Chapter: Affect Detection in Texts

Carlo Strapparava and Rada Mihalcea

Abstract

The field of affective NLP, and in particular the emotion recognition in texts is a challenging topic. Nonetheless with current NLP techniques it is possible to approach the problem with interesting results, opening exciting applicative perspectives for the future. In this chapter we presented some explorations in dealing with automatic recognition of affect in text. We start describing some available lexical resources, the problem of emotion annotations to create a gold standard, and the Affective Text task at SemEval-2007. That task focused on the classification of emotions in news headlines, and was meant as an exploration of the connection between emotions and lexical semantics. Then we approach the problem of recognizing emotions in texts, presenting some state-of-the-art knowledge-based and corpus-based methods. We conclude the chapter presenting two promising lines of research in the field of affective NLP. The first one approaches the related task of humor recognition; the second proposes the exploitation of extra-linguistic features (e.g. music) for emotion detection.

Keywords: Affective Natural Language Processing, emotion annotation, affective lexicon.

1. Introduction

Emotions have been widely studied in psychology and behavior sciences, as they are an important element of human nature. For instance, emotions have been studied with respect to facial expressions (Ekman, 1977), action tendencies (Frijda, 1982), physiological activity (Ax, 1953), or subjective experience (Rivera, 1998). They have also attracted the attention of researchers in computer science, especially in the field of human computer interaction, where studies have been carried out on the

recognition of emotions through a variety of sensors (e.g., Picard, 1997). In contrast to the considerable work focusing on the nonverbal expression of emotions, surprisingly little research has explored how emotions are reflected verbally (Fussell, 2002; Ortony, Clore, & Foss, 1987b). Important contributions come from social psychologists, studying language as a way of expressing emotions (Osgood et al., 1975) (Pennbaker, 2002). From the perspective of computational linguistics, the determination of what an emotion is, is not an easy problem. Emotions are not linguistic constructs. However the most convenient access we have to them is *through the language*. This is very much true nowadays, in the web age, in which large quantity of texts (and some of them particularly affectively oriented e.g., blogs) are easily at disposal.

In computational linguistics, the automatic detection of emotions in texts is also becoming increasingly important from an applicative point of view. Consider for example the tasks of opinion mining and market analysis, affective computing, or natural language interfaces such as e-learning environments or educational/edutainment games. Possible beneficial effects of emotions on memory and attention of the users, and in general on fostering their creativity are also well known in the field of psychology.

For instance, the following represent examples of applicative scenarios in which affective analysis could make valuable and interesting contributions:

- *Sentiment Analysis*. Text categorization according to affective relevance, opinion exploration for market analysis, etc., are examples of applications of these techniques. While positive/negative valence annotation is an active area in sentiment analysis, we believe that a fine-grained emotion annotation could increase the effectiveness of these applications.

- *Computer Assisted Creativity*. The automated generation of evaluative expressions with a bias on certain polarity orientation is a key component in automatic personalized advertisement and persuasive communication. Possible applicative contexts can be creative computational environments that help producing what human graphic designers sometime completely manually do for TV/Web presentations (e.g., advertisements, news titles)
- *Verbal Expressivity in Human Computer Interaction*. Future human-computer interaction is expected to emphasize naturalness and effectiveness, and hence the integration of models of possibly many human cognitive capabilities, including affective analysis and generation. For example, the expression of emotions by synthetic characters (e.g., embodied conversational agents) is now considered a key element for their believability. Affective words selection and understanding is crucial for realizing appropriate and expressive conversations.

This chapter presents some explorations in dealing with automatic recognition of affect in text. We start describing some available lexical resources, the problem of emotion annotations to create gold standard, and the "Affective Text" task, presented at SemEval-2007. That task focused on the classification of emotions in news headlines, and was meant as an exploration of the connection between emotions and lexical semantics. Then we approach the problem of recognizing emotions expressed in texts, presenting some state-of-the-art knowledge-based and corpus-based methods. We conclude the chapter presenting two promising lines of research in the field of affective NLP. The first one approaches the related task of humor recognition; the second proposes the exploitation of extra-linguistic features (e.g. music) for emotion detection.

2. Affective Lexical Resources

The starting point of a computational linguistic approach to the study of emotion in text is the use of specific affective lexicons. (Ortony, Clore, & Foss, 1987a) is one of the first works that introduced the problem of the referential structure of the affective lexicon. In that work the authors conducted an analysis of about 500 words taken from the literature on emotions. Then they developed a taxonomy that helps isolating terms that explicitly refer to emotions.

In recent years, the research community developed several interesting resources that can be operatively exploited in natural language processing tasks that deal with affect. We briefly review some of them below.

General Enquirer. The General Inquirer (Stone et al., 1966) is basically a mapping tool, which maps dictionary-supplied categories to lists of words and word senses. The currently distributed version combines the "Harvard IV-4" dictionary content-analysis categories, the "Lasswell" dictionary content-analysis categories, and five categories based on the social cognition work of Semin and Fiedler (1988), making for 182 categories in all. Each category is a list of words and word senses. It uses stemming and disambiguation, for example it distinguishes between *race* as a contest, *race* as moving rapidly, and race as a group of people of common descent. A sketch of some categories from the General Inquirer is shown below.

- ...
- XI. **Emotions** (EMOT): anger, fury, distress, happy, etc.
- XII. **Frequency** (FREQ): occasional, seldom, often, etc.
- XIII. **Evaluative Adjective** (EVAL): good, bad, beautiful, hard, easy, etc.
- XIV. **Dimensionality Adjective** (DIM): big, little, short, long, tall, etc.
- XV. **Position Adjective** (POS): low, lower, upper, high, middle, first, fourth, etc.
- XVI. **Degree Adverbs** (DEG): very, extremely, too, rather, somewhat. . .
- ...

SentiWordNet. SentiWordNet (Esuli & Sebastiani, 2006) is a lexical resource that focuses on polarity of subjective terms, i.e. whether a term that is a marker of

opinionated content has a *positive* or a *negative* connotation. In practice each synset *s* in WordNet is associated to three numerical scores *Obj(s)*, *Pos(s)* and *Neg(s)*, describing how Objective, Positive, and Negative the terms contained in the synset are. These three scores are derived by combining the results produced by a committee of eight ternary classifiers. These scores are interconnected, in particular the objectivity score can be calculated as: $Obj(s) = 1-[Pos(s)+Neg(s)]$. The rationale behind this formula is that a given text has a factual nature (i.e. describes objectively a given situation or event), if there is no presence of a positive or a negative opinion on it; otherwise it expresses an opinion on its subject matter.

While SentiWordNet does not address emotions directly, it can be exploited whenever detection along the positive vs. negative dimension is required.

Affective Norms for English Words. The Affective Norms for English Words (ANEW) provides a set of normative emotional ratings for a large number of words in the English language (Bradley & Lang, 1999). In particular, the goal was to develop a set of verbal materials that have been rated, as perceived by readers[1], in terms of *pleasure*, *arousal*, and *dominance*. This view is founded on the semantic differential, in which factor analyses conducted on a wide variety of verbal judgments indicated that the variance in emotional assessments was accounted for by those three major dimensions: the two primary dimensions were one of affective valence (from pleasant to unpleasant) and one of arousal (from calm to excited). A third, less strongly related dimension was variously called 'dominance' or 'control'. To assess these three dimensions, an affective rating system (the Self-Assessment Manikin) originally formulated by (Lang, 1980) was exploited. Bradley and Lang (1994) had determined that this rating system correlates well with factors of pleasure and arousal obtained

[1] In ANEW the communicative perspective is that the term acts as a stimulus to elicit a particular emotion in the reader.

using the more extended verbal Semantic Differential Scale (Mehrabian & Russell 1974). For an example of how the resource has been exploited computationally, see (Calvo and Kim, 2012).

There are 1034 words currently normed in ANEW, with the ratings respectively for pleasure, arousal, and dominance. Each rating scale runs from 1 to 9, with a rating of '1' indicating a low value on each dimension (e.g., low pleasure, low arousal, low dominance) and '9' indicating a high value on each dimension (high pleasure, high arousal, high dominance). An excerpt from the ANEW lexical resource is presented below.

Word	Valence (Mean)	Valence (SD)	Arousal (Mean)	Arousal (SD)	Dominance (Mean)	Dominance (SD)
abduction	2.76	2.06	5.53	2.43	3.49	2.38
abortion	3.50	2.30	5.39	2.80	4.59	2.54
absurd	4.26	1.82	4.36	2.20	4.73	1.72
abundance	6.59	2.01	5.51	2.63	5.80	2.16
abuse	1.80	1.23	6.83	2.70	3.69	2.94
acceptance	7.98	1.42	5.40	2.70	6.64	1.91
accident	2.05	1.19	6.26	2.87	3.76	2.22
ace	6.88	1.93	5.50	2.66	6.39	2.31
ache	2.46	1.52	5.00	2.45	3.54	1.73
achievement	7.89	1.38	5.53	2.81	6.56	2.35

WordNet Affect. The development of WordNet-Affect (Strapparava & Valitutti, 2004a; Strapparava, Valitutti, & Stock, 2006) was motivated by the need of a lexical resource with explicit fine-grained emotion annotations.

As claimed by Ortony et al. (1987b), we have to distinguish between words directly referring to emotional states (e.g. "fear", "cheerful") and those having only an indirect reference that depends on the context (e.g., words that indicate possible emotional causes as "monster" or emotional responses as "cry"). We call the former *direct* affective words and the latter *indirect* affective words. The rationale behind WordNet-Affect is to provide a resource with fine-grained emotion annotations only for the direct affective lexicon, leaving to other techniques the capabilities to classify

the emotional load of the indirect affective words. All words can potentially convey affective meaning. Each of them, even those more apparently neutral, can evoke pleasant or painful experiences. While some words have emotional meaning with respect to the individual story, for many others the affective power is part of the collective imagination (e.g., “mum”, “ghost”, “war” etc.). Thus, in principle it could be incorrect to conduct an a-priori annotation on the whole lexicon. Strapparava, Valitutti, and Stock (2006) suggest to use corpus-based driven (possibly exploiting specific corpora for particular purposes) for inferring the emotional load of generic words. More specifically they proposed a semantic similarity function, acquired automatically in an unsupervised way from a large corpus of texts, which allows us to put into relation generic concepts with direct emotional categories. We will describe a similar approach in Section 4.

WordNet-Affect is an extension of the WordNet database (Fellbaum, 1998), including a subset of synsets suitable to represent affective concepts. Similarly to the annotation method for domain labels (Magnini & Cavaglià, 2000), a number of WordNet synsets were assigned to one or more affective labels (*a-labels*). In particular, the affective concepts representing emotional state are individuated by synsets marked with the a-label EMOTION. There are also other a-labels for those concepts representing moods, situations eliciting emotions, or emotional responses. WordNet-Affect is freely available for research purpose at `http://wndomains.fbk.eu`. See (Strapparava & Valitutti, 2004b) for a complete description of the resource.

The emotional categories are hierarchically organized, in order to specialize synsets with a-label EMOTION. Regarding emotional valence, four additional a-labels are introduced: POSITIVE, NEGATIVE, AMBIGUOUS, NEUTRAL. The first one corresponds

to “positive emotions”, related to words expressing positive emotional states. It includes synsets such as **joy#1** or **enthusiasm#1**. Similarly the NEGATIVE a-label identifies “negative emotions”,

	# Synsets	# Words	# Senses
Nouns	280	539	564
Adjectives	342	601	951
Verbs	142	294	430
Adverbs	154	203	270
Total	918	1637	2215

Table 1: Number of elements in the emotional hierarchy.

for example labeling synset such as **anger#1** or **sadness#1**. Synsets representing affective states whose valence depends on semantic context (e.g., **surprise#1**) were marked with the tag AMBIGUOUS. Finally, synsets referring to mental states that are generally considered affective but are not characterized by valence, were marked with the tag NEUTRAL.

A-Labels	*Valence*	*Examples of word senses*
JOY	positive	noun **joy#1**, adjective **elated#2**, verb **gladden#2**, adverb **gleefully#1**
LOVE	positive	noun **love#1**, adjective **loving#1**, verb **love#1**, adverb **fondly#1**
APPREHENSION	negative	noun **apprehension#1**, adjective **apprehensive#3**, adverb **anxiously#1**
SADNESS	negative	noun **sadness#1**, adjective **unhappy#1**, verb **sadden#1**, adverb **deplorably#1**
SURPRISE	ambiguous	noun **surprise#1**, adjective **surprised#1**, verb **surprise#1**
APATHY	neutral	noun **apathy#1**, adjective **apathetic#1**, adverb **apathetically#1**
NEGATIVE-FEAR	negative	noun **scare#2**, adjective **afraid#1**, verb **frighten#1**, adverb **horryfyingly#1**
POSITIVE-FEAR	positive	noun **frisson#1**
POSITIVE-EXPECTATION	positive	noun **anticipation#1**, adjective **cliff-hanging#1**, verb **anticipate#1**

Table 2: Some of the emotional categories in WordNet-Affect and some corresponding word senses

Positive	Negative	Ambiguous	Neutral	Total
97	156	20	7	*280*

Table 3: Valence distribution of emotional categories.

3. Annotating Texts with Emotions

In order to explore the classification of emotions in texts, gold standards are required, consisting of manual emotion annotations. This is a rather difficult task, in particular given its subjectivity, where humans themselves often disagree on the emotions present in a given text. The task can also be very time consuming, even more so when

for the purpose of reaching higher inter-annotator agreements, a large number of annotations are sought. Because of its specificity, the granularity of the task is on short texts (e.g. single sentences, news headlines).

Previous work on emotion annotation of text (Alm, Roth, & Sproat, 2005; Strapparava & Mihalcea, 2007; Aman & Szpakowicz, 2008) has usually relied on the six basic emotions proposed by (Ekman, 1993): ANGER, DISGUST, FEAR, JOY, SADNESS, SURPRISE. We also focus on these six emotions in this chapter, and review two annotation efforts: one that targeted the annotation of emotions in lyrics using crowdsourcing, and one that aimed at building a gold standard consisting of news headlines annotated for emotion.

3.1 Emotion Annotations via Crowdsourcing

In a recent project concerned with the classification of emotions in songs (Strapparava, Mihalcea, & Battocchi, 2012a; Mihalcea & Strapparava, 2012a), we introduced a novel corpus consisting of 100 songs annotated for emotions. The songs were sampled from among some of the most popular pop, rock, and evergreen songs, such as *Dancing Queen* by ABBA, *Hotel California* by Eagles, *Let it Be* by The Beatles.

To collect the annotations, we used the Amazon Mechanical Turk service, which was previously found to produce reliable annotations with a quality comparable to those generated by experts (Snow et al. 2008).

The annotations were collected at line level, with a separate annotation for each of the six emotions. We collected numerical annotations using a scale between 0 and 10, with 0 corresponding to the absence of an emotion, and 10 corresponding to the highest intensity. Each HIT (i.e., annotation session) contains an entire song, with a number of lines ranging from 14 to 110, for an average of 50 lines per song.

Annotation Guidelines. The annotators were instructed to: (1) Score the emotions from the writer perspective, not their own perspective; (2) Read and interpret each line in context; i.e., they were asked to read and understand the entire song before producing any annotations; (3) Produce the six emotion annotations independent from each other, accounting for the fact that a line could contain none, one, or multiple emotions. In addition to the lyrics, the song was also available online, so they could listen to it in case they were not familiar with it. The annotators were also given three different examples to illustrate the annotation.

Controlling for Annotation Errors. While the use of crowdsourcing for data annotation can result in a large number of annotations in a very short amount of time, it also has the drawback of potential spamming that can interfere with the quality of the annotations. To address this aspect, we used two different techniques to prevent inappropriate annotations. First, in each song we inserted a "checkpoint" at a random position in the song – a fake line that reads "Please enter 7 for each of the six emotions." Those annotators who did not follow this concrete instruction were deemed as spammers who produce annotations without reading the content of the song, and thus removed. Second, for each remaining annotator, we calculated the Pearson correlation between her emotion scores and the average emotion scores of all the other annotators. Those annotators with a correlation with the average of the other annotators below 0.4 were also removed, thus leaving only the reliable annotators in the pool.

For each song, we started by asking for ten annotations. After spam removal, we were left with about two-five annotations per song. The final annotations were produced by averaging the emotions scores produced by the reliable annotators. Figure 2 shows an example of the emotion scores produced for two lines. The overall

correlation between the remaining reliable annotators was 0.73, which represents a strong correlation.

Emotions in the Corpus of 100 Songs. For each of the six emotions, Table 4 shows the number of lines that had that emotion present (i.e., the score of the emotion was different from 0), as well as the average score for that emotion over all 4,976 lines in the corpus. Perhaps not surprisingly, the emotions that are dominant in the corpus are JOY and SADNESS – which are the emotions that are often invoked by people as the reason behind a song.

Emotion	Number lines	Average
ANGER	2,516	0.95
DISGUST	2,461	0.71
FEAR	2,719	0.77
JOY	3,890	3.24
SADNESS	3,840	2.27
SURPRISE	2,982	0.83

Table 4: Emotions in the corpus of 100 songs: number of lines including a certain emotion, and average emotion score computed over all the 4,976 lines.

Note that the emotions do not exclude each other: i.e., a line that is labeled as containing joy may also contain a certain amount of SADNESS, which is the reason for the high percentage of songs containing both JOY and SADNESS. The emotional load for the overlapping emotions is however very different. For instance, the lines that have a joy score of 5 or higher have an average SADNESS score of 0.34. Conversely, the lines with a SADNESS score of 5 or higher have a joy score of 0.22.

3.2 SemEval 2007 "Affective Text" Task

In the context of SemEval 2007[2], we organized a task focused on the classification of emotions and valence (i.e., positive/negative polarity) in news headlines, and was

[2]http://nlp.cs.swarthmore.edu/semeval/

meant as an exploration of the connection between emotions and lexical semantics. In this section, we describe the data set used in the evaluation.

Task Definition. We proposed to focus on the emotion classification of news headlines extracted from news web sites. The news headlines typically consist of a few words and are often written by creative people with the intention to "provoke" emotions, and consequently to attract the readers' attention. These characteristics make the news headlines particularly suitable for use in an automatic emotion recognition setting, as the affective/emotional features (if present) are guaranteed to appear in these short sentences.

The structure of the task was as follows:

Corpus: News titles, extracted from news web sites (such as Google news, CNN) and/or newspapers. In the case of web sites, we can easily collect a few thousand titles in a short amount of time.

Objective: Provided a set of predefined six emotion labels (i.e. Anger, Disgust, Fear, Joy, Sadness, Surprise), classify the titles with the appropriate emotion label and/or with a valence indication (i.e. positive/negative) (positive/negative)

The emotion labeling and valence classification were seen as independent tasks, and thus a team was able to participate in one or both tasks. The task was carried out in an unsupervised setting, and consequently no training was provided. The reason behind this decision is that we wanted to emphasize the study of emotion lexical semantics, and avoid biasing the participants toward simple "text categorization" approaches. Nonetheless supervised systems were not precluded and in this case participating teams were allowed to create their own supervised training sets.

Participants were free to use any resources they wished. We provided a set words extracted from WordNet Affect (Strapparava & Valitutti, 2004a), relevant to the six emotions of interest. However the use of this list of words was entirely optional.

Data Set. The data set consists of news headlines drawn from major newspapers such as New York Times, CNN, and BBC News, as well as from the Google News search engine. We decided to focus our attention on headlines for two main reasons. First, news have typically a high load of emotional content, as they describe major national or worldwide events, and are written in a style meant to attract the attention of the readers. Second, the structure of headlines was appropriate for our goal of conducting sentence-level annotations of emotions.

Two data sets were made available: a development data set consisting of 250 annotated headlines, and a test data set with 1,000 annotated headlines.

Data Annotation. To perform the annotations, we developed a Web-based annotation interface that displayed one headline at a time, together with six slide bars for emotions and one slide bar for valence. The interval for the emotion annotations was set to [0, 100], where 0 means the emotion is missing from the given headline, and 100 represents maximum emotional load. The interval for the valence annotations was set to [−100, 100], where 0 represents a neutral headline, −100 represents a highly negative headline, and 100 corresponds to a highly positive headline.

Unlike previous annotations of sentiment or subjectivity (Wiebe, Wilson, & Cardie, 2005; Pang & Lee, 2004), which typically relied on binary 0/1 annotations, we decided to use a finer-grained scale, hence allowing the annotators to select different degrees of emotional load.

Six annotators independently labeled the test data set. The annotators were instructed to select the appropriate emotions for each headline based on the presence of words or phrases with emotional content, as well as the overall feeling invoked by the headline. Annotation examples were also provided, including examples of headlines bearing two or more emotions to illustrate the case where several emotions were jointly applicable. Finally, the annotators were encouraged to follow their "first intuition," and to use the full-range of the annotation scale bars.

The final annotation labels were created as the average of the six independent annotations, after normalizing the set of annotations provided by each annotator for each emotion to the 0-100 range. Table 5 shows three sample headlines in the data set, along with their final gold standard annotations.

	Emotions						
	Anger	Disgust	Fear	Joy	Sadness	Surprise	Valence
Inter Milan set Serie A win record	2	0	0	50	0	9	50
After Iraq trip, Clinton proposes war limits	8	0	8	53	13	25	38
7 dead in apartment building fire	14	2	47	0	86	10	-86

Table 5: Sample headlines and manual annotations of emotions

Inter-Annotator Agreement. We conducted inter-tagger agreement studies for each of the six emotions and for the valence annotations. The agreement evaluations were carried out using the Pearson correlation measure, and are shown in Table 6. To measure the agreement among the six annotators, we first measured the agreement between each annotator and the average of the remaining five annotators, followed by an average over the six resulting agreement figures.

Emotions	
Anger	49.55
Disgust	44.51
Fear	63.81
Joy	59.91
Sadness	68.19
Surprise	36.07
Valence	
Valence	78.01

Table 6: Inter-annotator agreement

Fine-grained and Coarse-grained Evaluations. Fine-grained evaluations were conducted using the Pearson measure of correlation between the system scores and the gold standard scores, averaged over all the headlines in the data set.

We have also run a coarse-grained evaluation, where each emotion was mapped to a 0/1 classification (0 = [0,50), 1 = [50,100]), and each valence was mapped to a -1/0/1 classification (-1 = [-100,-50], 0 = (-50,50), 1 = [50,100]). For the coarse-grained evaluations, we calculated accuracy, precision, and recall. Note that the accuracy is calculated with respect to all the possible classes, and thus it can be artificially high in the case of unbalanced datasets (as some of the emotions are, due to the high number of neutral headlines). Instead, the precision and recall figures exclude the neutral annotations.

4. Recognizing Emotions in Texts

In this section we present several algorithms for detecting emotion in texts, ranging from simple heuristics (e.g., directly checking specific affective lexicons) to more refined algorithms (e.g., checking similarity in a latent semantic space in which explicit representations of emotions are built, and exploiting Naïve Bayes classifiers trained on mood-labeled blogposts). It is worth noting that the proposed methodologies are either completely unsupervised or, when supervision is used, the training data can be easily collected from online mood-annotated materials. To give an idea of difficulties of the task we present the evaluation of the algorithms and a comparison with the systems that participated in the SemEval 2007 task on "Affective Text." As noted in Section 3, the focus is on short texts (e.g. news titles, single sentences, lines of lyrics).

4.1 Affective Semantic Similarity

As we have seen above, a crucial issue is to have a mechanism for evaluating the emotional load of generic terms. We introduce in this section a possible methodology to deal with the problem, based on the similarity among generic terms and affective lexical concepts. To this aim we estimated term similarity from a large-scale corpus. In particular we implemented a variation of Latent Semantic Analysis (LSA) in order to obtain a vector representation for words, texts and synsets.

In LSA (Deerwester et al., 1990), term co-occurrences in the documents of the corpus are captured by means of a dimensionality reduction operated by a Singular Value Decomposition (SVD) on the term-by-document matrix. SVD is a well-known operation in linear algebra, which can be applied to any rectangular matrix in order to find correlations among its rows and columns. In our case, SVD decomposes the term-by-document matrix **T** into three matrices $\mathbf{T} = \mathbf{U}\mathbf{\Sigma_k}\mathbf{V}^{\mathbf{T}}$ where $\mathbf{\Sigma_k}$ is the diagonal $k \times k$ matrix containing the k singular values of **T**, $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_k$, and **U** and **V** are column-orthogonal matrices. When the three matrices are multiplied together the original term-by-document matrix is re-composed. Typically we can choose $k' \ll k$ obtaining the approximation $\mathbf{T} \simeq \mathbf{U}\mathbf{\Sigma_{k'}}\mathbf{V}^{\mathbf{T}}$.

LSA can be viewed as a way to overcome some of the drawbacks of the standard vector space model (sparseness and high dimensionality). In fact, the LSA similarity is computed in a lower dimensional space, in which second-order relations among terms and texts are exploited. For the experiments reported in this paper, we run the SVD operation on the British National Corpus[3], using $k' = 400$ dimensions.

[3]The British National Corpus is a very large (over 100 million words) corpus of modern English, both spoken and written (BNC-Consortium 2000).

The resulting LSA vectors can be exploited to estimate both term and document similarity. Regarding document similarity, Latent Semantic Indexing (LSI) is a technique that allows us to represent a document by means of a LSA vector. In particular, we used a variation of the *pseudo-document* methodology described in (Berry, 1992). This variation takes into account also a *tf-idf* weighting schema (see Gliozzo & Strapparava, 2005 for more details). Each document can be represented in the LSA space by summing up the normalized LSA vectors of all the terms contained in it. Also a synset in WordNet (and then an emotional category) can be represent in the LSA space, performing the pseudo-document technique on all the words contained in the synset. Thus it is possible to have a vectorial representation of each emotional category in the LSA space (i.e., the *emotional vectors*). With an appropriate metric (e.g., cosine), we can compute a similarity measure among terms and affective categories. We defined the *affective weight* as the similarity value between an emotional vector and an input term vector.

For example, the term "sex" shows high similarity with respect to the positive emotional category AMOROUSNESS, with the negative category MISOGYNY, and with the ambiguous valence tagged category AMBIGUOUS_EXPECTATION. The noun "gift" is highly related to the emotional categories: LOVE (with positive valence), COMPASSION (with negative valence), SURPRISE (with ambiguous valence), and INDIFFERENCE (with neutral valence).

In conclusion, the vectorial representation in the Latent Semantic Space allows us to represent in a uniform way emotional categories, terms, concepts and possibly full documents. The affective weight function can be used in order to select the emotional categories that can best express or evoke valenced emotional states with respect to input term. Moreover, it allows us to individuate a set of terms that are

semantically similar to the input term and that share with it the same affective constraints (e.g. emotional categories with the same value of valence).

For example, given the noun university as input-term, it is possible to check for related terms that have a positive affective valence, possibly focusing only on some specific emotional categories (e.g. sympathy). On the other hand, given two terms, it is possible to check whether they are semantically related, and with respect to which emotional category. Table 7 shows a portion of the affective lexicon related to "university" with some emotional categories grouped by valence.

Variations of this technique, i.e. exploiting non-negative matrix factorization (NMF) or probabilistic LSA are reported in (Calvo and Kim, 2012).

Related Generic Terms	*Positive Emotional Category*	*Emotional Weight*
university	ENTHUSIASM	0.36
professor	SYMPATHY	0.56
scholarship	DEVOTION	0.72
achievement	ENCOURAGEMENT	0.76
	Negative Emotional Category	
university	DOWNHEARTEDNESS	0.33
professor	ANTIPATHY	0.46
study	ISOLATION	0.49
scholarship	MELANCHOLY	0.53

Table 7: Some terms related to "university" through some emotional categories

4.2 Knowledge-based Emotion Classification

We can also approach the task of emotion recognition by exploiting the use of words in a text, and in particular their co-occurrence with words that have explicit affective meaning.

For this method, as far as direct affective words are concerned, we followed the classification found in WordNet-Affect. In particular, we collected six lists of affective words by using the synsets labeled with the six emotions considered in our data set. Thus, as a baseline, we implemented a simple algorithm that checks the

presence of this direct affective words in the headlines, and computes a score that reflects the frequency of the words in this affective lexicon in the text.

A crucial aspect is the availability of a mechanism for evaluating the semantic similarity among "generic" terms and affective lexical concepts. For this purpose, we exploited the affective semantic similarity described in the previous section. We acquired a LSA space from the British National Corpus[4]. As we have seen, LSA yields a vector space model that allows for a *homogeneous* representation (and hence comparison) of words, word sets, sentences and texts. Then, regardless of how an emotion is represented in the LSA space, we can compute a similarity measure among (generic) terms in an input text and affective categories. In the LSA space, an emotion can be represented at least in three ways: (i) the vector of the specific word denoting the emotion (e.g. **anger**), (ii) the vector representing the synset of the emotion (e.g., {**anger**, **choler**, **ire**}), and (iii) the vector of all the words in the synsets labeled with the emotion. Here we describe experiments with all these representations.
We have implemented four different systems for emotion analysis by using the knowledge-based approaches.

1. WN-AFFECT PRESENCE, which is used as a baseline system, and which annotates the emotions in a text simply based on the presence of words from the WordNet Affect lexicon.
2. LSA SINGLE WORD, which calculates the LSA similarity between the given text and each emotion, where an emotion is represented as the vector of the specific word denoting the emotion (e.g., joy).
3. LSA EMOTION SYNSET, where in addition to the word denoting an emotion, its synonyms from the WordNet synset are also used.
4. LSA ALL EMOTION WORDS, which augments the previous set by adding the words in all the synsets labeled with a given emotion, as found in WordNet Affect.

[4]Other more specific corpora could also be considered, to obtain a more domain oriented similarity.

The results obtained with each of these methods, on the corpus of news headlines described in Section 3.2, are presented below in Table 10.

4.3 Corpus-based Emotion Classification

In addition to the experiments based on WordNet-Affect, we also present corpus-based experiments relying on blog entries from LiveJournal.com. We used a collection of blogposts annotated with moods that were mapped to the six emotions used in the classification. While every blog community practices a different genre of writing, LiveJournal.com blogs seem to more closely recount the goings-on of everyday life than any other blog community.

The indication of the mood is optional when posting on LiveJournal, therefore the mood-annotated posts used here are likely to reflect the true mood of the blog authors, since they were explicitly specified without particular coercion from the interface. Our corpus consists of 8,761 blogposts, with the distribution over the six emotions shown in Table 8. This corpus is a subset of the corpus used in the experiments reported in (Mishne, 2005).

In a pre-processing step, all the SGML tags were removed, and only the body of the blogposts was kept, which was then passed through a tokenizer. Only blogposts with a length within a range comparable to the one of the headlines, i.e. 100-400 characters, were kept. The average length of the blogposts in the final corpus is 60 words / entry. Six sample entries are shown in Table 9.

The blogposts were then used to train a Naïve Bayes classifier, where for each emotion we used the blogs associated with it as positive examples, and the blogs associated with all the other five emotions as negative examples.

Emotion	LiveJournal mood	Number of blogposts
ANGER	angry	951
DISGUST	disgusted	72
FEAR	scared	637
JOY	happy	4,856
SADNESS	sad	1,794
SURPRISE	surprised	451

Table 8: Blogposts and mood annotations extracted from LiveJournal

4.4 Evaluation on SemEval 2007 task

The five systems (four knowledge-based and one corpus-based) are evaluated on the data set of 1,000 newspaper headlines. As mentioned earlier, both fine-grained and coarse-grained evaluations can be conducted. Table 10 shows the results obtained by each system for the annotation of the six emotions. The best results obtained according to each individual metric are marked in bold.

As expected, different systems have different strengths. The system based exclusively on the presence of words from the WordNet-Affect lexicon has the highest precision at the cost of low recall. Instead, the LSA system using all the emotion words has by far the largest recall, although the precision is significantly lower. In terms of performance for individual emotions, the system based on blogs gives the best results for joy, which correlates with the size of the training data set (joy had the largest number of blogposts). The blogs are also providing the best results for anger (which also had a relatively large number of blogposts). For all the other emotions, the best performance is obtained with the LSA models.

We also compared our results with those obtained by three systems participating in the Semeval emotion annotation task: SWAT, UPAR7 and UA. Table 11 shows the results obtained by these systems on the same data set, using the same evaluation metrics. We briefly describe below each of these three systems:

UPAR7 (Chaumartin, 2007) is a rule-based system using a linguistic approach. A first pass through the data "uncapitalizes" common words in the news title. The system then used the Stanford syntactic parser on the modified titles, and identifies what is being said about the main subject by exploiting the dependency graph obtained from the parser. Each word is first rated separately for each emotion and then the main subject rating is boosted. The system uses a combination of SentiWordNet (Esuli & Sebastiani, 2006) and WordNet-Affect (Strapparava & Valitutti, 2004b), which were semi-automatically enriched on the basis of the original trial data provided during the Semeval task.

UA (Kozareva et al., 2007) uses statistics gathered from three search engines (MyWay, AlltheWeb and Yahoo) to determine the kind and the amount of emotion in each headline. Emotion score are obtained by using Pointwise Mutual Information (PMI). First, the number of documents obtained from the three Web search engines using a query that contains all the headline words and an emotion (the words occur in an independent proximity across the Web documents) is divided by the number of documents containing only an emotion and the number of documents containing all the headline words. Second, an associative score between each content word and an emotion is estimated and used to weight the final PMI score. The final results are normalized to the 0-100 range.

ANGER
I am so angry. Nicci can't get work off for the Used's show on the 30th, and we were stuck in traffic for almost 3 hours today, preventing us from seeing them. bastards
DISGUST
It's time to snap out of this. It's time to pull things together. This is ridiculous. I'm going nowhere. I'm doing nothing.
FEAR
He might have lung cancer. It's just a rumor...but it makes sense. is very depressed and that's just the beginning of things
JOY
This week has been the best week I've had since I can't remem ber when! I have been so hyper all week, it's been awesome!!!
SADNESS
Oh and a girl from my old school got run over and died the other day which is horrible, especially as it was a very small village school so everybody knew her.
SURPRISE
Small note: French men shake your hand as they say good morning to you. This is a little shocking to us fragile Americans, who are used to waving to each other in greeting.

Table 9: Sample blogposts labeled with moods corresponding to the six emotions

SWAT (Katz, Singleton, & Wicentowski, 2007) is a supervised system using an unigram model trained to annotate emotional content. Synonym expansion on the emotion label words is also performed, using the Roget Thesaurus. In addition to the development data provided by the task organizers, the SWAT team annotated an additional set of 1000 headlines, which was used for training.

For an overall comparison, the average over all six emotions for each system was calculated. Table 12 shows the overall results obtained by the five systems described above and by the three Semeval systems. The best results in terms of fine-grained evaluations are obtained by the UPAR7 system, which is perhaps due to the deep syntactic analysis performed by this system. Our systems give however the best performance in terms of coarse-grained evaluations, with the WordNet-Affect presence providing the best precision, and the LSA all emotion words leading to the highest recall and F-measure.

	Fine	Coarse		
	r	Prec.	Rec.	F1
ANGER				
WordNet-Affect presence	12.08	**33.33**	3.33	6.06
LSA single word	8.32	6.28	63.33	11.43
LSA emotion synset	17.80	7.29	86.67	13.45
LSA all emotion words	5.77	6.20	**88.33**	11.58
NB trained on blogs	**19.78**	13.68	21.67	**16.77**
DISGUST				
WordNet-Affect presence	-1.59	0	0	-
LSA single word	**13.54**	**2.41**	70.59	**4.68**
LSA emotion synset	7.41	1.53	64.71	3.00
LSA all emotion words	8.25	1.98	**94.12**	3.87
NB trained on blogs	4.77	0	0	-
Fear				
WordNet-Affect presence	24.86	**100.00**	1.69	3.33
LSA single word	**29.56**	12.93	**96.61**	**22.80**
LSA emotion synset	18.11	12.44	94.92	22.00
LSA all emotion words	10.28	12.55	86.44	21.91
NB trained on blogs	7.41	16.67	3.39	5.63
JOY				
WordNet-Affect presence	10.32	**50.00**	0.56	1.10
LSA single word	4.92	17.81	47.22	25.88
LSA emotion synset	6.34	19.37	72.22	30.55
LSA all emotion words	7.00	18.60	**90.00**	30.83
NB trained on blogs	**13.81**	22.71	59.44	**32.87**
SADNESS				
WordNet-Affect presence	8.56	**33.33**	3.67	6.61
LSA single word	8.13	13.13	55.05	21.20
LSA emotion synset	13.27	14.35	58.71	**23.06**
LSA all emotion words	10.71	11.69	**87.16**	20.61
NB trained on blogs	**16.01**	20.87	22.02	21.43
SURPRISE				
WordNet-Affect presence	3.06	**13.04**	4.68	6.90
LSA single word	9.71	6.73	67.19	12.23
LSA emotion synset	12.07	7.23	89.06	13.38
LSA all emotion words	**12.35**	7.62	**95.31**	**14.10**
NB trained on blogs	3.08	8.33	1.56	2.63

Table 10: Performance of the proposed algorithms

	Fine		Coarse	
	r	Prec.	Rec.	F1
ANGER				
SWAT	24.51	12.00	5.00	7.06
UA	23.20	12.74	**21.6**	16.03
UPAR7	**32.33**	**16.67**	1.66	3.02
DISGUST				
SWAT	**18.55**	0.00	0.00	-
UA	16.21	0.00	0.00	-
UPAR7	12.85	0.00	0.00	-
FEAR				
SWAT	32.52	25.00	14.40	18.27
UA	23.15	16.23	**26.27**	20.06
UPAR7	**44.92**	**33.33**	2.54	4.72
JOY				
SWAT	**26.11**	35.41	**9.44**	14.91
UA	2.35	40.00	2.22	4.21
UPAR7	22.49	**54.54**	6.66	11.87
SADNESS				
SWAT	38.98	32.50	11.92	17.44
UA	12.28	25.00	0.91	1.76
UPAR7	**40.98**	**48.97**	**22.02**	30.38
SURPRISE				
SWAT	11.82	11.86	10.93	11.78
UA	7.75	**13.70**	**16.56**	15.00
UPAR7	**16.71**	12.12	1.25	2.27

Table 11: Results of the systems participating in the SemEval task for emotion annotations

5. Further Directions

Affect detection from text only started to be explored quite recently, so several new directions will be probably developed in the future. In the following section we present two promising lines of research. The first one approaches the related task of humor recognition, the second proposes the exploitation of extra-linguistic features (e.g. music) for emotion detection.

5.1 Humor recognition

Of all the phenomena, which fall under the study of emotions, humor is one of the least explored from a computational point of view. Humor involves both cognitive and emotional processes, and understanding its subtle mechanisms is certainly a challenge. Nonetheless, given the importance of humor in our everyday life, and the increasing importance of computers in our work and entertainment, we believe that

studies related to computational humor will become increasingly important in fields such as human-computer interaction, intelligent interactive entertainment, and computer-assisted education.

Previous work in computational humor has focused mainly on the task of humor generation (Stock & Strapparava, 2003; Binsted & Ritchie, 1997), and very few attempts have been made to develop systems for automatic humor recognition (Taylor & Mazlack, 2004). In (Mihalcea & Strapparava, 2006), the authors explored the applicability of computational approaches to the recognition and use of verbally expressed humor.

Since a deep comprehension of humor in all of its aspects is probably too ambitious and beyond the existing computational capabilities, the investigation was restricted to the type of humor found in one-liners. A one-liner is a short sentence with comic effects and an interesting linguistic structure: simple syntax, deliberate use of rhetoric devices (e.g. alliteration, rhyme), and frequent use of creative language constructions meant to attract the readers' attention.

To test the hypothesis that automatic classification techniques represent a viable approach to humor recognition, we needed in the first place a data set consisting of both humorous (positive) and non-humorous (negative) examples. Such data sets can be used to automatically learn computational models for humor recognition, and at the same time evaluate the performance of such models.
We tested two different sets of "negative" examples (see Table 13):

1. *Reuters* titles, extracted from news articles published in the Reuters newswire over a period of one year (8/20/1996 – 8/19/1997) (Lewis et al., 2004). The titles consist of short sentences with simple syntax, and are often phrased to catch the readers attention (an effect similar to the one rendered by one-liners).

2. *Proverbs* extracted from an online proverb collection. Proverbs are sayings that transmit, usually in one short sentence, important facts or experiences that are considered true by many people. Their property of being condensed, but memorable sayings make them very similar to the one-liners. In fact, some one-liners attempt to reproduce proverbs, with a comic effect, as in e.g. "*Beauty is in the eye of the beer holder*", derived from "*Beauty is in the eye of the beholder*".

The dimension of the datasets is 16,000 one-liners, with the same number respectively for titles and proverbs.

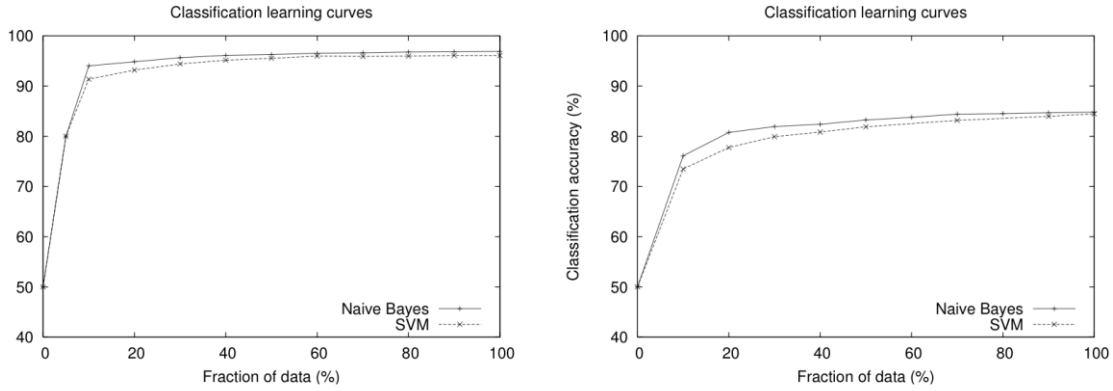
One-liners
Take my advice; I don't use it anyway.
I get enough exercise just pushing my luck.
I just got lost in thought, it was unfamiliar territory.
Beauty is in the eye of the beer holder.
I took an IQ test and the results were negative.
Reuters titles
Trocadero expects tripling of revenues.
Silver fixes at two-month high, but gold lags.
Oil prices slip as refiners shop for bargains.
Japanese prime minister arrives in Mexico.
Chains may raise prices after minimum wage hike.
Proverbs
Creativity is more important than knowledge.
Beauty is in the eye of the beholder.
I believe no tales from an enemy's tongue.
Do not look at the coat, but at what is under the coat.
A man is known by the company he keeps.

Table 13: Sample examples of one-liners, Reuters titles, and proverbs.

To test the feasibility of automatically differentiating between humorous and non-humorous texts using *content-based* features, we performed experiments where the humor-recognition task is formulated as a traditional text classification problem. We decided to use two of the most frequently employed text classifiers, Naïve Bayes (Yang & Liu, 1999; McCallum & Nigam, 1998) and Support Vector Machines (Vapnik, 1995; Joachims, 1998), selected based on their performance in previously reported work and for their diversity of learning methodologies.

The classification experiments are performed using stratified ten-fold cross validations, for accurate evaluations. The baseline for all the experiments is 50%, which represents the classification accuracy obtained if a label of "humorous" (or "non-humorous") would be assigned by default to all the examples in the data set. Table 14 shows results obtained using the two datasets, using the Naïve Bayes and SVM classifiers. Learning curves are plotted in Figure 1.

The results obtained in the automatic classification experiments reveal the fact that computational approaches represent a viable solution for the task of humor-recognition, and good performance can be achieved using classification techniques based on stylistic and content features.



(a) One-liners / Reuters (b) One-liners / Proverbs

Figure 1: Learning curves for humor-recognition using text classification techniques

Classifier	One-liners Reuters	One-liners Proverbs
Naïve Bayes	96.67%	84.81%
SVM	96.09%	84.48%

Table 14: Humor-recognition accuracy using content-based features and Naïve Bayes and SVM text classifiers.

Figure 1 shows that regardless of the type of negative data or classification methodology, there is significant learning only until about 60% of the data (i.e. about 10,000 positive examples, and the same number of negative examples). The rather steep ascent of the curve, especially in the first part of the learning, suggests that

humorous and non-humorous texts represent well distinguishable types of data. The plateau toward the end of the learning is also suggesting that more data is not likely to help improve the quality of an automatic humor-recognizer, and more sophisticated features are probably required. Linguistic theories of humor (Attardo, 1994) have suggested many *stylistic features* that characterize humorous texts. (Mihalcea & Strapparava, 2006) tried to identify a set of features that were both significant and feasible to implement using existing machine-readable resources. Specifically, they focused on alliteration, antonymy, and adult slang, which were previously suggested as potentially good indicators of humor (Ruch, 2002; Bucaria, 2004).

5.2 Exploiting Extra-linguistic Features

Extra-linguistic features refer to anything in the world outside language, but which is very relevant to the meaning and the pragmatics of an utterance. The careful use of these features can be exploited to the automatic processing of the language, improving or even making possible some tasks. This issue becomes quite important, especially if we are dealing with any form of emotion classification of language.

As an example, we can mention the CORPS corpus, (CORpus of tagged Political Speeches), a resource freely available for research purposes (Guerini, Strapparava, & Stock, 2008), which contains political speeches tagged with *audience reactions* (e.g. applause, standing-ovation, booing). The collected texts come from various Web sources (e.g. politicians' official sites, News web sites) to create a specific resource useful for the study of *persuasive language*. The corpus was built relying on the hypothesis that tags about public reaction, such as APPLAUSE, are indicators of hot-spots where persuasion attempts succeeded or, at least, a persuasive attempt had been recognized by the audience. Exploiting that corpus, (Strapparava, Guerini, & Stock, 2010) explored the possibility of classifying the transcript of

political discourses, according to their persuasive power, predicting the sentences that possibly trigger applause.

5.2.1 Music and Lyrics: a parallel corpus-based perspective

As another example of the usefulness of extra linguistic features, we can analyze the case of emotion classification of lyrics. After introducing a parallel corpus of music and lyrics, annotated with emotions at line level, we then describe some experiments on emotion classification using the music and the lyrics representations of the songs. Popular songs exert a lot of power on people, both at an individual level as well as on groups, mainly because of the message and emotions they communicate. Songs can lift our moods, make us dance, or move us to tears. Songs are able to embody deep feelings, usually through a combined effect of both the music and the lyrics. Songwriters know that music and lyrics have to be coherent, and the art of shaping words for music involves precise techniques of creative writing, using elements of grammar, phonetics, metrics, or rhyme, which make this genre a suitable candidate to be investigated by NLP techniques.

The computational treatment of music is a very active research field. The increasing availability of music in digital format (e.g., MIDI) has motivated the development of tools for music accessing, filtering, classification, and retrieval. For instance, the task of music retrieval and music recommendation has received a lot of attention from both the arts and the computer science communities (see for instance (Orio, 2006) for an introduction to this task).

There are several works on MIDI analysis. We report mainly those that are relevant for the purpose of the present work. For example (Das, Howard, & Smith, 2000) describes an analysis of predominant up-down motion types within music,

through extraction of the kinematic variables of music velocity and acceleration from MIDI data streams. (Cataltepe, Yaslan, & Sonmez, 2007) addresses music genre classification using MIDI and audio features, while (Wang et al., 2004) automatically aligns acoustic musical signals with their corresponding textual lyrics.

MIDI files are typically organized into one or more parallel "tracks" for independent recording and editing. A reliable system to identify the MIDI track containing the *melody*[5] is very relevant for music information retrieval, and there are several approaches that have been proposed to address this issue (Rizo et al., 2006; Velusamy, Thoshkahna, & Ramakrishnan, 2007).

Regarding natural language processing techniques applied to lyrics, there have been a few studies that mainly exploit the song lyrics components only, while ignoring the musical component. For instance, (Mahedero, Martinez, & Cano, 2005) deals with language identification, structure extraction, and thematic categorization for lyrics. (Yang & Lee, 2009) approach the problem of emotion identification in lyrics.

Despite the interest of the researchers in music and language, and despite the long history of the interaction between music and lyrics, there is little scholarly research that explicitly focuses on the connection between music and lyrics. Here, we focus on the connection between the musical and linguistic representations in popular songs, and their role in the expression of *affect*. Strapparava, Mihalcea, and Battocchi (2012b) introduced a corpus of songs with a strict alignment between notes and words, which can be regarded and used as a *parallel* corpus suitable for common parallel corpora techniques previously used in computational linguistics. The corpus

[5]A melody can be defined as a 'cantabile' sequence of notes, usually the sequence that a listener can remember after hearing a song.

consists of 100 popular songs, such as "*On Happy Days*" or "*All the Time in the World*," covering famous interpreters such as the Beatles or Sting. For each song, both the music (extracted from MIDI format) and the lyrics (as raw text) are included, along with an alignment between the MIDI features and the words. Moreover, because of the important role played by emotions in songs, the corpus also embeds manual annotations of six basic emotions collected via crowdsourcing, as described earlier in Section 3. Table 15 shows some statistics collected on the entire corpus.

SONGS	100
SONGS IN "MAJOR" KEY	59
SONGS IN "MINOR" KEY	41
LINES	4,976
ALIGNED SYLLABLES / NOTES	34,045

Table 15: Some statistics of the corpus

Figure 2 shows an example from the corpus, consisting of the first two lines in the Beatles' song *A hard day's night*.

```
<song filename=AHARDDAY.m2a>
<key time=0>G major</key>
<line pvers=1 raising=3 anger=1.5 disgust=0.7 sadness=2.5 surprise=0.8 >
<token time=5040 orig−note=B degree=3 duration=210>IT</token>
<token time=5050 orig−note=B degree=3 duration=210>'S </token>
<token time=5280 orig−note=C′ degree=4 duration=210>BEEN </token>
<token time=5520 orig−note=B degree=3 duration=210>A </token>
<token time=5760 orig−note=D′ degree=5 duration=810>HARD </token>
<token time=6720 orig−note=D′ degree=5 duration=570>DAY</token>
<token time=6730 orig−note=D′ degree=5 duration=570>'S </token>
<token time=7440 orig−note=D′ degree=5 duration=690>NIGHT</token>
</line>
<line pvers=2 raising=5 anger=3.5 disgust=2 sadness=1.2 surprise=0.2 >
<token time=8880 orig−note=C′ degree=4 duration=212>AND </token>
<token time=9120 orig−note=D′ degree=5 duration=210>I</token>
<token time=9130 orig−note=D′ degree=5 duration=210>'VE </token>
<token time=9360 orig−note=C′ degree=4 duration=210>BEEN </token>
<token time=9600 orig−note=D′ degree=5 duration=210>WOR</token>
<token time=9840 orig−note=F′ degree=7− duration=930>KING </token>
<token time=10800 orig−note=D′ degree=5 duration=210>LI</token>
<token time=11040 orig−note=C′ degree=4 duration=210>KE </token>
<token time=11050 orig−note=C′ degree=4 duration=210>A </token>
<token time=11280 orig−note=D′ degree=5 duration=330>D</token>
<token time=11640 orig−note=C′ degree=4 duration=90>O</token>
<token time=11760 orig−note=B degree=3 duration=330>G</token>
</line>
```

Figure 2: Two lines of a song in the corpus: *It-'s been a hard day-'s night, And I-'ve been wor-king li-ke a d-o-g*

We explicitly encode the following features. At the song level, the key of the song (e.g., G major, C minor). At the line level, we represent the *raising*, which is the musical interval (in half-steps) between the first note in the line and the most important note (i.e., the note in the line with the longest duration), as well as the manual emotion annotations. Finally, at the note level, we encode the time code of the note with respect to the beginning of the song; the note aligned with the corresponding syllable; the degree of the note with relation to the key of the song; and the duration of the note.

Mihalcea and Strapparava (2012b) described some experiments that display the usefulness of the joint music/text representation in this corpus of songs. In this chapter we outline an experiment for emotion recognition in songs, which relies on both music and text features. The authors used a corpus of 100 songs, which at this stage has full lyrics, text, and emotion annotations. Using a simple bag-of-words representation, which is fed to a machine learning classifier, they run two comparative experiments: one that uses only the lyrics and one that uses both the lyrics and the notes for a joint model of music and lyrics. The task was transformed into a binary classification task by using a threshold empirically set at 3. If the score for an emotion is below 3, it was recorded as "absent," whereas if the score is equal to or above 3, it was recorded as "present."

For the classification, Support Vector Machines (SVM) was used, binary classifiers that seek to find the hyperplane that best separates a set of positive examples from a set of negative examples, with maximum margin (Vapnik, 1995). Applications of SVM classifiers to text categorization led to some of the best results reported in the literature (Joachims, 1998).

Table 16 shows the results obtained for each of the six emotions, and for the three major settings that we consider: textual features only, musical features only, and a classifier that jointly uses the textual and the musical features. The classification accuracy for each experiment is reported as the average of the accuracies obtained during a ten-fold cross-validation on the corpus. The table also shows a baseline, computed as the average of the accuracies obtained when using the most frequent class observed on the training data for each fold.

Emotion	Baseline	Textual	Musical	Textual and Musical
ANGER	89.27%	91.14%	89.63%	92.40%
DISGUST	93.85%	94.67%	93.85%	94.77%
FEAR	93.58%	93.87%	93.58%	93.87%
JOY	50.26%	70.92%	61.95%	75.64%
SADNESS	67.40%	75.84%	70.65%	79.42%
SURPRISE	94.83%	94.83%	94.83%	94.83%
AVERAGE	81.53%	86.87%	84.08%	88.49%

Table 16: Evaluations using a coarse-grained binary classification.

As seen from the table, on average, the joint use of textual and musical features is beneficial for the classification of emotions. Perhaps not surprisingly, the effect of the classifier is stronger for those emotions that are dominant in the corpus, i.e., JOY and SADNESS (see Table 4). The improvement obtained with the classifiers is much smaller for the other emotions (or even absent, e.g., for SURPRISE), which is also explained by their high baseline of over 90%.

6. Conclusions

The field of affective NLP, and in particular the recognition of emotions in texts is a challenging topic. Nonetheless with current NLP techniques it is possible to approach the problem with interesting results, opening exciting applicative perspectives for the future.

In this chapter we presented some explorations in dealing with automatic recognition of affect in text. We start describing some available lexical resources, the problem of emotion annotations to create gold standard, and the Affective Text task at SemEval-2007. That task focused on the classification of emotions in news headlines, and was meant as an exploration of the connection between emotions and lexical semantics. Then we approach the problem of recognizing emotions in texts, presenting some state-of-the-art knowledge-based and corpus-based methods. We conclude the chapter presenting two promising lines of research in the field of affective NLP. The first one approaches the related task of humor recognition; the second proposes the exploitation of extra-linguistic features (e.g. music) for emotion detection.

Acknowledgements

Carlo Strapparava was partially supported by the PerTe project (Trento RISE).
This material is based in part upon work supported by National Science Foundation award #0917170

References

Alm, C., D. Roth, and R. Sproat (2005). “Emotions from Text: Machine learning for text-based emotion prediction.” In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Vancouver, Canada, pp. 347–354.

Aman, S. and S. Szpakowicz (2008). “Using Roget’s Thesaurus for Fine-grained Emotion Recognition.” In: Proceedings of the International Joint Conference on Natural Language Processing. Hyderabad, India.

Attardo, S. (1994). Linguistic Theory of Humor. Berlin: Mouton de Gruyter. Ax, A. F. (1953). “The Physiological Differentiation between Fear and Anger in Humans.” In: Psychosomatic Medicine 15, pp. 433–442.

Berry, M. (1992). “Large-Scale Sparse Singular Value Computations.” In: International Journal of Supercomputer Applications 6.1, pp. 13–49. Binsted, K. and G. Ritchie (1997). “Computational rules for punning riddles.” In: Humor 10.1.

BNC-Consortium (2000). British National Corpus http://www.hcu.ox.ac.uk/BNC/. Humanities Computing Unit of Oxford University.

Bradley, M. M. and P. J. Lang (1994). "Measuring emotion: The self-assessment manikin and the semantic differential." In: Journal of Behavioral Therapy and Experimental Psychiatry 25, pp. 49–59.

Bradley, M.M. and P.J. Lang (1999). Affective norms for English words (ANEW): Instruction manual and affective ratings. Tech. rep. The Center for Research in Psychophysiology, University of Florida.

Bucaria, C. (2004). "Lexical and syntactic ambiguity as a source of humor." In: Humor 17.3.

Calvo R. and M. Kim (2012) "Emotions in text: dimensional and categorical models". Computational Intelligence.

Cataltepe, Z., Y. Yaslan, and A. Sonmez (2007). "Music Genre Classification Using MIDI and Audio Features." In: Journal on Advances in Signal Processing.

Chaumartin, F.R. (2007). "UPAR7: A knowledge-based system for headline sentiment tagging." In: Proceedings of SemEval-2007. Prague, Czech Republic.

Das, M., D. Howard, and S. Smith (2000). "The kinematic analysis of motion curves through MIDI data analysis." In: Organised Sound 5.1, pp. 137– 145.

Deerwester, S. et al. (1990). "Indexing by latent semantic analysis." In: Journal of the American Society for Information Science 41.6, pp. 391–407.

Ekman, P. (1977). "Biological and cultural contributions to body and facial movement." In: Anthropology of the Body. Ed. by J. Blacking. London: Academic Press, pp. 34–84.

— (1993). "Facial expression of emotion." In: American Psychologist 48, pp. 384–392.

Esuli, A. and F. Sebastiani (2006). "SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining." In: Proceedings of the 5th Conference on Language Resources and Evaluation. Genova, IT.

Fellbaum, C. (1998). WordNet. An Electronic Lexical Database. The MIT Press.

Frijda, N. (1982). The Emotions (Studies in Emotion and Social Interaction). New York: Cambridge University Press.

Fussell, S.R. (2002). "The verbal communication of emotion." In: The Verbal communication of emotion: Interdisciplinary perspective. Ed. by S.R. Fussell. Lawrence Erlbaum Associates.

Gliozzo, A. and C. Strapparava (2005). “Domains Kernels for Text Categorization.” In: Proc. of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005). Ann Arbor.

Guerini, M., C. Strapparava, and O. Stock (2008). “CORPS: A Corpus of Tagged Political Speeches for Persuasive Communication Processing.” In: Journal of Information Technology & Politics 5.1, pp. 19–32.

Joachims, T. (1998). “Text Categorization with Support Vector Machines: learning with many relevant features.” In: Proceedings of the European Conference on Machine Learning.

Katz, P., M. Singleton, and R. Wicentowski (2007). “SWAT-MP:The SemEval-2007 Systems for Task 5 and Task 14.” In: Proceedings of SemEval-2007. Prague, Czech Republic.

Kim, S., & Calvo, R. A. (2011). Sentiment-Oriented Summarisation of Peer Reviews. In G. Biswas, S. Bull, J. Kay, & A. Mitrovic (Eds.), Artificial Intelligence in Education (pp. 491-493). Auckland, New Zealand: Springer, LNAI Vol 6738.

Kozareva, Z. et al. (2007). “UA-ZBSA: A Headline Emotion Classification through Web Information.” In: Proceedings of SemEval-2007. Prague, Czech Republic.

Lang, P. J. (1980). “Behavioral treatment and bio-behavioral assessment: Computer applications.” In: Technology in mental health care delivery systems. Ed. by J. B. Sidowski, J. H. Johnson, and T. A. Williams. Ablex Publishing, pp. 119–137.

Lewis, D. et al. (2004). “RCV1: A New Benchmark Collection for Text Categorization Research.” In: The Journal of Machine Learning Research 5, pp. 361–397.

Magnini, B. and G. Cavaglia` (2000). “Integrating Subject Field Codes into WordNet.” In: Proceedings of LREC-2000, Second International Conference on Language Resources and Evaluation. Athens, Greece, pp. 1413– 1418.

Mahedero, J., A. Martinez, and P. Cano (2005). “Natural Language Processing of Lyrics.” In: Proceedings of MM’05. Singapore.

McCallum, A. and K. Nigam (1998). “A comparison of event models for Naive Bayes text classification.” In: Proceedings of AAAI-98 Workshop on Learning for Text Categorization.

Mehrabian, A. and J. A. Russell (1974). An approach to environmental psychology. MIT Press.

Mihalcea, R. and C. Strapparava (2006). “Learning to Laugh (Automatically): Computational Models for Humor Recognition.” In: Journal of Computational Intelligence 22.2, pp. 126–142.

— (2012a). “Lyrics, Music, and Emotions.” In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Jeju Island, Korea.

— (2012b). “Lyrics, Music, and Emotions.” In: Proceedings of Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2012). Jeju, Korea.

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). Brazile.

Orio, N. (2006). “Music Retrieval: A Tutorial and Review.” In: Foundations and Trends in Information Retrieval 1.1, pp. 1–90.

Ortony, A., G. Clore, and M. Foss (1987a). “The Referential Structure of the Affective Lexicon.” In: Cognitive Science 11.3, pp. 341–364.

Ortony, A., G. L. Clore, and M. A. Foss (1987b). “The psychological foundations of the affective lexicon.” In: Journal of Personality and Social Psychology 53, pp. 751–766.

Osgood, C. E., W. H. May and M. S. Miron (1975). Cross-Cultural Universals of Affective Meaning. Urbana, University of Illinois Press.

Pang, B. and L. Lee (2004). “A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts.” In: Proceedings of the 42nd Meeting of the Association for Computational Linguistics. Barcelona, Spain.

Pennbaker, J. (2002). Emotion, Disclosure, and Health. Washington, D.C.: American Psychological Association.

Picard, R. (1997). Affective computing. Cambridge, MA, USA: MIT Press. isbn: 0-262-16170-2.

Rivera, J. de (1998). A Structural Theory of the Emotions. New York: International Universities Press.

Rizo, D. et al. (2006). “A Pattern Recognition Approach for Melody Track Selection in MIDI Files.” In: Proceedings of 7th International Symposyum on Music Information Retrieval (ISMIR-06). Victoria, Canada, pp. 61– 66.

Ruch, W. (2002). “Computers with a personality? Lessons to be learned from studies of the psychology of humor.” In: Proceedings of the The April Fools Day Workshop on Computational Humour.

Semin, G. R. and K. Fiedler (1988). “The cognitive functions of linguistic categories in describing persons: Social cognition and language.” In: Journal of Personality and Social Psychology 54.4, pp. 558–568.

Snow, R. et al. (2008). “Cheap and Fast – But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks.” In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Honolulu, Hawaii.

Stock, O. and C. Strapparava (2003). “Getting Serious about the Development of Computational Humour.” In: Proceedings of the 8th International Joint Conference on Artificial Intelligence (IJCAI-03). Acapulco, Mexico.

Stone, P. et al. (1966). The General Inquirer: A Computer Approach to Content Analysis. The MIT Press.

Strapparava, C., M. Guerini, and O. Stock (2010). “Predicting Persuasiveness in Political Discourses.” In: Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC’10). Valletta, Malta: European Language Resources Association (ELRA), pp. 1342– 1345. isbn: 2-9517408-6-7.

Strapparava, C. and R. Mihalcea (2007). “SemEval-2007 Task 14: Affective Text.” In: Proceedings of the 4th International Workshop on the Semantic Evaluations (SemEval 2007). Prague, Czech Republic.

Strapparava, C., R. Mihalcea, and A. Battocchi (2012a). “A Parallel Corpus of Music and Lyrics Annotated with Emotions.” In: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC’12). Istanbul, Turkey.

— (2012b). “A Parallel Corpus of Music and Lyrics Annotated with Emotions.” In: Proceedings of the 8th international conference on Language Resources and Evaluation (LREC-2012). Istanbul, Turkey.

Strapparava, C. and A. Valitutti (2004a). “WordNet-Affect: an affective extension of WordNet.” In: Proceedings of the 4th International Conference on Language Resources and Evaluation. Lisbon.

— (2004b).“WordNet-Affect:anAffectiveExtensionofWordNet.”In:Proc. of 4th International Conference on Language Resources and Evaluation. Lisbon.

Strapparava, C., A. Valitutti, and O. Stock (2006). “The Affective Weight of Lexicon.” In: Proceedings of the Fifth International Conference on Language Resources and Evaluation. Genoa, Italy.

Taylor, J. and L. Mazlack (2004). “Computationally Recognizing Wordplay in Jokes.” In: Proceedings of CogSci 2004. Chicago. url: http://www. cogsci.northwestern.edu/cogsci2004/.

Vapnik, V. (1995). The Nature of Statistical Learning Theory. Springer, New York. 27

Velusamy, S., B. Thoshkahna, and K. Ramakrishnan (2007). “Novel melody line identification algorithm for polyphonic MIDI music.” In: Proceedings of 13th International Multimedia Modeling Conference (MMM 2007). Singapore.

Wang, Y. et al. (2004). “LyricAlly: Automatic Synchronization of Acoustic Musical Signals and Textual Lyrics.” In: Proceedings of MM’04. New York.

Wiebe, J., T. Wilson, and C. Cardie (2005). “Annotating expressions of opinions and emotions in language.” In: Language Resources and Evaluation 39.2-3.

Yang, D. and W. Lee (2009). “Music Emotion Identification from Lyrics.” In: Proceedings of 11th IEEE Symposium on Multimedia.

Yang, Y. and X. Liu (1999). “A reexamination of text categorization methods.” In: Proceedings of the 22nd ACM SIGIR Conference on Research and Development in Information Retrieval.