Annotating and Identifying Emotions in Text

Carlo Strapparava and Rada Mihalcea

Abstract. This paper focuses on the classification of emotions and polarity in news headlines and it is meant as an exploration of the connection between emotions and lexical semantics. We first describe the construction of the data set used in evaluation exercise "Affective Text" task at SEMEVAL 2007, annotated for six basic emotions: ANGER, DISGUST, FEAR, JOY, SADNESS and SURPRISE, and for POSITIVE and NEGATIVE polarity. We also briefly describe the participating systems and their results. Second, exploiting the same data set, we propose and evaluate several knowledge-based and corpus-based methods for the automatic identification of emotions in text.

1 Introduction

Emotions have been widely studied in psychology and behavior sciences, as they are an important element of human nature. 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 facial expressions (e.g., [13]) or on the recognition of emotions through a variety of sensors (e.g., [35]).

Although only relatively little work has been carried out so far on the automatic identification of emotions in text [31, 1], the automatic detection of emotions in texts is 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.

Carlo Strapparava
FBK-IRST
e-mail: strappa@fbk.edu

Rada Mihalcea
University of North Texas
e-mail: rada@cs.unt.edu

G. Armano et al. (Eds.): Intelligent Information Access, SCI 301, pp. 21–38.
springerlink.com

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

- *Sentiment Analysis.* tracking sentiment timelines in on-line forums and news [25, 5], review classification [43, 34], mining opinions from product reviews [20], 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 [8].
- *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 [11]) is now considered a key element for their believability. Affective words selection and understanding is crucial for realizing appropriate and expressive conversations [7].

This paper describes experiments concerned with the emotion analysis of news headlines. In Section 3, we describe the construction of a data set of news titles annotated for emotions, and we propose a methodology for fine-grained and coarse-grained evaluations. This data set was proposed at the "Affective Text" SEMEVAL task. Section 4 reports briefly the descriptions of the participating systems. In Section 5, we introduce several algorithms for the automatic classification of news headlines according to a given emotion. In particular we present several algorithms, 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 emotion-annotated blogposts). Section 5.3 presents the evaluation of the algorithms and a comparison with the systems that participated in the SEMEVAL 2007 task on "Affective Text."

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 emotion-annotated materials such as blogs.

2 Background and Related Work

The characterization of emotions through linguistic analysis is a notoriously difficult task. On the one hand, emotions are not linguistic entities, and thus many of the previously proposed approaches for emotion detection were developed in a variety of other fields, including psychology, sociology, or philosophy. For instance, emotions have been studied with respect to facial expressions [13], action tendencies [17], physiological activity [4], or subjective experience [36].

On the other hand, one of the most convenient ways to access emotional content is through the use and analysis of language, and thus a number of previous efforts have been concentrated on the development of affective lexical resources.

One the first studies dealing with the referential structure of an affective lexicon is that of [29], consisting of an analysis of 500 words taken form the literature on emotions. Their goal was to develop a taxonomy of affective words, with special attention paid to the isolation of terms referring to emotions.

A well-known resource is General Inquirer [39]. The General Inquirer[1] is a mapping tool, which maps an input text file to counts in dictionary-supplied categories. 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 [38], for a total of 182 categories. Each category is a list of words and word senses. Currently, the category "negative" is the largest, with 2291 entries.

SentiWordNet[2] [14] is a lexical resource in which each synset s from WORDNET [16] is associated to three numerical scores $Obj(s)$, $Pos(s)$ and $Neg(s)$, indicating whether a synset term is objective, positive, or negative. The three scores are derived by combining the results produced by a committee of eight ternary classifiers.

The Affective Norms for English Words (ANEW) [9] provides a set of normative emotional ratings for a large number of words in the English language. This resource was built from analyses conducted on a wide variety of verbal judgments indicating the variance in emotional assessments along three major dimensions. The two main dimensions are "affective valence" (ranging from pleasant to unpleasant) and "arousal" (ranging from calm to excited). The third dimension is referred to as either "dominance" or "control."

Finally, WORDNET AFFECT[3] [41] is an extension of the WORDNET database that assesses a fine-grained emotion labeling of a subset of synsets suitable to represent affective concepts. In particular, one or more emotion labels (e.g. FEAR, JOY, LOVE) are assigned to a number of WORDNET synsets. There are also other labels for those concepts representing moods, situations eliciting emotions, or emotional responses. In this paper, we use WORDNET AFFECT in several of our experiments, as described in Section 5.

In addition to the task of emotion recognition and construction of affective lexical resources, related work was also concerned with opinion analysis and genre classification. Opinion analysis is a topic at the crossroads of text mining and computational linguistics, concerned with the identification of opinions (either positive or negative) expressed in a document [46, 44, 10, 33]. While opinion analysis deals with texts that are often affectively loaded, its focus is on subjectivity and polarity recognition, which is a coarser-grained level as compared to emotion recognition. Finally, related work in text genre classification was concerned with humor

[1] http://www.wjh.harvard.edu/~inquirer/

[2] http://sentiwordnet.isti.cnr.it

[3] WORDNET AFFECT is freely available for research purpose at http://wndomains.itc.it

recognition [27], male/female writing differences [22, 24], and happiness recognition in blogs [26].

3 Building a Data Set for Emotion Analysis

For the experiments reported in this paper we use the data set we developed for the SEMEVAL 2007 task on "Affective Text" [40].

The task was focused on the emotion classification of news headlines extracted from news web sites. 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 this type of text 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 predefined set of emotions (ANGER, DISGUST, FEAR, JOY, SADNESS, SURPRISE), classify the titles with the appropriate emotion label and/or with a valence indication (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 from participation, and in such cases the teams were allowed to create their own supervised training sets.

Participants were free to use any resources they wanted. We provided a set of words extracted from WORDNET AFFECT [41], relevant to the six emotions of interest. However, the use of this list was entirely optional.

3.1 Data Set

The data set consisted 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 consisting of 1,000 annotated headlines.[4]

3.2 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 [45, 32], which typically rely on binary $0/1$ annotations, we decided to use a finer-grained scale, hence allowing the annotators to select different degrees of emotional load.

The test data set was independently labeled by six annotators. 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 1 shows three sample headlines in our data set, along with their final gold standard annotations.

Table 1 Sample headlines and manual annotations of emotions

	EMOTIONS						
	Anger	Disgust	Fear	Joy	Sadness	Surprise	Valence
Inter Milan set Serie A win record	2	0	0	50	0	9	50
Cisco sues Apple over iPhone name	48	8	10	0	11	19	-56
Planned cesareans not risk-free, group warns	0	0	61	0	15	11	-60

3.3 Inter-annotator Agreement

We conducted inter-tagger agreement studies for each of the six emotions. The agreement evaluations were carried out using the Pearson correlation measure, and are shown in Table 2. 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.

[4] The data set and more information about the task can be found at the SEMEVAL 2007 web site `http://nlp.cs.swarthmore.edu/semeval`

Table 2 Pearson correlation for inter-annotator agreement

EMOTIONS	
ANGER	49.55
DISGUST	44.51
FEAR	63.81
JOY	59.91
SADNESS	68.19
SURPRISE	36.07
VALENCE	
Valence	78.01

3.4 Fine-Grained and Coarse-Grained Evaluations

Provided a gold-standard data set with emotion annotations, we used both fine-grained and coarse-grained evaluation metrics for the evaluation of systems for automatic emotion annotation.

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 also ran coarse-grained evaluations, where each emotion was mapped to a 0/1 classification (0 = [0,50), 1 = [50,100]). For the coarse-grained evaluations, we calculated precision, recall, and F-measure.

4 Systems and Results Obtained in the AFFECTIVE TEXT Task

Five teams have participated in the "Affective Text" task as SEMEVAL, with five systems for valence classification and three systems for emotion labeling.

4.1 Participating Systems

The following represents a short description of the systems.

UPAR7:

This is a rule-based system [12] 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 title, and tried to identify what is being said about the main subject by exploiting the dependency graph obtained from the parser.

Each word was first rated separately for each emotion (the six emotions plus COMPASSION) and for its valence. Next, the main subject rating was boosted. Contrasts and accentuations between "good" or "bad" were detected, making it possible

to identify surprising good or bad news. The system also takes into account: human will (as opposed to illness or natural disasters); negation and modals; high-tech context; celebrities.

The lexical resource used was a combination of SentiWordNet [15] and WORDNET AFFECT[41], which were semi-automatically enriched on the basis of the original trial data.

SICS:

The SICS team used a very simple approach for valence annotation based on a word-space model and a set of seed words [37]. The idea was to create two points in a high-dimensional word space - one representing positive valence, the other representing negative valence - and then projecting each headline into this space, choosing the valence whose point was closer to the headline.

The word space was produced from a lemmatized and stop list filtered version of the LA times corpus (consisting of documents from 1994, released for experimentation in the Cross Language Evaluation Forum (CLEF)) using documents as contexts and standard *tf.idf* weighting of frequencies. No dimensionality reduction was used, resulting in a 220,220-dimensional word space containing predominantly syntagmatic relations between words. Valence vectors were created in this space by summing the context vectors of a set of manually selected seed words (8 positive and 8 negative words).

For each headline in the test data, stop words and words with frequency above 10,000 in the LA times corpus were removed. The context vectors of the remaining words were then summed, and the cosine of the angles between the summed vector and each of the valence vectors were computed, and the headline was ascribed the valence value (computed as [cosine * 100 + 50]) of the closest valence vector (headlines that were closer to the negative valence vector were assigned a negative valence value). In 11 cases, a value of -0.0 was ascribed either because no words were left in the headline after frequency and stop word filtering, or because none of the remaining words occurred in the LA times corpus and thus did not have any context vector.

Table 3 System results for valence annotations

	Fine	Coarse			
	r	Acc.	Prec.	Rec.	F1
CLaC	**47.70**	**55.10**	**61.42**	9.20	16.00
UPAR7	36.96	55.00	57.54	8.78	15.24
SWAT	35.25	53.20	45.71	3.42	6.36
CLaC-NB	25.41	31.20	31.18	**66.38**	**42.43**
SICS	20.68	29.00	28.41	60.17	38.60

Table 4 System results for emotion annotations

	Fine	Coarse			
	r	Acc.	Prec.	Rec.	F1
Anger					
SWAT	24.51	92.10	12.00	5.00	7.06
UA	23.20	86.40	12.74	**21.6**	**16.03**
UPAR7	**32.33**	**93.60**	**16.67**	1.66	3.02
Disgust					
SWAT	**18.55**	97.20	0.00	0.00	-
UA	16.21	**97.30**	0.00	0.00	-
UPAR7	12.85	95.30	0.00	0.00	-
Fear					
SWAT	32.52	84.80	25.00	14.40	18.27
UA	23.15	75.30	16.23	**26.27**	**20.06**
UPAR7	**44.92**	**87.90**	**33.33**	2.54	4.72
Joy					
SWAT	**26.11**	80.60	35.41	**9.44**	**14.91**
UA	2.35	81.80	40.00	2.22	4.21
UPAR7	22.49	**82.20**	**54.54**	6.66	11.87
Sadness					
SWAT	38.98	87.70	32.50	11.92	17.44
UA	12.28	88.90	25.00	0.91	1.76
UPAR7	**40.98**	**89.00**	**48.97**	**22.02**	**30.38**
Surprise					
SWAT	11.82	**89.10**	11.86	10.93	11.78
UA	7.75	84.60	**13.70**	**16.56**	**15.00**
UPAR7	**16.71**	88.60	12.12	1.25	2.27

CLaC:

This team submitted two systems [3] to the competition: an unsupervised knowledge-based system (*CLaC*) and a supervised corpus-based system (*CLaC-NB*). Both systems were used for assigning positive/negative and neutral valence to headlines on the scale [-100,100].

CLaC:

The CLaC system relies on a knowledge-based domain-independent unsupervised approach to headline valence detection and scoring. The system uses three main kinds of knowledge: a list of sentiment-bearing words, a list of valence shifters and a set of rules that define the scope and the result of the combination of sentiment-bearing words and valence shifters. The unigrams used for sentence/headline classification were learned from WordNetdictionary entries. In order to take advantage of the special properties of WordNetglosses and relations, we developed a system that used the list of human-annotated adjectives from [19] as a seed list and

learned additional unigrams from WORDNETsynsets and glosses. The list was then expanded by adding to it all the words annotated with Positive or Negative tags in the General Inquirer. Each unigram in the resulting list had the degree of membership in the category of positive or negative sentiment assigned to it using the fuzzy net overlap score method described in the team's earlier work [2]. Only words with fuzzy membership score not equal to zero were retained in the list. The resulting list contained 10,809 sentiment-bearing words of different parts of speech.

The fuzzy net overlap score counts were complemented with the capability to discern and take into account some relevant elements of syntactic structure of the sentences. Two components were added to the system to enable this capability: (1) valence shifter handling rules and (2) parse tree analysis. The list of valence shifters was a combination of a list of common English negations and a subset of the list of automatically obtained words with increase/decrease semantics, complemented with manual annotation. The full list consists of 450 words and expressions. Each entry in the list of valence shifters has an action and scope associated with it, which are used by special handling rules that enable the system to identify such words and phrases in the text and take them into account in sentence sentiment determination. In order to correctly determine the scope of valence shifters in a sentence, the system used a parse tree analysis using MiniPar.

As a result of this processing, every headline received a system score assigned based on the combined fuzzy Net Overlap Score of its constituents. This score was then mapped into the [-100 to 100] scale as required by the task.

CLaC-NB:

In order to assess the performance of basic Machine Learning techniques on headlines, a second system ClaC-NB was also implemented. This system used a Naïve Bayes classifier in order to assign valence to headlines. It was trained on a small corpus composed of the development corpus of 250 headlines provided for this competition, plus an additional 200 headlines manually annotated and 400 positive and negative news sentences. The probabilities assigned by the classifier were mapped to the [-100, 100] scale as follows: all negative headlines received the score of -100, all positive were assigned the score of +100, and the neutral headlines obtained the score of 0.

UA:

In this system [23], in order to determine the kind and the amount of emotions in a headline, statistics were gathered from three different web Search Engines: MyWay, AlltheWeb and Yahoo. This information was used to observe the distribution of the nouns, the verbs, the adverbs and the adjectives extracted from the headline and the different emotions.

The emotion scores were obtained through 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) was divided by the number

of documents containing only an emotion and the number of documents containing all the headline words. Second, associative score between a content word and an emotion was estimated and used to weight the final PMI score. The obtained results were normalized in the range 0-100.

SWAT:

SWAT [21] is a supervised system using an unigram model trained to annotate emotional content. Synonym expansion on the emotion label words was 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.

4.2 Results

Tables 3 and 4 show the results obtained by the participating systems. The tables show both the fine-grained Pearson correlation measure and the coarse-grained accuracy, precision and recall figures.

The results indicate that the task of emotion annotation is difficult. Although the Pearson correlation for the inter-tagger agreement is not particularly high, the gap between the results obtained by the systems and the upper bound represented by the annotator agreement suggests that there is room for future improvements.

5 Automatic Emotion Analysis

In this section, we propose and evaluate several knowledge-based and corpus-based methods for the automatic identification of emotions in text, and compare the results with those obtained by the systems participating in the "Affective Text" task at SEMEVAL.

5.1 Knowledge-Based Emotion Annotation

We 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. As suggested by Ortony et al. [30], 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 such as "killer" or emotional responses such as "cry"). We call the former *direct affective words* and the latter *indirect affective words* [42].

As far as direct affective words are concerned, we follow the classification found in WORDNET AFFECT. This is an extension of the WORDNETdatabase [16], including a subset of synsets suitable to represent affective concepts. In particular, one or more affective labels (*a-labels*) are assigned to a number of WORDNETsynsets. There are also other a-labels for those concepts representing moods, situations

eliciting emotions, or emotional responses. Starting with WORDNET AFFECT, 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.

Table 5 Blogposts and mood annotations extracted from LiveJournal

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

A crucial aspect in the task of sentiment analysis is the availability of a mechanism for evaluating the semantic similarity among "generic" terms and affective lexical concepts. To this end we implemented a semantic similarity mechanism automatically acquired in an unsupervised way from a large corpus of texts (e.g., British National Corpus[5]). In particular we implemented a variation of Latent Semantic Analysis (LSA). LSA yields a vector space model that allows for a *homogeneous* representation (and hence comparison) of words, word sets, sentences and texts. For representing word sets and texts by means of an LSA vector, we used a variation of the *pseudo-document* methodology described in [6]. This variation takes into account also a *tf-idf* weighting schema (see [18] for more details). In practice, each document can be represented in the LSA space by summing up the normalized LSA vectors of all the terms contained in it. Thus a synset in WORDNET (and even all the words labeled with a particular emotion) can be represented in the LSA space, performing the pseudo-document technique on all the words contained in the synset. 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. In this paper we performed experiments with all these three representations.

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. For example in a LSA space built form the BNC, the noun "gift" is highly related to the emotional categories JOY and SURPRISE. In summary, the vectorial representation in the LSA allows us to represent, in a *uniform* way, emotional categories,

[5] BNC is a very large (over 100 million words) corpus of modern English, both spoken and written (see `http://www.hcu.ox.ac.uk/bnc/`). Other more specific corpora could also be considered, to obtain a more domain oriented similarity.

generic terms and concepts (synsets), and eventually full sentences. See [42] for more details.

5.2 Corpus-Based Emotion Annotation

In addition to the experiments based on WORDNET AFFECT, we have also conducted 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 we are using 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 5. This corpus is a subset of the corpus used in the experiments reported in [28].

Table 6 Sample blogposts labeled with moods corresponding to the six emotions

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 remember 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.

In a pre-processing step, we removed all the SGML tags and kept only the body of the blogposts, which was then passed through a tokenizer. We also kept only blogposts with a length within a range comparable to the one of the headlines,

i.e. 100-400 characters. The average length of the blogposts in the final corpus is 60 words / entry. Six sample entries are shown in Table 6.

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.

5.3 Evaluations and Results

We have implemented five different systems for emotion analysis by using the knowledge-based and corpus-based approaches described above.

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 WORDNETsynset 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.
5. NB TRAINED ON BLOGS, which is a Naive Bayes classifier trained on the blog data annotated for emotions.

The five systems were evaluated on the data set of 1,000 newspaper headlines. As mentioned earlier, we conduct both fine-grained and coarse-grained evaluations. Table 7 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 compare our results with those obtained by three systems participating in the SEMEVAL emotion annotation task: SWAT, UPAR7 and UA. Table 4 shows the results obtained by these systems on the same data set, using the same evaluation metrics.

For an overall comparison, we calculated the average over all six emotions for each system. Table 8 shows the overall results obtained by our five systems and by the three SEMEVAL systems. The best results in terms of fine-grained evaluations

Table 7 Performance of the proposed algorithms

	Fine	Coarse		
	r	Prec.	Rec.	F1
ANGER				
WN-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				
WN-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				
WN-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				
WN-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				
WN-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				
WN-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

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 WN-AFFECT PRESENCE providing the best precision, and the LSA ALL EMOTION WORDS leading to the highest recall and F-measure.

Table 8 Overall average results obtained by the five proposed systems and by the three SEMEVAL systems

	Fine	Coarse		
	r	Prec.	Rec.	F1
WN-AFFECT PRESENCE	9.54	**38.28**	1.54	4.00
LSA SINGLE WORD	12.36	9.88	66.72	16.37
LSA EMOTION SYNSET	12.50	9.20	77.71	13.38
LSA ALL EMOTION WORDS	9.06	9.77	**90.22**	**17.57**
NB TRAINED ON BLOGS	10.81	12.04	18.01	13.22
SWAT	25.41	19.46	8.61	11.57
UA	14.15	17.94	11.26	9.51
UPAR7	**28.38**	27.60	5.68	8.71

6 Conclusions

Affective computing deals with the automatic recognition and interpretation of emotions. While many studies have been carried out in the field of human-computer interaction, attempting to capture the user's physical state or behavior, only relatively little work has been carried out on the detection of emotions in texts. Written language is one of our main means of communication and, besides informative content, it also transmits attitudinal information, including emotional states. Thus, we believe that it is worthwhile to explore the task through existing state-of-the-art natural language processing techniques.

In this paper, we described the "Affective Text" task, presented at SEMEVAL-2007. The task focused on the classification of emotions in news headlines, and was meant as an exploration of the connection between emotions and lexical semantics.

After illustrating the data set, the rationale of the task, and a brief description of the participating systems, we presented several experiments in the automatic annotation of emotions in text. Through comparative evaluations of several knowledge-based and corpus-based methods carried out on the data set of 1,000 deadlines, we tried to identify the methods that work best for the annotation of emotions in text. The evaluation showed that different methods have different strengths, especially with respect to individual emotions. For instance, it seems that a machine learning classifier trained on blog data has good performance for recognizing JOY and ANGER, whereas a method based on semantic similarity is generally better for FEAR and SADNESS.

In future work, we plan to explore the lexical structure of emotions, and integrate deeper semantic processing of the text into the knowledge-based and corpus-based classification methods.

References

1. Aman, S., Szpakowicz, S.: Using roget's thesaurus for fine-grained emotion recognition. In: Proceedings of the International Joint Conference on Natural Language Processing, Hyderabad, India (2008)

2. Andreevskaia, A., Bergler, S.: Senses and sentiments: Sentiment tagging of adjectives at the meaning level. In: In: Proceedings of the 19th Canadian Conference on Artificial Intelligence, AI 2006. Quebec, Canada (2006)
3. Andreevskaia, A., Bergler, S.: CLaC and CLaC-NB: Knowledge-based and corpus-based approaches to sentiment tagging. In: Proceedings of SemEval-2007, Prague, Czech Republic (2007)
4. Ax, A.F.: The physiological differentiation between fear and anger in humans. Psychosomatic Medicine 15, 433–442 (1953)
5. Balog, K., Mishne, G., de Rijke, M.: Why are they excited? identifying and explaining spikes in blog mood levels. In: Proceedings of the 11th Meeting of the European Chapter of the Association for Computational Linguistics, EACL 2006 (2006)
6. Berry, M.: Large-scale sparse singular value computations. International Journal of Supercomputer Applications 6(1), 13–49 (1992)
7. Beskow, J., Cerrato, L., Granstrvm, B., House, D., Nordenberg, M., Nordstrand, M., Svanfeldt, G.: Expressive animated agents for affective dialogue systems. In: Proceedings of the Research Workshop on Affective Dialogue Systems, Kloster Irsee, Tyskland (2004)
8. Bozios, T., Lekakos, G., Skoularidou, V., Chorianopoulos, K.: Advance techniques for personalized advertising in a digital tv environment: the imedia system. In: eBusiness and eWork Conference, Venice, Italy, pp. 1025–1031 (2001)
9. Bradley, M., Lang, P.: Affective norms for english words (anew): Instruction manual and affective ratings. Tech. rep., The Center for Research in Psychophysiology, University of Florida (1999)
10. Breck, E., Choi, Y., Cardie, C.: Identifying expressions of opinion in context. In: Proceedings of IJCAI 2007, Hyderabad, India (2007)
11. Cassell, J.: Embodied conversational agents: Representation and intelligence in user interface. AI Magazine 22(3) (2001)
12. Chaumartin, F.: Upar7: A knowledge-based system for headline sentiment tagging. In: Proceedings of SemEval 2007, Prague, Czech Republic (2007)
13. Ekman, P.: Biological and cultural contributions to body and facial movement. In: Blacking, J. (ed.) Anthropology of the Body, pp. 34–84. Academic Press, London (1977)
14. Esuli, A., Sebastiani, F.: SentiWordNet: A publicly available lexical resource for opinion mining. In: Proceedings of the 5th Conference on Language Resources and Evaluation, Genova, IT (2006)
15. Esuli, A., Sebastiani, F.: SentiWordNet: A publicly available lexical resource for opinion mining. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006), Genoa, Italy (2006)
16. Fellbaum, C.: WordNet. An Electronic Lexical Database. The MIT Press, Cambridge (1998)
17. Frijda, N.: The Emotions (Studies in Emotion and Social Interaction). Cambridge University Press, New York (1982)
18. Gliozzo, A., Strapparava, C.: Domains kernels for text categorization. In: Proc. of the Ninth Conference on Computational Natural Language Learning (CoNLL 2005), Ann Arbor (2005)
19. Hatzivassiloglou, V., McKeown, K.: Predicting the semantic orientation of adjectives. In: Proceedings of the 35th Annual Meeting of the ACL, Madrid, Spain (1997)
20. Hu, M., Liu, B.: Mining and summarizing customer reviews. In: Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining 2004 (KDD 2004), Seattle, Washington, pp. 168–177 (2004)

21. Katz, P., Singleton, M., Wicentowski, R.: SWAT-MP: The semeval-2007 systems for task 5 and task 14. In: Proceedings of SemEval-2007, Prague, Czech Republ. (2007)
22. Koppel, M., Argamon, S., Shimoni, A.: Automatically categorizing written texts by author gender. Literary and Linguistic Computing 4(17), 401–412 (2002)
23. Kozareva, Z., Navarro, B., Vazquez, S., Montoyo, A.: UA-ZBSA: A headline emotion classification through web information. In: Proceedings of SemEval-2007, Prague, Czech Republic (2007)
24. Liu, H., Mihalcea, R.: Of men, women, and computers: Data-driven gender modeling for improved user interfaces. In: International Conference on Weblogs and Social Media (2007)
25. Lloyd, L., Kechagias, D., Skiena, S.: Lydia: A system for large-scale news analysis. In: Consens, M.P., Navarro, G. (eds.) SPIRE 2005. LNCS, vol. 3772, pp. 161–166. Springer, Heidelberg (2005)
26. Mihalcea, R., Liu, H.: A corpus-based approach to finding happiness. In: Proceedings of the AAAI Spring Symposium on Computational Approaches to Weblogs, Stanford, CA, pp. 139–144 (2006)
27. Mihalcea, R., Strapparava, C.: Learning to laugh (automatically): Computational models for humor recognition. Computational Intelligence 22(2), 126–142 (2006)
28. Mishne, G.: Experiments with mood classification in blog posts. In: Proceedings of the 1st Workshop on Stylistic Analysis Of Text For Information Access (Style 2005), Brazile (2005)
29. Ortony, A., Clore, G., Foss, M.: The referential structure of the affective lexicon. Cognitive Science 11(3), 341–364 (1987)
30. Ortony, A., Clore, G.L., Foss, M.A.: The psychological foundations of the affective lexicon. Journal of Personality and Social Psychology 53, 751–766 (1987)
31. Ovesdotter, C., Roth, D., Sproat, R.: Emotions from text: machine learning for text-based emotion prediction. In: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, pp. 579–586 (2005)
32. Pang, B., Lee, L.: 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 (2004)
33. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval 2(1-2), 1–135 (2008), http://dx.doi.org/10.1561/1500000011
34. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment classification using machine learning techniques. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2002), Philadelphia, Pennsylvania, pp. 79–86 (2002)
35. Picard, R.: Affective computing. MIT Press, Cambridge (1997)
36. de Rivera, J.: A Structural Theory of the Emotions. International Universities Press, New York (1998)
37. Sahlgren, M., Karlgren, J., Eriksson, G.: SICS: Valence annotation based on seeds in word space. In: Proceedings of SemEval-2007, Prague, Czech Republic (2007)
38. Semin, G., Fiedler, K.: The cognitive functions of linguistic categories in describing persons: Social cognition and language. Journal of Personality and Social Psychology 54, 558–568 (1988)
39. Stone, P., Dunphy, D., Smith, M., Ogilvie, D.: The General Inquirer: A Computer Approach to Content Analysis. The MIT Press, Cambridge (1966)
40. Strapparava, C., Mihalcea, R.: SemEval-2007 task 14: Affective Text. In: Proceedings of SemEval 2007, Prague, Czech Republic (2007)

41. Strapparava, C., Valitutti, A.: WordNet-Affect: an affective extension of WordNet. In: Proc. of 4th International Conference on Language Resources and Evaluation, Lisbon (2004)
42. Strapparava, C., Valitutti, A., Stock, O.: The affective weight of lexicon. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation, Genoa, Italy (2006)
43. Turney, P.: 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 2002), Philadelphia, pp. 417–424 (2002)
44. Wiebe, J., Mihalcea, R.: Word sense and subjectivity. In: Proceedings of 21st International Conference on Computational Linguistics, ACL 2006 (2006)
45. Wiebe, J., Wilson, T., Cardie, C.: Annotating expressions of opinions and emotions in language. Language Resources and Evaluation 39(2-3) (2005)
46. Wilson, T., Wiebe, J., Hwa, R.: Just how mad are you? Finding strong and weak opinion clauses. In: Proceedings of AAAI, pp. 761–769 (2004)