	A↔D	Average
kappa	0.73	0.84	0.71	0.76

Next, I calculated the agreement among the judges on fine-grained labeling of emotion sentences. Within the emotion sentences, there are seven possible categories of emotion to which a sentence can be assigned. Table 3.8 shows the value of kappa for each of these emotion categories for each annotator pair. The values vary a lot across the categories. The agreement was found to be highest for *fear* and *happiness* categories. From this, it may be surmised that writers express these emotions in more explicit and unambiguous terms, which makes them easy to identify. The *mixed emotion* category showed least agreement. This was expected given the fact that this category was added to account for the sentences which had more than one emotions, or which would not fit into any of the six basic emotion categories.

Table 3.8 Pairwise agreement in emotion categories

Category	A↔B	A↔C	A↔D	average
happiness	0.76	0.84	0.71	0.77
sadness	0.68	0.79	0.56	0.68
anger	0.62	0.76	0.59	0.66
disgust	0.64	0.62	0.74	0.67
surprise	0.61	0.72	0.48	0.60
fear	0.78	0.80	0.78	0.79
mixed emotion	0.24	0.61	0.44	0.43

Emotion Intensity

Agreement on emotion intensities was also be measured using kappa, as there are distinct categories – *high*, *medium*, and *low* in this case. Table 3.9 shows the values of inter-annotator

agreement in terms of kappa for each of the emotion intensity levels. The agreement varies a lot depending on the perceived intensity of the sentence. It was found that the judges agreed more when the emotion intensity was high, and this agreement declined with decrease in the intensity of emotion. A major factor in disagreement is the fact that what one judge perceives as a low-intensity sentence, another judge may consider a no-emotion sentence.

Table 3.9 Pairwise agreement in emotion intensities

Intensity	A↔B	A↔C	A↔D	average
High (h)	0.69	0.82	0.65	0.72
Medium (m)	0.39	0.61	0.38	0.46
Low (l)	0.31	0.50	0.29	0.37

Emotion Indicators

Emotion indicators consist of individual words or strings of words selected by annotators as indicators of emotion in a sentence. Since there are no predefined categories in this case, kappa cannot be used to calculate the agreement between judges. In this case, the agreement is to be calculated for the sets of text spans selected by the two judges for each sentence. The text spans marked by one judge may be a subset of, or overlap with, or be disjoint with those marked by the second judge.

Several methods of measuring agreement between sets have been proposed in the literature. In this study, I chose two measures of calculating this agreement - the Measure of Agreement on Set-valued Items (MASI) and the IO (In-Out) Method. Each of these methods is described below.

Measure of Agreement on Set-valued Items (MASI)

A popularly used metric for calculating similarity between sets is the Jaccard metric (Jaccard, 1908), which is defined as the ratio of the cardinality of the intersection to that of the union of the two sets. MASI extends the Jaccard metric by introducing a monotonicity factor.

Figure 3.10 shows how MASI is calculated for two sets A and B. It is defined as a distance between two sets; its value is 1 for identical sets, and 0 for disjoint sets. MASI has been previously used for measuring agreement on co-reference annotation (Passonneau, 2004) and for evaluation of automatic summarization (Passonneau, 2006).

MASI = J * M

where the Jaccard metric is

$J = |A \cap B| / |A \cup B|$

and monotonicity is

$$M = \begin{cases} 1, if\ A = B \\ 2/3, if\ A \subset B\ or\ B \subset A \\ 1/3, if\ A \cap B \neq \phi, A - B \neq \phi, and\ B - A \neq \phi \\ 0, if\ A \cap B = \phi \end{cases}$$

Figure 3.10 Calculation of MASI on sets A and B

The monotonicity factor in the definition of MASI can be explained as follows. If one set is monotonic with respect to another, it indicates that one set's elements always match those of the other set – for instance, in annotation sets {*crappy*} and {*crappy*, *best*} for the sentence shown in Figure 3.11. However, in non-monotonic sets, as in {*crappy*, *relationship*} and {*crappy, best*}, there are elements in each set ("*relationship*" in the first set and "*best*" in the second set) that are not contained in the other set, indicating a greater degree of disagreement. The presence of the monotonicity factor in MASI therefore ensures that the latter cases are penalized more heavily than the former.

We've both had our share of crappy relationship, and are now trying to be the best we can for each other.

Figure 3.11 Sample sentence to illustrate agreement measurement using MASI

While looking for emotion indicators in a sentence, often it is likely that the judges may identify the same expression but differ in marking text span boundaries. For example in the sentence shown in Figure 3.11, one annotators could identify "*crappy*" as the emotion indicator, while the other could identify "*crappy relationship*", both of which essentially refer to the same item, but disagree on the placement of the span boundary. This leads to strings of varying lengths. To simplify the agreement measurement using MASI, I split all strings into words to ensure that members of the set are all individual words. MASI was then calculated for each pair of annotations for all sentences in the corpus. In Table 3.10, I report the values of MASI for each annotator pair. The average over all values was found to be 0.61.

Table 3.10 Pairwise agreement in emotion indicators (using MASI)

Metric	A↔B	A↔C	A↔C	Average
MASI	0.59	0.66	0.59	0.61

IO (In-Out) Method

The second method I adopted to measure agreement between the emotion indicators is the IO method. This method is a variant of the IOB encoding (Ramshaw and Marcus, 1995) used in text chunking and named entity recognition tasks. I used IO encoding, in which each word in the sentence is labeled as being either **I**n or **O**utside an emotion indicator text span, as illustrated in the sentence in Figure 3.12.

Sorry/**I** for/**O** the/**O** ranting/**I** post/**O**, but/**O** I/**O** am/**O** just/**O** really/**I** annoyed/**I**.

Figure 3.12 Sample sentence to illustrate the IO method of agreement measurement

Binary IO labeling of each word in a way reduces the sentence-level task of emotion labeling to that of word-level classification into non-emotion and emotion indicator categories. It follows that kappa can now be used for measuring agreement; pair-wise kappa values using this method are shown in Table 3.11. The average kappa value of 0.66 is lower than that observed for sentence level classification. This is in line with the common observation that agreement on lower levels of granularity is generally found to be lower.

Table 3.11 Pairwise agreement in emotion indicators (using kappa)

Metric	**A↔B**	**A↔C**	**A↔C**	**average**
Kappa	0.61	0.73	0.65	0.66

3.5 Conclusion

In this chapter, I described the process of and the decisions behind the data selection for this research. Next, I introduced the annotation scheme for adding different kinds of emotion-related information to the data. These include emotion categories – *happiness*, *sadness*, *anger*, *disgust*, *surprise*, *fear*, *mixed-emotion*, and *no-emotion*; emotion intensity levels – *high*, *medium*, and *low*, as well as emotion indictor words or phrases for each sentence.

The results of the annotation agreement study show variation in agreement among the human judges. Among all emotion categories, I found that the annotators tend to agree most in identifying instances of *fear* and *happiness*. Among intensity levels, I found that agreement on sentences with *high* emotion intensity surpassed that on the sentences with *medium* and *low* intensity. Furthermore, finding emotion indicators in a sentence was found to be a hard task, with judges disagreeing in identifying precisely the spans of text that indicate emotion in a sentence.

Chapter 4
Differentiating Emotion Sentences from Non-emotion Sentences

Each problem that I solved became a rule,
which served afterwards to solve other problems.
***René Descartes** (1596-1650)*

While the long-term goal of this research is fine-grained automatic classification of sentences on the basis of the emotion type expressed in them, many applications can benefit from high-level differentiation of emotion and non-emotion sentences, regardless of the type of emotion. In this chapter, I describe the experiments conducted to this end.

Section 4.1 describes the decisions that went into selecting the features to represent sentences in the ML-based classification experiments, Section 4.2 presents the experimental setup and classification results, while the results are discussed in Section 4.3. The chapter ends with conclusions in Section 4.4.

4.1 Defining the Feature Set

In defining the feature set for automatic classification of emotional sentences, I was looking for those features, which distinctly characterize emotional expressions, but are not likely to be found in the non-emotional ones. The most appropriate features that distinguish emotional expressions from non-emotional ones are the obvious emotion words present in the former. By obvious emotion words, I mean those words that are quite unambiguously affective. To

recognize these words in the sentence, I utilized two publicly available lexical resources – the General Inquirer (Stone et al., 1966) and WordNet-Affect (Strapparava and Valitutti, 2004).

Features from General Inquirer

The General Inquirer (GI) is a useful resource for content analysis of text. It has been previously utilized as a source of external knowledge in sentiment classification experiments (Kennedy and Inkpen, 2006; Nadeau et al., 2006; Andreevskaia, et al., 2007). The GI consists of words drawn from several dictionaries that are grouped into various semantic categories. It lists the different senses of a term and for each sense it provides several tags indicating the different semantic categories it belongs to. For example, see Figure 4.1 for different senses of some emotion words. GI includes the obvious emotion words such as "afraid" and "glad", as well as the words idiomatically used for affective expression, such as "fed" in "fed up" for expressing *disgust*. As shown in the Figure 4.1, the latter category of words generally comprises the secondary senses of the words.

The point of interest in the GI entries are the emotion-related tags, which can lead to the identification of appropriate semantic categories that can be used as features in ML-based experiments. The most common tags are found to be *EMOT* (emotion) – used with obvious emotion words (such as "afraid" and "glad") and *Pos/Pstv* (positive) and *Neg/Ngtv* (negative) – used to indicate the valence of emotion-related words. Interjections are also often used in the context of emotion expression. For example, "what" is used to express *surprise*; it is included in GI under the tag *Intrj* (See Figure 4.1). Finally, two more emotion-related tags, namely *Pleasure* and *Pain*, were included in the feature group drawn from GI.

AFRAID#1 H4Lvd Subm Psv Pain EMOT WLBPSYC WLBTOT Modif PFREQ | adjective: Feeling fear, filled with apprehension
AFRAID#2 H4Lvd Neg Ngtv Weak Negate NOT LY | 1% idiom: "Afraid not"

FED#1 H4Lvd Natpro DAV WLBGAIN WLBTOT SUPV ED | 72% verb-adj: Provided something necessary for growth
FED#2 H4Lvd Neg Ngtv Hostile Psv Pain EMOT NEGAFF Modif | 28% idiom-adj-adv: "Fed up" - disgusted

GLAD#1 H4Lvd Pos Pstv Psv Pleasure EMOT WLBPSYC WLBTOT Modif PFREQ | 96% adjective: Pleased
GLAD#2 H4Lvd Pos Pstv Affil Psv Pleasure WLBPSYC WLBTOT LY | 4% adv: "Gladly" with pleasure

WHAT#1 H4Lvd PRON INDEF INT PFREQ Rltvi | 91% pron: Interrogative and relative (interrogative about 20)
WHAT#2 H4Lvd DET PRE PRE2 INT | 6% adj: Interrogative and relative--"what time is it," "tell me what suit to
WHAT#3 H4Lvd Ovrst Intrj INTJ | 2% adv: Intensifier--"what a pity," "what lovely flowers"

Figure 4.1 Sample GI entries for some emotion-related words

Features from WordNet-Affect

The second resource used to find emotion words is WordNet-Affect (Strapparava and Valitutti, 2004), which assigns a variety of affect labels to each synset in WordNet (Fellbaum, 1998). I utilized the publicly available lists[18] extracted from WordNet-Affect (WNA), consisting of emotion-related words. There are six lists corresponding to the six basic emotion categories identified by Ekman (1992). A detailed description of these lists appears in Table 4.1.

[18] http://www.cse.unt.edu/~rada/affectivetext/data/WordNetAffectEmotionLists.tar.gz

Table 4.1Description of emotion word lists extracted from WordNet-Affect

Category	#synsets	Sample Words
Happiness	227	joy, love, rejoicing, glee, happiness, euphoria, enthusiasm, admiration, cheerful, content, happy, merrily
Sadness	123	sorrow, misery, woe, gloom, grief, depression, heartbreak, mournful, hapless, guilty, sadly, remorsefully, repent, bored
Anger	127	wrath, fury, rage, angry, annoyed, pissed, mad, sore, livid, displeasingly, aggressive, hateful, hostile, malice, spite, resentfully
Disgust	19	repulsion, nauseous, foul, abhorrent, fed_up, abominably, revolt, sicken
Surprise	28	wonder, fantastic, marvelous, baffle, bewilder, astonishing, awestruck, stupefy, dumbfound, staggering
Fear	82	scary, fright, panic, terror, chilling, frightful, terrible, intimidate, dread, anxiously, apprehension

Other Features

Beyond emotion-related lexical features, emotion information in text can also be expressed through the use of symbols such as punctuation and emoticons. Many kinds of texts are characterized by increased use of punctuations (Say and Akman, 1996), which justifies their use as features in emotion classification. In the case of blogs, emotion is quite frequently emphasized through repeated usage of punctuations (as in "it was awesome!!!!"). Emoticons have evolved as textual representations of human facial expressions in online conversations, and have become quite popular in emails, blogs and chats. The use of emoticons for expressing sentiment in online communication has previously been studied in the context of sentiment classification works (Mishne, 2005; Read, 2005). Therefore, emoticons and special punctuation, exclamation marks ("!") and question marks ("?") were also included as features.

Table 4.2 summarizes the features (14 in all) that were used in the ML-based classification experiments. The feature vector for a sentence represented the counts for all the features in the sentence.

Table 4.2 Summary of features used in emotion/non-emotion classification

GI Features	WN-Affect Features	Other Features
Emotion words	Happiness words	Emoticons
Positive words	Sadness words	Punctuation Symbols ("!" and "?")
Negative words	Anger words	
Interjection words	Disgust words	
Pleasure words	Surprise words	
Pain words	Fear words	

4.2 Experiments and Results

In this section, I describe the experiments performed for classifying sentences in the blog data into two high-level categories, namely, *emotion* and *non-emotion*. I begin with an overview of the data and the classification models used in the experiments, and follow with a description of the results.

The dataset used in this classification experiment comprises all those sentences from the blog corpus (described earlier in Chapter 3) for which there was consensus among the judges on their emotion category. In this manner, a benchmark was created, which could be used to evaluate the results of automatic classification. For the purpose of the binary categorization addressed here, I assigned all emotion category sentences (labeled *hp*, *sd*, *ag*, *dg*, *sp*, *fr*, or *me*) to the class "EM" (emotion), while all *no emotion* sentences were assigned to the class "NE" (non-emotion). The distribution of this dataset is shown in Table 4.3.

Table 4.3 Class distribution in the dataset used in emotion/non-emotion classification

Class	Number of Sentences	Percentage
EM	1466	34.4%
NE	2800	65.6%
Total	4266	100%

For the classification experiments, I used two machine-learning methods, namely, Naïve Bayes (NB), and Support Vector Machines (SVM), which have been popularly used in sentiment classification tasks (Pang et al., 2002; Mullen and Collier, 2004; Kennedy and Inkpen, 2006; Mihalcea and Liu, 2006).

The naïve baseline for the experiments was set at 65.6%, which represents the accuracy achieved by assigning the label of the most frequent class (which in this case is "NE") to all the instances in the dataset. Each sentence was represented by a 14-value vector, representing the number of occurrences of each feature type in the sentence. (The features used in the experiments are listed in Table 4.2.)

Four different sets of experiments were performed to test the effectiveness and contribution of the different feature groups:

1. Using only features from the General Inquirer (GI)
2. Using only features from WordNet-Affect (WNA)
3. Combining features from the GI and WNA
4. Combining all features (including the "other" features comprising of punctuations and emoticons)

Table 4.4 presents the results of ten-fold cross validation for emotion/non-emotion classification experiments performed using the NaïveBayes and SMO (SVM implementation) classifiers in the WEKA (Witten and Frank, 2005) machine learning package.

Table 4.4 Results of emotion/non-emotion classification

Features	Naïve Bayes Accuracy	SVM Accuracy
GI	71.45%	71.33%
WNA	70.16%	70.58%
GI+WNA	71.70%	73.89%
ALL	**72.08%**	**73.89%**

Overall the performance of the SVM classifier was found to be better than that of the Naïve Bayes classifier for this task. The highest accuracy achieved was 73.89%, which surpasses the baseline accuracy of 65.6%. The improvement is statistically significant (on the basis of a t-test, p=0.05).

The best results were achieved when all the feature groups were combined together. While the use of non-lexical features does not seem to affect the output of the SVM classifier, it did increase the accuracy of the Naïve Bayes classifier. These results indicate that a combination of features is needed to improve emotion classification results.

4.3 Discussion

The results of the automatic emotion classification experiments indicate how external knowledge resources can be leveraged in identifying emotion-related words in text. It may, however, be noted that the lexical coverage of these resources may be limited, given the informal nature of online discourse. For instance, one of the most frequent words used for expressing *happiness* in the corpus is the acronym "lol" (for "laughing out loud"), which does not appear in any of these resources. This indicates the need to augment the word lists obtained from the General Inquirer and WordNet-Affect with such words.

In the same context, it is important to note that in these experiments, I have not addressed the case of typographical errors and other orthographic features that express or emphasize emotion in text. They affect the performance results, as they are not counted in the relevant

feature attribute where they should be. Sokolova et al. (2005b) have discussed the various kinds of noise present in online text based communication. They have used spell checking based on frequency counts to use the most appropriate replacement for misspelled words. In our case, however, the task is complicated by the absence of means of distinguishing between inadvertent and intentional misspellings. For instance, "soo sweeet" could be an example of emotion emphasis, or a typographical error, or a combination of both.

Finally, I would like to bring attention to the fact that the use of emotion-related words is not the sole means of expressing emotion. Often a sentence, which otherwise may not have an emotional word, may become emotion-bearing depending on the context or the underlying semantic meaning. Consider the sentence shown in Figure 4.2, for instance, which implicitly expresses *fear* without the use of any emotion bearing word.

What if nothing goes as planned?

Figure 4.2 Sample emotive sentence with no emotion-bearing word

4.4 Conclusion

In this chapter, I described the experiments performed for high-level classification of sentences into *emotion* and *non-emotion* categories. I presented the results of automatic emotion classification experiments using the Naïve Bayes and SVM classifiers. It was found that the use of knowledge resources in identifying the emotion bearing words in the sentences, and consequently in distinguishing the emotional sentences from the non-emotional ones, gives encouraging results. The classification accuracy was 73.89%, significantly higher than the baseline accuracy. I also discussed the factors that affect classification performance. Particularly notable is the conclusion that to be able to accurately classify emotion, we need to consider wider means of identifying the underlying semantics of text.

Chapter 5
Fine-grained Emotion Classification

We can only see a short distance ahead,
but we can see plenty there that needs to be done.
***Alan Turing** (1912-1954)*

Having addressed the problem of high-level classification of sentences into *emotion* and *non-emotion* categories in the previous chapter, I now shift the focus to the identification of fine-grained emotion categories in text, which is also one of the main objectives of this research. In this chapter I describe the efforts that went into finding an effective methodology for the fine-grained automatic classification of sentences on the basis of the emotion type expressed in them.

Section 5.1 describes the dataset used in the experiments reported in this chapter. Section 5.2 introduces a rule-based semi-supervised approach to emotion classification that serves as a baseline for comparing the performance of the experiments performed later. In Sections 5.3, 5.4 and 5.5, I describe in detail the different feature groups used in the experiments. The first type of features I experimented with is a corpus-based unigram representation of text. The second group of features is constructed from words that appear in emotion lexicons. One such lexicon consists of words that I automatically extracted from *Roget's* Thesaurus (Jarmasz and Szpakowicz, 2001) based on their semantic similarity to a basic set of terms that represent each emotion category. Another emotion lexicon builds on lists of words for each emotion category, extracted from WordNet-Affect (Strapparava and Valitutti, 2004).

The ML experiments and results are presented in Section 5.6. Finally, discussion appears in Section 5.7, and conclusions in Section 5.8.

5.1 The Data

For fine-grained emotion classification experiments reported in this chapter, I used only those sentences from the corpus described in chapter 3 for which there was agreement between both the judges. This was done to form a benchmark of emotion-labeled sentences for training and evaluation of classifiers. The *mixed-emotion* category sentences were not chosen as the objective here is to find distinct emotion categories, so overlap of emotions in sentences is beyond the current scope. Further, I chose lower number of *no-emotion* sentences than available in the corpus so as to avoid any skew in the data in favor of the *no-emotion* category. The case of unbalanced data is a problem domain in its own right, and out of the scope of the task considered here. The resulting dataset is rich in all emotion types; the distribution of emotion categories is shown in Table 5.1.

Table 5.1 Distribution of emotion classes in the dataset

Emotion Class	Number of instances
Happiness (hp)	536
Sadness (sd)	173
Anger (ag)	179
Disgust (dg)	172
Surprise (sp)	115
Fear (fr)	115
No-emotion (ne)	600

5.2 Developing a Baseline

An object of interest here is to investigate if emotion in text can be discerned on the basis of its lexical content. A naïve approach to determining the emotional orientation of text is to look for obvious emotion words, such as "happy", "afraid", and "astonished" in text. The presence of one or more words of a particular emotion category in a sentence provides a good premise for interpreting the overall emotion of the sentence. This kind of approach relies on a list of words with prior knowledge about their emotion type, and uses it for sentence-level classification. The obvious advantage of this approach is that it requires no training data.

For evaluation purposes, I utilized this rule-based approach to develop a baseline system that counts the number of emotion words of each category in a sentence, and then assigns the category with the largest number of words to the sentence. Ties were resolved by choosing the emotion label based on an arbitrary predefined ordering of emotion classes. A sentence containing no emotion word of any type was assigned the *no-emotion* category. For obtaining prior knowledge about emotion-bearing words, I utilized the word lists[19] extracted from WordNet-Affect (Strapparava and Valitutti, 2004) for six basic emotion categories.

Table 5.2 shows the precision, recall, and F-Measure values for the baseline system. As there are seven classes in these experiments, the class imbalance makes accuracy values less informative compared to precision, recall, and F-measure values (Nastase et al., 2006). Hence, I do not report accuracy values in the results.

The baseline system shows precision values above 50% for all but two classes, which reaffirms the premise that looking for emotion words is an effective means for determining the emotion category of a sentence. However, this method fails in the absence of obvious emotion words in the sentence, as indicated by low recall values for all "emotion" classes. Recall is higher for *no-emotion* class, which indicates that most sentences without any emotion words were correctly recognized by the baseline system. This is because the presence of any emotion word in the sentence would lead the baseline system to count that sentence in the relevant emotion category. On the other hand, low precision for *no-emotion*

[19] The same word lists were utilized earlier in Chapter 4.

class can be explained by the fact that in many instances the sentence was an emotion sentence without any obvious emotion-bearing words, and hence was incorrectly labeled as a *no-emotion* sentence by the system. It can therefore be concluded that emotion may be expressed in a sentence without the employment of explicit emotion-bearing words. Thus, in order to improve recall, there is a need to increase the ambit of words that are considered emotion-related, including as well those which may not be directly but conceptually related to emotions.

Table 5.2 Performance metrics of the baseline system

Class	Precision	Recall	F-Measure
Happiness	0.589	0.390	0.469
Sadness	0.527	0.283	0.368
Anger	0.681	0.262	0.379
Disgust	0.944	0.099	0.179
Surprise	0.318	0.296	0.306
Fear	0.824	0.365	0.506
No-emotion	0.434	0.867	0.579

Another pitfall of this baseline system is that it is unable to handle negations. Consider for example, the sentence shown in Figure 5.1. The baseline system will inevitably categorize this sentence as *happy* due to the presence of the emotion word "enjoying". The baseline system can be improved by addition of new rules to handle negation and other shortcomings. However, handcrafted rules cannot be generalized for all domains and applications. An alternative approach employed in research is to use the machine learning approach, which can automatically learn rules from training data for characterizing emotion in text. Machine learning methods can also to some extent capture negations and other valence shifters if they appear frequently in the training data, as noted previously by Kennedy and Inkpen (2006).

I knew she was not enjoying it.

Figure 5.1 Sample emotion sentence illustrating negation

In the following pages, I describe the ML experiments performed for fine-grained emotion classification. For ML-based methods to be effective, first an appropriate set of features need to determined for characterizing the text. To this end, I experimented with two types of features – corpus-based features and emotion lexicon-based features. They are described next.

5.3 Corpus-based features

The corpus-based features exploit the statistical characteristics of the data on the basis of the distribution of n-grams in the data. In the experiments, I utilized unigrams (n=1) as features. Unigram models have widely been used in text classification tasks and previously shown to provide good results in sentiment classification tasks (Pang et al., 2002; Kennedy and Inkpen, 2006). This is because unigram representations can capture a variety of lexical combinations and distributions, including those of emotion words. This is particularly important in the case of blogs, whose language is often characterized by frequent use of new words, acronyms (such as "lol"), onomatopoeic words ("haha", "grrr"), and slang, all of which can be captured in unigram representation. Another advantage of using corpus-based representation is that it leverages entirely the statistical properties of the data and does not require any prior knowledge about the data under investigation or the classes to be identified. It can thus be considered as an automatic approach that does not rely on expert knowledge (Sokolova et al., 2005a).

For the ML experiments, I selected all unigrams that occur more than three times in the corpus. A similar approach has been used in (Pang et al., 2002) and (Kennedy and Inkpen, 2006). This represents a kind of automatic feature selection that removes non-informative terms based on corpus statistics. It is based on the assumption that rare terms are non-

informative for category prediction and do not influence the global performance of the classifier (Yang and Pederson, 1997). In case of data drawn from blogs, this method of selecting unigrams leads to elimination of rare words, as well as foreign-language words and spelling mistakes, which are quite common in blogs. Yang and Pederson (1997) have indicated that when rare terms are noise terms, their removal leads to improvement in classification performance.

I also excluded from the unigrams words that occur in a list of stopwords, comprising primarily of function words that do not generally have emotional connotations. For this purpose, I used the SMART list of stopwords[20], with minor modifications. For instance, I removed from the stop list words such as "what" and "why", which may be used in the context of expressing surprise.

5.4 Features derived from *Roget's* Thesaurus

The second type of features that I experimented with is selected using a knowledge-based approach. More specifically, these features consist of words that are in some way related to one or more of the emotion categories being investigated in this research. I utilized *Roget's* Thesaurus[21] (Jarmasz and Szpakowicz, 2001) to automatically build a lexicon of emotion-related words. The emotion lexicon-based features, in contrast to corpus-based features require prior knowledge about emotion-relatedness of words. In this case, the knowledge is extracted from *Roget's* Thesaurus.

Roget's classification system (Roget, 1852) groups together related concepts into eight levels of hierarchy. The hierarchy extends from eight top-level classes (shown in Figure 5.2) to the semi-colon groups at the lowest level consisting of semantically related words. For detailed account of the Roget's classification structure, see (Jarmasz and Szpakowicz, 2001).

---

[20] Used with the SMART information retrieval system at Cornell University. Available at: ftp://ftp.cs.cornell.edu/pub/smart/english.stop

[21] Several versions of *Roget's* Thesaurus are available. In the experiments reported in this thesis, the 1987 Penguin's Roget's Thesaurus (Kirkpatrick, 1987) was used.

1. Abstract Relations
2. Space
3. Matter
4. Intellect: formation of ideas
5. Intellect: communication of ideas
6. Volition: individual volition
7. Volition: social volition
8. Emotion, religion and morality

Figure 5.2 Top-level Classes in the Roget's Thesaurus

Roget's structure can be utilized for calculation of semantic relatedness between words, based on the path length between the nodes in the structure that represent those words. In case of multiple paths, the shortest path is considered for determining relatedness. Jarmasz and Szpakowicz (2004) have introduce a similarity measure derived from path length, which assigns scores ranging from a maximum of 16 to most semantically related words to a minimum of 0 to least related words. They have shown that this measure of similarity outperforms several other methods on semantic similarity tests.

Many semantic similarity methods (Hirst and St-Onge, 1998; Leacock and Chodorow, 1998; Resnik, 1995) have been proposed utilizing WordNet (Fellbaum, 1998). They utilize the hypernym/hyponym hierarchy of WordNet, which is not uniform as some domains are presented in more detail than others leading to finer distinctions in the hierarchy (Nastase et al., 2006). Therefore, the similarity measures based on distance between nodes in WordNet hierarchy may not best reflect the relatedness of words. In contrast, the structure of Roget's Thesaurus is more uniform and based on the idea of concept classification, which may lead to better measurement of relatedness. Another advantage of Roget's is that, unlike WordNet, it can be used to trace links between different parts of speech (Jarmasz and Szpakowicz, 2004), which is important in this case, as emotion-related words can belong to different parts of speech.

To build a lexicon of emotion-related words utilizing Roget's structure, I had to make two decisions: select a primary set of emotion words starting with which I could extract other

similar words, and decide on an appropriate similarity score to serve as cutoff for determining semantic relatedness between words.

The primary set of words I selected consists of one word for each emotion category, representing the base form of the *name* of each emotion category. Thus I had the following elements in the set: {*happy, sad, anger, disgust, surprise, fear*}.

The second decision regarding the similarity score (in terms of path length) to be chosen as cutoff for selection of words from the *Roget's* Thesaurus was made on the basis of previous experiments to determine semantic similarity of words. Experiments performed on Miller and Charles similarity data (Miller and Charles, 1991) (reported in Jarmasz and Szpakowicz, 2004) have shown that pairs of words with a semantic similarity value of 16 have high similarity, while those with a score of 12 to 14 have intermediate similarity. Therefore, I selected the score of 12 as cutoff, and included in the emotion lexicon all words that had similarity scores of 12 or higher with respect to the words in the primary set. This selection of cutoff therefore serves as a form of feature selection. The total number of words thus selected for inclusion in the lexicon was 1710.

In Table 5.3, I present sample words from the lexicon with similarity scores of 16, 14, and 12 for each emotion category. These words represent three different levels of relatedness to each emotion category. As these words indicate, I was able to identify a large variety of emotion related words belonging to different parts of speech. Among the obvious emotion-bearing words recognized through this method are "smiling", "crying", "boring", "hate", "bitter", "shock", "dislike", "hurt", "amazing", "stunned", "panic", etc. Many of the words in this lexicon go well beyond the stereotypical words associated with different emotions, signifying the wide variety of words that may contextually evoke some type of emotion in humans[22]. These include "house", "bed", "divine", "work", "increase", "black", "power",

[22] It is interesting to find words such as "house", "bed", and "rest" related to *happiness*, while "work" and "leadership" associated with *disgust*.

Table 5.3 Sample words from the emotion lexicon built using Roget's Thesaurus

Emotion Class	Similarity Score=16	Similarity Score=14	Similarity Score=12
Happiness (hp)	family, home, friends, life, house, loving, partying, bed, pleasure, rest, close, event, lucks, times, played, cards	love, like, feel, pretty, lovely, better, smiling, nice, warm, beautiful, hope, cutest celebrations, increase, desires	gift, treats, adorable, fun, hug, kidding, bigger, great, hero, lighting, won, stars, enjoy, favourite, social, divine, found
Sadness (sd)	crying, lost, wounds, bad, pills, falling, messed, spot, unhappy, pass, black, events, hurts, shocked	ill, bored, feeling, ruin, blow, down, wrong, awful, evil, worry, crushing, bug, death, trouble, dark	defeat, nasty, boring, ugly, loser, end, victim, sick, hard, serious, aggravating, bothering, burning, buried
Anger (ag)	pride, fits, stormed, abandoned, bothered, mental, anger, feelings, distractions, tripped, states	hate, burn, upset, dislike, wrong, blood, ill, flaws, bar, defects, bitter, black, growled, slow, passion	lose, throw, offended, hit, power, feel, flaring, pills, broken, life, forgot, ranting
Disgust (dg)	shock, disgust, dislike, loathing	hate, pain, horrifying, ill, pills, sad, wear, blood, appalling, end, work, weighed, regrets, bad, leadership	feel, fun, lies, drawn, lose, missed, deprived, lack, sighs, defeat, down, hurt, tears, insulted, criticized
Surprise (sp)	plans, catch, expected, early, slid, slipped, earlier, caught, act	left, swing, noticed, worry, times, amazing, stolen, break, interesting, attention	realize, pick, wake, sense, jumped, new, late, magic, omen, forget, popped, feel, question, throw
Fear (fr)	nervous, cry, terror, panic, feelings, run, fog, fire, turn, police, faith, battle, war, sounds	falling, life, stunned, pay, broken, hate, blast, times, hanging, hope, broken, blood, blue	fearful, spy, night, upset, feel, chased, hazardous, tomorrow, victim, grim, terrorists, apprehensive, dreams, freak

"new", "dreams", etc. Particularly notable are some generic neutral words, such as "life", and "times" associated with many emotion categories, indicating their relevance to emotions. When considered individually, most of these words evoke a particular type of emotion. The surrounding context, however, may influence a word's emotional orientation. For example, in "broken house" or "ruined house", the context provided by the words "broken" and "ruined" changes the original emotion of *happiness* associated with "house".

In some cases, this method could capture the different senses of some words. For instance, one sense of the word "fun" (as in "we had lot of fun") indicates its association with *happiness*, while another sense (as in "make fun of") is related to *disgust*. Both these senses are correctly categorized in the lexicon. However, I found that the word "blast" is only included in the *fear* category. Another sense of the word (as in "we had a blast") indicating *happiness* does not find a place in the lexicon. There were many other words whose assignment to some categories cannot be intuitively explained. However, no effort was made to manually clean the lexicon. The classification system of *Roget's* was considered as the sole criterion for inclusion of words in this lexicon. This is because some words may not be obviously emotional, but may be potentially emotional depending on the context. Ortony *et al.* (1988) have previously explored this distinction between words directly referring to emotions, and those referring indirectly to emotions or evoking emotions (such as the word "defeat", which may evoke *sadness* or *anger*).

5.5 Features derived from WordNet-Affect

Besides the automatically acquired lexicon describe din the previous section, I also derived features from WordNet-Affect (Strapparava and Valitutti, 2004), which is an affective lexical resource that assigns a variety of affect-related labels to a subset of WordNet synsets that convey affective concepts. I utilized the publicly available lists of words extracted from it for each of the six emotion categories.

5.6 Experiments and Results

I experimented with the different feature groups described earlier in sections 5.3, 5.4 and 5.5 using the Support Vector Machine (SVM) machine learning method for predicting the emotion category of the sentences in my dataset. I trained classifiers with features for each emotion class. In all experiments, a sentence was represented by a vector containing values indicating the number of times each feature occurred in the sentence. In Table 5.4, I report results from ten-fold cross-validation experiments conducted using the SMO implementation of SVM in WEKA (Witten and Frank, 2005). (The baseline F-measure values for all classes are shown in the last column for comparison.)

- In the first set of experiments, I used only corpus-based unigram features. This setup gives high precision values for all emotion classes (as shown in Table 5.4), and the recall and F-measure values surpass baseline values for all classes except the *no-emotion* class. This validates the earlier premise that unigram representation can capture lexical distributions well for accurate prediction of emotion categories.

- In the second set of experiment, I used as features all words in the emotion lexicon acquired from Roget's Thesaurus (RT). (These features were described previously in Section 5.4) The F-measure values beat the baseline for four out of seven classes. These four classes are: *happiness*, *disgust*, *surprise*, and *fear*. When compared with corpus-based features, I found that the latter perform better than the RT features.

- However, in the third set of experiment, when I combined both corpus-based unigram features and the RT features, I found that this resulted in an increase in recall values across all seven classes. The recall values are better than those obtained by using any one group of features alone. This improvement in recall is statistically significant, as shown by paired t-test, $p<0.05$, when performance is compared with corpus-based unigram features, and $p<0.001$ when compared with RT features.

Table 5.4 Results of fine-grained classification using SVM

Model	Class	Precision	Recall	F-Measure	Baseline F-Measure
Corpus-based Unigrams	Happiness	**0.840**	0.675	0.740	0.469
	Sadness	**0.619**	0.301	0.405	0.368
	Anger	0.634	0.358	0.457	0.379
	Disgust	**0.772**	0.453	**0.571**	0.179
	Surprise	**0.813**	0.339	0.479	0.306
	Fear	**0.889**	0.487	0.629	0.506
	No-emotion	0.581	0.342	0.431	0.579
Roget's Thesaurus (RT) Features	Happiness	0.772	0.562	0.650	0.469
	Sadness	0.574	0.225	0.324	0.368
	Anger	0.638	0.246	0.355	0.379
	Disgust	0.729	0.297	0.421	0.179
	Surprise	0.778	0.243	0.371	0.306
	Fear	0.857	0.470	0.607	0.506
	No-emotion	0.498	0.258	0.340	0.579
Corpus-based Unigrams + RT Features	Happiness	0.809	0.705	**0.754**	0.469
	Sadness	0.577	0.370	**0.451**	0.368
	Anger	0.636	0.419	0.505	0.379
	Disgust	0.686	0.471	0.559	0.179
	Surprise	0.717	0.374	0.491	0.306
	Fear	0.831	0.513	0.634	0.506
	No-emotion	0.586	0.512	0.546	0.579
Corpus-based Unigrams + RT Features + WNA Features	Happiness	0.813	**0.698**	0.751	0.469
	Sadness	0.605	**0.416**	0.493	0.368
	Anger	**0.650**	**0.436**	**0.522**	0.379
	Disgust	0.672	**0.488**	0.566	0.179
	Surprise	0.723	**0.409**	**0.522**	0.306
	Fear	0.868	**0.513**	**0.645**	0.506
	No-emotion	**0.587**	**0.625**	**0.605**	0.579

* Highest precision, recall, and F-measure values for each class are shown in bold

- In the last set of experiments, I added features from WordNet-Affect[23] (WNA) to the feature set containing corpus-based unigrams and RT features. This led to further improvement in overall performance. Combining all features, I achieved highest recall values across all classes among all the methods tested so far. The resulting F-measure values (ranging from 0.493 to 0.751) surpass the baseline values across all seven classes. This improvement over baseline performance is statistically significant as shown by paired t-test, p=0.005.

I repeated all the aforementioned experiments with the Naïve Bayes classifier. The results from ten-fold cross-validation experiments conducted using the WEKA (Witten and Frank, 2005) machine-learning package are shown in Table 5.5. The performance using the Naïve Bayes classifier was found to be worse than that of SVM. A similar pattern has been earlier observed by Joachims (1998) and Pang et al. (2002). Text classification domain quite often deals with high dimensionality of feature space. Many learning algorithms do not scale to a high dimensional feature space (Yang and Pederson, 1997). SVMs have been shown to give good performance in text classification experiments as they scale well to the large amounts of features often encountered in text classification (Joachims, 1998). They have been shown to outperform other types of classifiers in this domain (Yang and Liu, 1999; Pang et al, 2002).

5.7 Discussion

I observe that corpus-based features and emotion-related features together contribute to improved performance, better than given by any type of feature group alone.

[23] I also performed experiments using the WNA features alone, however, the results were worse than all other results reported here.

Table 5.5 Results of fine-grained classification using Naive Bayes

Model	Class	Precision	Recall	F-Measure	Baseline F-Measure
Corpus-based Unigrams	Happiness	0.743	0.377	0.500	0.469
	Sadness	0.476	0.341	0.397	0.368
	Anger	0.344	0.302	0.321	0.379
	Disgust	0.529	0.320	0.399	0.179
	Surprise	0.337	0.243	0.283	0.306
	Fear	0.538	0.374	0.441	0.506
	No-emotion	0.394	0.022	0.041	0.579
Roget's Thesaurus (RT) Features	Happiness	0.687	0.319	0.436	0.469
	Sadness	0.388	0.289	0.331	0.368
	Anger	0.400	0.201	0.268	0.379
	Disgust	0.604	0.169	0.264	0.179
	Surprise	0.388	0.226	0.286	0.306
	Fear	0.672	0.391	0.495	0.506
	No-emotion	0.267	0.013	0.025	0.579
Corpus-based Unigrams + RT Features	Happiness	0.690	0.386	0.495	0.469
	Sadness	0.368	0.434	0.398	0.368
	Anger	0.270	0.346	0.303	0.379
	Disgust	0.387	0.308	0.343	0.179
	Surprise	0.256	0.287	0.270	0.306
	Fear	0.360	0.426	0.390	0.506
	No-emotion	0.471	0.055	0.099	0.579
Corpus-based Unigrams + RT Features + WNA Features	Happiness	0.698	0.384	0.0496	0.469
	Sadness	0.361	0.422	0.389	0.368
	Anger	0.268	0.358	0.306	0.379
	Disgust	0.402	0.308	0.349	0.179
	Surprise	0.283	0.296	0.289	0.306
	Fear	0.366	0.426	0.394	0.506
	No-emotion	0.493	0.062	0.11	0.579

Any automatic approach to recognize emotion should inevitably take into account a wide variety of words that are semantically connected to emotions. While some words are obviously affective, many more are potentially affective. The latter derive their affective property from their associations with emotional concepts. For instance, words like "family", "friends", "home", are not inherently emotional, but – because of their well-known semantic association with emotion concepts – their presence in a sentence can be taken as an indicator of emotion expression in the sentence.

I interpret the results as indicative of how much correlation the classifiers can learn between the features and the predicted class. Considering the best results using all features, I find that this correlation is highest for the *happy* class, indicated by a precision of 0.813 and recall of 0.698, the highest among all classes. It can therefore be concluded that it is easiest to discern happy emotion from text compared to other emotions. It is interesting to note that combining emotion-related words with corpus-based unigrams increases precision and recall for no-emotion class as well. This improvement may be considered as a consequence of improvement in the recognition of emotion categories.

5.8 Conclusion

Working on a corpus of blog sentences annotated with emotion labels, I demonstrated that a combination of corpus-based unigram features and features derived from emotion lexicons can help automatically distinguish basic emotion categories in written text. When used together in an SVM-based learning environment, these features increased recall in all cases and the resulting F-measure values significantly surpassed the baseline scores for all emotion categories.

In addition, I described a method of building an emotion lexicon derived from Roget's Thesaurus on the basis of semantic relatedness of words to a set of basic emotion words for each emotion category. The effectiveness of this emotion lexicon was demonstrated in the emotion classification tasks.

Chapter 6
Emotion Intensity Recognition

All progress is precarious, and the solution of one problem
brings us face to face with another problem.
Martin Luther King Jr. *(1929-1968)*

This chapter focuses on the problem of automatic recognition of the intensity of emotion expressed in a sentence. I describe the various syntactic constructs used in the English language to express and modify the intensity of emotion. I use those as features in the machine learning experiments to classify sentences on the basis of emotion intensity. Section 6.1 describes the dataset used in the experiments reported in this chapter. Section 6.2 presents in detail the different syntactic features used in the experiments. The ML experiments and results are presented in Section 6.3. The discussion appears in Section 6.4, and conclusion in Section 6.5.

6.1 The Data

For the emotion intensity classification experiments reported in this chapter, I used only those sentences from the corpus described in chapter 3 for which there was agreement on emotion intensity between both the judges. This was done to form a benchmark of emotion intensity-labeled sentences for training and evaluation of classifiers. The distribution of emotion intensity levels in the data is shown in Table 6.1. I consider four categories – *high*, *medium*, *low*, and *neutral*. The first three are marked as such in the annotated corpus, while

the *neutral* category consists of sentences drawn from the *no emotion* category. It is important to consider *neutral* category also as even the emotive texts contain a large number of no-emotion sentences.

Table 6.1 Distribution of emotion intensity levels in the data

Emotion Intensity	Number of Sentences
High	641
Medium	447
Low	193
Neutral	600

6.2 Emotion Intensity Expressions

There are various ways in which the intensity of an emotion can be expressed or modified. One of the ways is the use of relatively *strong* and *weak* words such as "hate" and "abhor", which express different gradations of intensity. Often the expression of emotion strength is relative and determined by the context. Another way of modifying emotion strength is by the use of modifiers such as "very", "not", "little", "somewhat", etc. These modifiers act to increase the strength (as in "very happy", "highly grateful", and "much disappointed") or decrease the strength (as in "little embarrassed", "somewhat apprehensive", and "not pathetic") of expressed emotion. Comparative and superlative forms of adjectives (as in "happier times" and "the greatest disaster" also express relatively higher-intensity emotions.

Several syntactic patterns can be recognized to affect the strength of the expressed emotion. In previous work, Whitelaw et. al. (2005b) have identified syntactic patterns of appraisal adjectives and their modifiers (as in "very good" and "not really very funny"). Benamara et al. (2007) have also used adverb-adjective combinations (AACs) using adverbs of degree ("extremely", "hardly", "really", etc.) in scoring schemes for determining the strength of a sentiment.

In my experiments, I used syntactic bigrams related to emotion intensity expression, identified using the Link Grammar Parser (Sleator and Temperley, 1991). The Link Parser recognizes a variety of syntactic links between words in a sentence. Figure 6.1 shows a sample output from the link parser for the sentence "This was the best summer I have ever experienced". I examined the various categories of links output by the parser, and out of them, I chose those types of links that I considered were relevant to emotion intensity expression. The syntactic bigrams consisted of pairs of words in a sentence that happen to be connected by one of those types of links. These bigrams were then used as features in the ML experiments described later in Section 6.2.

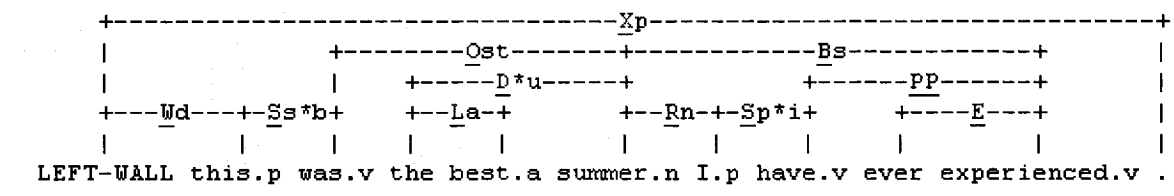

Figure 6.1Sample output from Link Parser

The first category of links I identified is "EA", which connects adverbs to adjectives in a sentence. Table 6.2 shows the various links (and sample sentences) belonging to this category. The different links within this category recognize finer distinctions in adverb-adjective links, for example, the use of word "so" as an adjectival adverb in "Eaxk". The second category of links I considered is "EE", which connects adverbs to other adverbs. This type of syntactic link can also be used for expressing different intensities of emotion. Similar to the "EA" category, there are finer distinctions within this category. Table 6.3 shows the various links (and sample sentences) in this category.

Table 6.2 Adverb-adjective links and examples

Link	Example
EA	She is very happy.
Eah	How proud must she be!
Eax	They have so many losses to cope with.
Eaxk	It was so disappointing to lose in this way.
Eam	She is more happy than anyone is.

Table 6.3 Adverb-adverb links and examples

Link Type	Example
EE	I will quite happily do it for you.
EEh	How fearlessly did he fight!
EEx	They were pretty much afraid of him.
EEXk	She replied so angrily.
EEm	They lived more happily thereafter.

Table 6.4 Other adverb related links and examples

Link Type	Example
E	She definitely is shaken by the news.
ECn	How much more hatred do they have?
ECa	They are a much happier couple today.
EF	Our losses are a good enough reminder for us.
EB	It was undoubtedly the best.
Em	The applause was somewhat muted.

Table 6.5 Other adjective related links and examples

Link Type	Example
A	They have an awful lot going on in their lives.
AL	We had all the fun you could think of.
AM	We had as much fun as you could think of.

Table 6.6 Some other emotion-related links and examples

Link Type	Example
IDBB	I am very very disappointed.
IDC	It was quite a show!
DD	It was the best!
Dmc	They suffered many losses.
Dmu	The news brought them little happiness.
MVa	He spoke angrily.
MVb	She sings better.
MVm	I enjoyed the play more than he did.
La	It is the happiest moment of my life.
N	It might not be bad.
Pam	I would have been happier if you had come.

I also considered several other adverb-related and adjective-related links output by the Link Parser. They are shown in Table 6.4 and Table 6.5. The link "E" is used for verb-modifying adverbs that precede the verb. The links "ECn" and "ECa" connect adverbs and comparative adjectives; the former is noun-focused while the latter is adjective-focused. The link "EF" connects the word "enough" to adjectives and adverbs. The link "A" connects pre-noun adjectives to nouns; "AL" connects determiners such as "all" and "both" to the following determiners; while "AM" is used in comparative constructions such as "as much" and "as many".

In addition to the links related to adjectives and adverbs, I also used some other links that may sometimes be used to modify the strength of an emotion. They are listed in Table 6.6.

The links "IDBB" and "IDC" are special types of links that deal with idiomatic expressions (which are preloaded in the parser's dictionary). The links "DD", "Dmc" and "Dmu" are used to connect determiners to nouns and adjectives acting as nouns. The links "MVa", "MVb" and "MVm" connect verbs (and adjectives) to various kinds of modifying phrases as shown in the example sentences in Table 6.6. The link "La" connects determiners to superlative adjectives; "N" connects the word "not" to preceding auxiliaries and modals, and "Pam" connects forms of the verb "be" with comparative forms of adjectives.

6.3 Experiments and Results

I used the Support Vector Machine (SVM) machine learning method for predicting the emotion intensity of the sentences in the dataset. I trained classifiers with features for each emotion intensity class. In all experiments, a sentence was represented by a vector containing values indicating the number of times each feature occurred in the sentence. In Table 6.7, I report the results of ten-fold cross-validation experiments conducted using the SMO implementation of SVM in WEKA (Witten and Frank, 2005).

As a baseline, I used the unigram representation for each sentence. The unigram representation exploits the statistical characteristics of the data on the basis of the distribution of words in the text. Those models have the advantage of being less dependent on the data under investigation, though they can capture a variety of lexical combinations that characterize each class. I made the same decisions about the selection of unigrams and the stopword list as in the emotion category classification experiments (see Section 5.3).

The baseline system shows precision values of over 0.6 for two classes – *high* and *neutral*. This is predictable because high-intensity expressions and no-emotion expressions are more pronounced and more recognizable by both automatic systems and humans. The precision values are low for *medium* and *low* intensity classes. The low recall values of the baseline system indicate that recognizing emotion intensity is a difficult task, as humans can express emotion linguistically in a variety of ways, many of which are difficult to capture in computational models.

In the next step, I used the syntactic bigrams (described in the previous section) as features along with the corpus-based unigram features. This setup led to an increase in the recall values for all intensity classes, thereby validating the premise that syntactic bigrams can capture many of the intensity-related expressions for accurate prediction of emotion intensities. The resulting F-measure values surpass the baseline values for all intensity classes. This increase, though modest, was statistically significant (based on a t-test, $p<0.05$)

Table 6.7 Results of emotion intensity classification using SVMesults of emotion intensity classification using SVM

Model	Intensity Class	Precision	Recall	F-Measure
Corpus-based Unigrams	High	**0.613**	0.376	0.466
	Medium	0.359	0.186	0.245
	Low	**0.230**	0.073	0.110
	Neutral	**0.624**	0.362	0.458
Corpus-based Unigrams + Syntactic bigrams	High	0.568	**0.435**	**0.493**
	Medium	**0.367**	**0.255**	**0.301**
	Low	0.225	**0.130**	**0.164**
	Neutral	0.591	**0.443**	**0.507**

6.4 Discussion

The experiments indicate that corpus-based unigram and syntactic bigram features together provide a better combination for predicting the emotion intensity of sentences than the unigram representation alone. This indicates the usefulness of syntactic bigrams. The results can be interpreted as indicative of how much association the classifiers can learn between the features and the predicted class. I found that the *high* and *neutral* classes are better characterized by the corpus-based features. It can therefore be concluded that it is easier to discern these intensities than the finer intensity gradations of *medium* and *low*.

6.5 Conclusion

In this chapter, I described the attempts to identify the emotion intensity levels using a combination of corpus-based unigram and syntactic bigram features. When used together in an SVM-based learning environment, these features increased recall in all cases and the resulting F-measure values significantly surpassed the baseline scores for all intensity categories. I also demonstrated the usefulness of syntactic bigrams in representing the constructs used to express emotion and modify emotion intensity. The advantage of using corpus-based syntactic bigrams is that this approach does not require any external knowledge resource, and exploits only the syntactic combinations used to express emotion.

Chapter 7
Conclusion

Not everything that can be counted counts,
and not everything that counts can be counted.
***Albert Einstein** (1879-1955)*

This thesis reports an investigation of expressions of emotion in text. The report covers manual works performed to identify textual expressions of emotion, as well as the computational methods adopted to learn emotions in text.

The presentation begins with an introduction of the problem in Chapter 1. The high-level objective of this work was to explore the applicability of automatic approaches to recognizing emotions expressed in written text. The scope of this work was limited to learning emotions that can be interpreted from textual expressions. Emotion research is an inter-disciplinary area and draws upon earlier works in Psychology, Linguistics, and Natural Language Processing. I provided a description of the previous works in these fields in Chapter 2.

One of the tasks addressed in this work has been to prepare a corpus annotated with emotion-related information. In Chapter 3, I described the process of how an emotion-annotated corpus was prepared. Four human judges were involved in the process; and an inter-annotator study was performed to investigate the reliability of the annotations produced. The differences in human judgment related to the assignment of *emotion category* and *emotion intensity* to individual sentences in the corpus indicate that interpretation of emotion expressed in text is subjective; in many cases, different judges interpreted differently the

emotive content of the same sentences. These differences also point to the inherent difficulty of the task of identifying the type of emotion expressed in text. Only those sentences for which there was an agreement among the judges were included to form a gold standard data for use in later experiments.

The primary objective of this work is to develop automatic methodologies for learning emotions from text. In Chapters 4 and 5, I presented several experiments performed to predict the emotive orientation of sentences. In my work, I adopted knowledge-based methods that make use of external lexical-semantic resources to determine the emotion-relatedness of individual words; this knowledge is further leveraged in machine-learning algorithms to predict emotion at sentence-level. The resources I used – General Inquirer, WordNet-Affect, and *Roget's* Thesaurus – are all publicly available[24]. The former two implicitly categorize words on the basis of semantic distinctions, many of which are relevant to emotion. The latter's classification system was utilized to automatically build a lexicon of emotion-related words based on the similarity of words to a small primary set of manually identified emotion words. The results of the experiments suggest that combining a variety of features – representing explicitly emotion words as well as words that may be conceptually related to emotion – helps distinguish the fine-grained emotions in text. The performance achieved in these experiments was significantly better than that using the baseline methods.

I also addressed the task of automatic recognition of emotion intensity level of text. In Chapter 6, I presented experiments that exploit the lexical and syntactic characteristics of the text for determining its emotion intensity. The results suggest that the use of syntactic bigrams helps improve performance of the emotion intensity recognition system.

I believe that the performance of the system could have been further improved if there were more annotated data available for training the system. Many studies have demonstrated that the accuracy of the learning methods can be improved by increasing the training data (Mihalcea and Strapparava, 2005). Emotion expression is a subjective task, and people use a variety of creative combinations of linguistic constructs to express emotion in text. For any

[24] The publicly available machine-readable Roget's Thesaurus uses the 1911 version of the thesaurus (available at http://www.nzdl.org/ELKB/); this thesis used the 1987 version.

automatic system to be able to recognize these constructs and learn associated emotions, it must be trained with maximum possible varieties of textual emotion expressions. Even after annotating a corpus of 5205-sentences, the number of instances of each emotion category and intensity was small. A large number of sentences in the corpus belonged to the *no-emotion* category. A similar phenomenon was observed by Wiebe et al. (2004) during manual annotation of opinion and non-opinion pieces from newspapers. A parallel may be drawn between opinion pieces and the blog data used in this work. The authors note that opinion pieces are not entirely composed of subjective sentences and that as many as 30% sentences in opinion pieces were found to be objective.

In the above passages, I summarized the work presented in this thesis. In next two sections, I present major contributions of this work and a roadmap for future research.

7.1 Contributions

This work addresses an important and less investigated area of sentiment research, that is, emotion detection in text. The major contribution of this work is to show that it is viable to apply computational methods to identify and distinguish various types of emotions in text. Similar works in fine-grained emotion detection lack mention of conventional performance metrics such as precision and recall, which prevents proper assessment of their approaches (Liu et al., 2003; Neviarouskaya et al., 2007a). They demonstrated the efficacy of their methods through user evaluation studies only. This thesis reports the first empirical results in this domain that addresses the problem of detecting basic emotion categories at the sentence level.

The experiments performed in automatic emotion detection also suggested that to achieve good performance that it is important to include in the ambit of consideration a wide variety of words that go beyond the stereotypical emotion words. This finding might appear commonplace, but the fact is that it is not emphasized in earlier works, and not particularly proven empirically.

Another significant contribution of this work is to produce a 5205-sentence corpus of emotion-annotated data. The annotations include fine-grained distinctions of various emotion

categories, the emotion intensity levels, as well as the emotion indicators in text. No comparable work focusing exclusively on emotions exists in the public domain.

This thesis also introduced a novel approach of automatically building emotion lexicon utilizing the classification system of *Roget's* Thesaurus. In earlier works, emotion (or sentiment) words were typically acquired using WordNet or corpus-based approaches. In the approach introduced here, a variety of emotion-related words were learned, and their usability demonstrated by their effectiveness in ML methods for emotion detection.

7.2 Future Work

The work presented in this thesis can be pursued further in several directions. One of the tasks to be addressed is to explore the relation between emotion categories and intensity. Some steps could be taken to address the special needs of the kind of informal language used in online communication. This would help improve performance.

Another direction for future work is to consider the emotion intensity classification problem as that of ordinal classification – that is, classification with ordered categories. The intensity levels of high, medium, low, and neutral form a natural ordering, which can be taken into account during classification. One of the approaches used for this kind of problem is to assign numerical labels to classes and then apply regression (Wilson et al., 2004). Another approach is to transform the ordinal problem into a series of binary class problems that incorporate ordering information (Frank and Hall, 2001). These methods could be applied to the emotion intensity classification task addressed in this thesis.

The data prepared as part of this work is rich in emotion annotations and offers several exciting possibilities for further research. Future work may attempt to automatically identify the emotion indicators in sentences. A corpus-driven approach can allow a lexicon of emotion words to be built starting from the set of emotion indicators identified during the annotation process. This set could be further extended based on existing syntactic patterns in the set and similarity measures to identify similar words.

Content Analysis of emotion-labeled data is yet another possible line of research. If the salient words in emotion sentences are analyzed on the basis of their affiliations to semantic categories as defined in resources such as *Roget's* Thesaurus, it can help pinpoint the sources that evoke particular emotions.

Appendix A

Lists of emotion-related seed words used to build blog corpus

Emotion Category	Seed Words
Happiness	awesome, happy, amused, fantastic, excited, pleased, cheerful, love, great, amazing
Sadness	sad, lonely, gloomy, depressed, unhappy, down, disheartened, sorrowful, painful, guilty
Anger	angry, annoyed, boiling, enraged, indignant, irate, furious, inflamed, livid, mad
Disgust	stupid, sucks, irritated, humiliated, disgusted, nauseating, sickening, contempt, repelling, unpleasant
Surprise	astonished, bewildered, surprised, confused, sudden, unaware, shocked, perplexed, what, unexpected
Fear	afraid, frightened, fearful, horrified, nervous, panicked, alarmed, phobia, scared, insecure

Appendix B

Annotation Instructions

Emotion Annotation Experiment -- Instructions

1. Introduction

The goal of the emotion annotation experiment is to manually add emotion information to each sentence in a dataset of blogs collected from the web. This manually annotated data can be used to train computer-based systems to automatically identify emotion information on a large-scale.

2. Annotation Scheme

This section describes the types of annotations that are to be added to each sentence in the corpus.

2.1 Emotion Category

Researchers have identified various categories of basic and secondary emotions. For the purpose of this experiment, we will use the six basic emotions identified by Ekman (Ekman, P. 1992. An Argument for Basic Emotions. Cognition and Emotion. 6, 169-200).

Annotators are required to read each sentence and identify which of the following emotion categories can be assigned to the sentence:

- happiness (hp)
- sadness (sd)
- anger (ag)
- disgust(dg)
- fear (fr)
- surprise (sp)
- mixed emotion (me)

- no emotion (ne)

If a sentence contains more than one emotion , or if its emotion doesn't fit well into any of the given six classes (hp, sd, ag, dg, fr, sp), then it is to be assigned to the "mixed emotion" (me) class. A sentence with no emotion content is to be assigned to "non-emotion" (ne) category.

2.2 Emotion Intensity

Annotators are also required to label the intensity of emotion expressed in a sentence. The intensity labels are – high (h), medium (m), and low (l).

2.3 Emotion Indicators

Finally, the annotators have to underline those words and/or phrases in a sentence that carry emotion information. (These are those words in the sentence that lead the annotators to believe that the sentence belongs to a particular emotion category).

3. Examples

The following examples serve to illustrate the kind of annotations expected in this experiment. (Annotators only need to apply their language skills, commonsense knowledge, and intuition in carrying out the annotations. No expertise or training is required for this task.)

In some cases, the expression of emotion may be explicit, and it may be easy to identify the emotion category as well as the emotion indicators, as in the following examples:

[1] Gosh, my exams start in one week, and boy I am nervous.

After annotation:

[1] Gosh, my exams start in one week, and boy I am nervous. (fr, h)

(Here "fr" indicates Emotion Category: "Fear", while "h" indicates Emotion Intensity: "High")

(Similarly, underlined words indicate Emotion Indicators: Gosh, nervous)

[2] And then something unexpected from your past will resurface.

After annotation:

[2] And then something unexpected from your past will resurface.(sp, m)

(Emotion Category: Surprise, Emotion Intensity: Medium, Emotion Indicators: unexpected)

Besides individual words, the annotators may also identify phrases as emotion indicators. (Here, phrase means any collection of consecutive words within a sentence). For example, consider the following sentence:

[3] It was very sickening to see people making fun of him.

After annotation:

[3] It was very sickening to see people making fun of him. (dg, h)

(Category: Disgust, Intensity: High, Indicators: "very sickening", "making fun")

[4] My new food is making me feel really healthy.

After annotation:

[4] My new food is making me feel really healthy. (hp, h)

[5] What if it doesn't go as planned?

After annotation:

[5] What if it doesn't go as planned? (fr,h)

In some cases, the sentence may not convey any emotion, as shown in the following example. In such cases, "emotion intensity" and "emotion indicators" are not marked.

[6] Actually, I just want to give the opportunity a chance.

After annotation:

[6] Actually, I just want to give the opportunity a chance. (ne)

Sometimes, the sentence may contain more than one emotion. For example, the following sentence contains both Happiness and Fear.

[7] I am enjoying the competition at my new job but also scared as everyone is ahead of me.

After annotation:

[7] I am enjoying the competition at my new job but also scared as everyone is ahead of me.(me,h)

The following examples show that sometimes emotion may be expressed in subtle ways, and therefore annotators may have to think hard to identify the appropriate emotion category:

[8] First of all, they were not the first to sell movies online, so why raise a fuss about it?

After annotation:

[8] First of all, they were not the first to sell movies online, so why raise a fuss about it? (ag, l)

[9] I won't ever run away when someone needs me.

After annotation:

[9] I won't ever run away when someone needs me. (me,l)

[10] Officially, this is my last day at work.

After annotation:

[10] Officially, this is my last day at work. (sd, l)

[11] One thing I learned from going to the hospital was to be thankful that hers wasn't a more serious case.

After annotation:

[11] One thing I learned from going to the hospital was to be thankful that hers wasn't a more serious case. (me, m)

[12] Four more days before I leave.

After annotation:

[12] Four more days before I leave. (me, m)

(Emotion Category: Mixed – as it could be happiness or sadness)

[13] What is my neighbor doing here?

After annotation:

[13] What is my neighbor doing here? (sp, m)

[14] He could not take any more, and walked out.

After annotation:

[14] He could not take any more, and walked out. (dg, h)

"Emotion Indicators" may not be marked, if no particular word or phrase sufficiently describes the emotion, as in the following example:

[15] We stayed there for one hour only to be told to come next day!

After annotation:

[15] We stayed there for one hour only to be told to come next day! (ag, h)

Bibliography

Alm, C.O., Roth, D. and Sproat, R. (2005). Emotions from text: machine learning for text-based emotion prediction. In *Proceedings of the Joint Conference on Human Language Technology / Empirical Methods in Natural Language Processing (HLT/EMNLP 2005),* Vancouver, Canada, pages 579-586.

Aman, S. and Szpakowicz, S. (2007). Identifying Expressions of Emotion in Text. In Proceedings of the 10th International Conference on Text, Speech, and Dialogue (TSD-2007), Plzeň, Czech Republic, Lecture Notes in Computer Science (LNCS), Springer-Verlag. (to appear).

Andreevskaia, A., Bergler, S. and Urseanu, M. (2007). All Blogs Are Not Made Equal: Exploring Genre Differences in Sentiment Tagging of Blogs, In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM 2007),* pages 293-294, Boulder, Colorado, USA.

Anscombe, E. and Geach, P. (1970). *Descartes Philosophical Writings.* Nelson: The Open University.

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

Benamara, F., Cesarano, C., Picariello, A., Reforgiato, D. and Subrahmanian, V.S. (2007). Sentiment Analysis: Adjectives and Adverbs are Better than Adjectives Alone, In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM 2007),* Boulder, Colorado, USA.

Bradley, M.M. and Lang, P.J. (1999). *Affective norms for English words (ANEW): Stimuli, instruction manual and affective ratings.* Technical Report C-1, Gainesville, FL. The Center for Research in Psychophysiology, University of Florida.

Carletta, J. (1996). Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics.* 22(2), pages 249-254.

Chesley, P., Vincent, B., Xu, L., and Srihari, R. (2006). Using Verbs and Adjectives to Automatically Classify Blog Sentiment. In *Proceedings of the AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs,* Stanford, California, USA.

Chinchor, N. (1998). MUC-7 test scores introduction. In *Proceedings of the Seventh Message Understanding Conference.*

Clore, G.L., Ortony, A., and Foss, M.A. (1987). The psychological foundations of the affective lexicon. *Journal of Personality and Social Psychology*, Vol. 53, pages 751–766.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20 (1): 37–46.

Cowie, R. (2000). Describing the emotional states expressed in speech. In *Proceedings of the ISCA Workshop on Speech and Emotion*, pages 11–18, Northern Ireland.

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

Curran, J.R. (2002). Ensemble methods for automatic thesaurus extraction. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP-2002)*, pages 222–229, Philadelphia, Pennsylvania, USA.

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6), pages 391–407.

Ekman, P. (1992). An Argument for Basic Emotions. *Cognition and Emotion*, 6, 169-200.

Fellbaum, C. (1998). *WordNet: An electronic lexical database*. MIT Press, Cambridge, MA.

Fleiss, J. L. (1981). *Statistical Methods for Rates and Proportions* (2nd edition). John Wiley & Sons, New York.

Frank, E., and Hall, M. (2001). A simple approach to ordinal classification. In *Proceedings of the 12th European Conference on Machine Learning*, pages 145-156. Berlin: Springer-Verlag.

Frijda, N., Ortony, A., Sonnemans, J. & Clore, G. L. (1992). The complexity of intensity. Issues concerning the structure of emotion intensity. In M. S. Clark (Ed.). *Review of Personality and social psychology* (13, 60-89). Newbury Park, CA: Sage.

Glance, N., Hurst, M., and Tomokiyo, T. (2004). Blogpulse: Automated trend discovery for weblogs. In *Proceedings of the WWW-2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*, New York, USA.

Hatzivassiloglou, V. and McKeown, K.R. (1997). Predicting the semantic orientation of adjectives. In P. R. Cohen and W. Wahlster, editors, In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics, and the 8th EACL*, Somerset, New Jersey, pages 174–181, Madrid, Spain.

Hatzivassiloglou, V. and Wiebe, J. (2000). Effects of adjective orientation and gradability on sentence subjectivity. In *Proceedings of the International Conference on Computational Linguistics (COLING-2000)*, Saarbrücken, Germany.

Hiroshima, N., Yamada, S., Furuse, O., and Kataoka, R. (2006). Searching for Sentences Expressing Opinions by Using Declaratively Subjective Clues. In *Proceedings of the ACL Workshop on Sentiment and Subjectivity in Text*, Sydney, Australia.

Hirst, G. and St-Onge, D. (1998). Lexical chains as representation of context for the detection and correction malapropisms. In Fellbaum, C., editor, *WordNet: An Electronic Lexical Database*, pages 305–322. MIT Press, Cambridge, MA.

Hoffman, L. (2005). *Two Thumbs Up*. Forbes.com. November 15, 2005.

Holzman, L. and Pottenger, W. (2003). *Classification of emotions in internet chat: An application of machine learning using speech phonemes*. Technical Report LU-CSE-03-002, Lehigh University.

Hu, M. and Liu, B. (2004). Mining and Summarizing Customer Reviews. In *Proceedings of the Tenth International Conference on Knowledge Discovery & Data Mining*, pages 168-177, Seattle, Washington, USA.

Hurst, M. and Nigam, K. (2004). Retrieving Topical Sentiments from Online Document Collections. In *Document Recognition and Retrieval XI*, San Jose, California, USA.

Izard, C.E. (1977). *Human emotions*. New York: Plenum Press, 1977.

Jaccard, P. (1908). Nouvelles recherches sur la distribution florale. *Bulletin de la Societe Vaudoise de Sciences Naturelles* 44:223-270.

Jakobson, R., (1960). Linguistics and Poetics, in T. Sebeok, ed., *Style in Language*, Cambridge, MA: MIT Press, pages 350-377.

Jarmasz, M. and Szpakowicz, S. (2001). The Design and Implementation of an Electronic Lexical Knowledge Base. In *Proceeding of the 14th Biennial Conference of the Canadian Society for Computational Studies of Intelligence (AI-2001)*, Ottawa, Canada, pages 325-333.

Jarmasz, M. and Szpakowicz, S. (2004). Roget's Thesaurus and Semantic Similarity. In N. Nicolov, K. Bontcheva, G. Angelova, R. Mitkov (eds.) *Recent Advances in Natural Language Processing III: Selected Papers from RANLP 2003*, John Benjamins, Amsterdam/Philadelphia, *Current Issues in Linguistic Theory*, 260, pages 111-120.

Java, A., Kolari, P., Finin, T., Mayfield, J., Joshi, A., Martineau, J. (2006). BlogVox: Separating Blog Wheat from Blog Chaff. In *Proceedings of the Workshop on Analytics for Noisy Unstructured Text Data, 20th International Joint Conference on Artificial Intelligence (IJCAI-2007),* Hyderabad, India.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of the Tenth European Conference on Machine Learning (ECML-98)*, pages 137–142, Chemnitz, Germany.

Johnson-Laird, P.N. and Oatley, K. (1989). The Language of Emotions: An Analysis of a Semantic Field. *Cognition and Emotion*, 3(2): 81-123.

Kamps, J., and Marx, M. (2002). Words with attitude. *In Proceedings of the First International Conference on Global WordNet*, Mysore, India.

Kennedy, A. and Inkpen, D. (2006). Sentiment Classification of Movie Reviews Using Contextual Valence Shifters. *Computational Intelligence* 22(2):110-125.

Kirkpatrick, B., editor (1987). *Roget's Thesaurus of English Words and Phrases*. Penguin, Harmondsworth, Middlesex, England.

Kim, S., Hovy, E. (2006). Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Text. In *Proceedings of the ACL Workshop on Sentiment and Subjectivity in Text*, pages 1-8, Sydney, Australia.

Krippendorff, K. (1980). *Content Analysis*. Newbury Park, CA: Sage Publications.

Leacock, C. and Chodorow, M. (1998). Combining local context and WordNet sense similarity for word sense disambiguation. In Fellbaum, C., editor, *WordNet: An Electronic Lexical Database*, pages 265–284. MIT Press, Cambridge, MA.

Liu, B., Hu, M. and Cheng, J. (2005). Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of WWW2005: the 14th international conference on World Wide Web*, pages 342–351, New York, NY, USA. ACM Press

Liu, H., Lieberman, H., Selker, T. (2003). A Model of Textual Affect Sensing using Real-World Knowledge. In *Proceedings of the International Conference on Intelligent User Interfaces, IUI 2003*, Miami, Florida, USA.

Liu, H., and Maes, P. (2004). What Would They Think? A Computational Model of Attitudes. In *Proceedings of the ACM International Conference on Intelligent User Interfaces, IUI 2004*, pages 38-45, Island of Madeira, Portugal. ACM Press.

Liu, H. (2006). *Computing Point-of-View: Modeling and Simulating Judgments of Taste*, Ph.D. Dissertation, Program in Media Arts and Sciences, School of Architecture and Planning, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA.

Martin, J.R. and White, P.R.R. (2005). *The Language of Evaluation: Appraisal in English.* Palgrave, London. (http://grammatics.com/appraisal/)

Mehrabian, A. (1995). Framework for a comprehensive description and measurement of emotional states. *Genetic, Social, and General Psychology Monographs*, 121, pages 339-361.

Mihalcea, R. and Liu, H. (2006). A corpus-based approach to finding happiness, in the *AAAI Spring Symposium on Computational Approaches to Weblogs*, Stanford, California, USA.

Mihalcea, R. and Strapparava, C. (2005). Making Computers Laugh: Investigations in Automatic Humor Recognition, In *Proceedings of the Joint Conference on Human Language Technology/Empirical Methods in Natural Language Processing (HLT/EMNLP)*, pages 531-538, Vancouver, Canada.

Mihalcea, R. and Strapparava. C. (2006). Learning to laugh (automatically): Computational models for humor recognition. *Computational Intelligence*, 22(2), pages 126–142.

Miller, G. and Charles, W. (1991) Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1-28.

Mishne, G. (2005). Experiments with Mood Classification in Blog Posts. In *Proceedings of the 1st Workshop on Stylistic Analysis Of Text For Information Access (Style 2005), Brazil.*

Mishne, G. (2006). Information Access Challenges in the Blogspace. In *Proceedings of the International Workshop on Intelligent Information Access.*

Mishne, G. and Glance, N. (2006). Predicting Movie Sales from Blogger Sentiment. In *Proceedings of the AAAI-2006 Spring Symposium on Computational Approaches to Analysing Weblogs*, Stanford, California, USA.

Mullen, T and Collier, N. (2004). Sentiment analysis using support vector machines with diverse information sources. In Dekang Lin and Dekai Wu, editors, *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004)*, pages 412–418, Barcelona, Spain.

Mullen, T. and Malouf, R. (2006). A preliminary investigation into sentiment analysis of informal political discourse. In *Proceedings of the AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs*, pages 159-162, Stanford, California, USA.

Nadeau, D., Sabourin, C., De Koninck, J., Matwin, S., and Turney, P. (2006). Automatic Dream Sentiment Analysis. In *Proccedings of the Workshop on Computational Aesthetics at the Twenty-First National Conference on Artificial Intelligence,* Boston, Massachusetts, USA. NRC 48725.

Nastase, V., Shirabad, J.S., Sokolova, M. and Szpakowicz, S. (2006). Learning Noun-Modifier Semantic Relations with Corpus-based and WordNet-based Features. In *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI 2006)*, pages 781-787, Boston, Massachusetts, USA.

Neviarouskaya, A., Prendinger, H., and Ishizuka, M. (2007a). Analysis of affect expressed through the evolving language of online communication. In *Proceedings of the 12th International Conference on Intelligent User Interfaces (IUI-07)*, pages 278-281, Honolulu, Hawaii, USA.

Neviarouskaya, A., Prendinger, H., and Ishizuka, M. (2007b). Narrowing the Social Gap among People involved in Global Dialog: Automatic Emotion Detection in Blog Posts, In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM 2007),* pages 293-294, Boulder, Colorado, USA.

Ortony, A., Clore, G.L., and Collins, A. (1988). *The Cognitive Structure of Emotions.* Cambridge University Press.

Osgood, C.E., Succi, G.J. and Tannenbaum, P.H. (1957). *The Measurement of Meaning.* Urbana: University of Illinois Press.

Owsley S., Sood S., Hammond K., Domain Specific Affective Classification of Documents. (2006). In *Proceedings of the AAAI-2006 Spring Symposium on Computational Approaches to Analysing Weblogs*, Stanford, California, USA.

Pang, B. and Lee, L. (2004). A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL'04),* pages 271-278, Barcelona, Spain.

Pang, B. and Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL'05),* pages 115-124, Ann Arbor, Michigan, USA.

Pang, B., Lee, L. and Vaithyanathan, S. (2002). Thumbs up? Sentiment Classification using Machine Learning Techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP-2002)*, pages 79-86, Philadelphia, Pennsylvania, USA.

Passonneau, R. (2004). Computing reliability for co-reference annotation. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC-2004)*, Lisbon, Portugal.

Passonneau, R. (2006). Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC-2006)*, Genoa, Italy.

Pennebaker, J.W., Francis, M. E. and Booth, R.J. (2001). *Linguistic Inquiry and Word Count LIWC2001*. Erlbaum Publishers, Mahwah, New Jersey.

Pennebaker, J.W., Mehl, M.R., & Niederhoffer, K. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*, 54, pages 547-577.

Picard, R.W. (1997). *Affective Computing*. MIT Press, ISBN 0262661152.

Plutchik, R. (1980). A General Psychoevolutionary Theory of Emotion. In Plutchik, R. and Kellerman, H. (eds.), *Emotion: Theory, Research and Experience: Vol. 1. Theories of Emotion* (3-33). New York: Academic.

Plutchik, R. (1980). *Emotion: A Psychoevolutionary Synthesis*. New York: Harper & Row.

Polanyi, L., and Zaenen, A. (2004). Contextual valence shifters. In *Computing Attitude and Affect in Text: Theory and Applications.* Edited by J. Shanahan, Y. Qu, and J. Wiebe. The Information Retrieval Series, Vol. 20, Springer, Dordrecht, The Netherlands, pages 1–9.

Quirk, R., S. Greenbaum, G. Leech, and J. Svartvik. (1985). *A Comprehensive Grammar of the English Language*. New York: Longman.

Ramshaw, L.A. and Marcus, M.P. (1995). Text chunking using transformation-based learning. In *Proceedings of the Third ACL Workshop on Very Large Corpora.*

Read, J. (2004). *Recognising affect in text using pointwise mutual information.* Master's thesis, University of Sussex.

Read, J. (2005). Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *Proceedings of the ACL 2005 Student Research Workshop*, pages 43-48, Ann Arbor, Michigan, USA.

Read, J., Hope D., and Carroll, J. (2007). Annotating expressions of Appraisal in English. To appear in the *Proceedings of the ACL-2007 Linguistic Annotation Workshop*, Prague, Czech Republic.

Resnik, P. (1995). Using information content to evaluate semantic similarity. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI-95)*, pages 448–453, Montreal, Canada.

Riloff, E. and Wiebe, J. (2003). Learning Extraction Patterns for Subjective Expressions. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP-2003)*, pages 105–112, Sapporo, Japan.

Riloff, E., Wiebe, J., and Wilson, T. (2003). Learning Subjective Nouns Using Extraction Pattern Bootstrapping. In *Proceedings of the 7th Conference on Natural Language Learning (CoNLL-2003)*, pages 25–32, Edmonton, Canada.

Riloff, E., Patwardhan, S. and Wiebe, J. (2006). Feature Subsumption for Opinion Analysis. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP-2006)*, pages 440-448, Sydney, Australia.

Roget, P.M. (1852). *Roget's Thesaurus of English Words and Phrases*. Harlow, Essex, England: Longman Group Limited.

Rubin, V.L., Stanton, J.M., Liddy, E.D. (2004). Discerning Emotions in Texts. In *Proceedings of the AAAI-2004 Symposium on Exploring Attitude and Affect in Text*, Stanford, California, USA.

Say, B. and Akman, V. (1996). Current Approaches to Punctuation in Computational Linguistics. *Computers and the Humanities*, 30(6):457–469.

Schlosberg, H. (1952). The description of facial expressions in terms of two dimensions. *Journal of Experimental Psychology,* 44, pages 229-237.

Singh, P., Lin, T., Mueller, E.T., Lim, G., Perkins, T. and Zhu, W.L. (2002). Open Mind Common Sense (OMCS): Knowledge acquisition from the general public. In *Proceedings of the First International Conference on Ontologies, Databases, and Applications of Semantics for Large Scale Information Systems*. LNCS. Springer.

Sleator, D. and Temperley, D. (1991). Parsing English with a Link Grammar. Carnegie Mellon University Computer Science technical report CMU-CS-91-196.

Sokolova, M., Nastase, V., Shah, M. and Szpakowicz, S. (2005a). Feature selection for electronic negotiation texts. In *Proceedings of Recent Advances in Natural Language Processing (RANLP'2005)*, pages 518 – 524, Borovets, Bulgaria.

Sokolova, M., Nastase, V., Szpakowicz, S. and Shah, M. (2005b). Analysis and Models of Language in Electronic Negotiations, *Issues in Intelligent Systems. Models and Techniques*, 197 - 211, EXIT, Warszawa.

Somasundaran, S., Wilson, T., Wiebe, J. and Stoyanov, V. (2007). QA with Attitude: Exploiting Opinion Type Analysis for Improving Question Answering in On-line Discussions and the News. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM-2007)*, Boulder, Colorado, USA.

Stone, P.J., Dunphy, D.C., Smith, M.S., Ogilvie, D.M., and associates. (1966). *The General Inquirer: A Computer Approach to Content Analysis*. The MIT Press.

Stoyanov, V., Cardie, C. and Wiebe, J. (2005). Multi-Perspective Question Answering Using the OpQA Corpus. In *Proceedings of the Joint Conference on Human Language Technology/Empirical Methods in Natural Language Processing (HLT/EMNLP 2005),* pages 923-930, Vancouver, Canada.

Strapparava, C. and Valitutti, A. (2004). WordNet-Affect: an affective extension of WordNet. In *Proceedings of 4th International Conference on Language Resources and Evaluation (LREC2004),* pages 1083 – 1086, Lisbon, Portugal.

Strapparava, C., Valitutti, A. and Stock, O. (2006). The affective weight of lexicon. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, pages 423–426, Genoa, Italy.

Strapparava, C., Valitutti, A. and Stock, O. (2007). Dances with Words. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI-07),* Hyderabad, India.

Taboada, M. and Grieve, J. (2004). Analyzing appraisal automatically. *In AAAI Spring Symposium on Exploring Attitude and Affect in Text.*

Tomkins, S.S. (1962). *Affect, imagery, consciousness* (Vol. 1). The positive affects. New York: Springer.

Turney, P.D. (2001). Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of the 12th European Conference on Machine Learning (ECML-2001)*, pages 491-502, Frieburg, Germany.

Turney, P.D. (2002). Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*, pages 417-424, Philadelphia, Pennsylvania, USA.

Turney, P.D. (2004). *Waterloo MultiText System: User's Guide.*

Turney, P.D., and Littman, M.L. (2003). Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems* (TOIS), 21(4), pages 315–346.

Watson, D. and Tellegen, A. (1985). Towards a consensual structure of mood. *Psychological Bulletin*, 98, pages 219-235.

Whitelaw, C., Garg, N., and Argamon, S. (2005a). Using Appraisal Taxonomies for Sentiment Analysis. In *Proceedings of MCLC-05, the Second Midwest Computational Linguistic Colloquium,* Columbus, Ohio, USA.

Whitelaw, C., Garg, N., and Argamon, S. (2005b). Using Appraisal Groups for Sentiment Analysis. In *Proceedings of the 14th ACM international conference on Information and Knowledge Management (CIKM'05)*, pages 625–631, Bremen, Germany.

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

Wiebe, J., Wilson, T., and Cardie, C. (2005). Annotating expressions of opinions and emotions in language. Language Resources and Evaluation, 39(2-3), pages 165-210.

Wilson, T. and Wiebe, J. (2003). Annotating opinions in the world press. In *Proceedings of the 4th SIGDIAL Workshop on Discourse and Dialog (SIGDIAL-03)*, Sapporo, Japan.

Wilson, T. and Wiebe, J. and Hoffmann, P. (2005). Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In *Proceedings of the Joint Conference on Human Language Technology/Empirical Methods in Natural Language Processing (HLT/EMNLP 2005),* pages 347-354, Vancouver, Canada.

Wilson, T., Wiebe, J., and Hwa, R. (2004). Just How Mad Are You? Finding Strong and Weak Opinion Clauses. In *Proceedings of the Nineteenth National Conference on Artificial Intelligence (AAAI-2004)*, pages 761–766, San Jose, California, USA.

Wilson, T., Wiebe, J., and Hwa, R. (2006). Recognizing strong and weak opinion clauses. *Computational Intelligence*, 22 (2): 73-99.

Witten, I.H. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques* (2nd Edition), Morgan Kaufmann, San Francisco

Yang, Y. and Pederson, J.O. (1997). A Comparative Study on Feature Selection in Text Categorization, In *Proceedings of the 14th International Conference on Machine Learning (ICML-1997)*, pages 412-420, Nashville, Tennessee, USA.

Yang, Y. and Liu, X. (1999). A re-examination of text categorization methods. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 42–49, New York, NY, USA, ACM Press.

Zhang, L., Barnden, J., Hendley, R., and Wallington, A. (2006). Exploitation in Affect Detection in Open-Ended Improvisational Text. In *Proceedings of the ACL Workshop on Sentiment and Subjectivity in Text,* pages 47-54, Sydney, Australia.