# SemEval-2007 Task 14: Affective Text

**Carlo Strapparava**
FBK – irst
Istituto per la Ricerca Scientifica e Tecnologica
I-38050, Povo, Trento, Italy
strappa@itc.it

**Rada Mihalcea**
Department of Computer Science
University of North Texas
Denton, TX, 76203, USA
rada@cs.unt.edu

## Abstract

The "Affective Text" task focuses on the classification of emotions and valence (i.e. positive/negative polarity) in news headlines, and is meant as an exploration of the connection between emotions and lexical semantics. In this paper, we describe the data set used in the evaluation and the results obtained by the participating systems.

## 1 Introduction

All words can potentially convey affective meaning. Every word, even those apparently neutral, can evoke pleasant or painful experiences due to their semantic relation with emotional concepts or categories. Some words have emotional meaning with respect to an individual story, while for many others the affective power is part of the collective imagination (e.g. words such as "mum", "ghost", "war").

The automatic detection of emotion 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. Possible beneficial effects of emotions on memory and attention of the users, and in general on fostering their creativity are also well-known in psychology field.

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

**Sentiment Analysis.** Text categorization according to affective relevance, opinion exploration for market analysis, etc. are just some examples of application of these techniques. While positive/negative valence checking is an active field of sentiment analysis, we believe that a fine-grained emotion checking would increase the effectiveness of the field.

**Computer Assisted Creativity.** The automated generation of evaluative expressions with a bias on some polarity orientation are a key component for automatic personalized advertisement and persuasive communication.

**Verbal Expressivity in Human Computer Interaction.** Future human-computer interaction, according to a widespread view, will emphasize naturalness and effectiveness and hence the incorporation of models of possibly many human cognitive capabilities, including affective analysis and generation. For example, emotions expression by synthetic characters (e.g. embodied conversational agents) is considered now a key element for their believability. Affective words selection and understanding is crucial for realizing appropriate and expressive conversations.

The "Affective Text" task is intended as an exploration of the connection between lexical semantics and emotions, and an evaluation of various automatic approaches to emotion recognition.

The task is not easy. Indeed, as (Ortony et al., 1987) indicates, besides words directly referring to emotional states (e.g. "fear", "cheerful") and for which an appropriate lexicon would help, there are words that act only as an indirect reference to emotions depending on the context (e.g. "monster", "ghost"). We can call the former *direct affective words* and the latter *indirect affective words* (Strapparava et al., 2006).

## 2 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 want 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 and Valitutti, 2004), relevant to the six emotions of interest. However that the use of this list of words was entirely optional.

### 2.1 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.

### 2.2 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 (Wiebe et al., 2005; Pang and 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.

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.

### 2.3 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 1. 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 1: Inter-annotator agreement

### 2.4 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.

## 3 Participating Systems

Five teams have participated in the task, with five systems for valence classification and three systems for emotion labeling. The following represents a short description of the systems.

**UPAR7:** This 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 uses a syntactic parser (the Stanford Parser (**?**)) on the modified title, and tries to identify 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 (the six emotions plus "compassion") and for valence. Next, the main subject rating is boosted. Contrasts and accentuations between "good" or "bad" things are detected, making it possible to identify surprising good ("xxx reduces risk") or bad news. The system also takes into account: human will (in contrast to illness or natural disasters); negations and modals; high-tech context; celebrities.

The lexical resource used is a combination of SentiWordNet (**?**) and WordNetAffect (Strapparava and Valitutti, 2004), 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. 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 TFIDF 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.

**CLaC:** This team submitted two systems 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 results of combination of sentiment-bearing words and valence shifters. The unigrams used for sentence/headline classification were learned from WordNet dictionary entries. In order to take advantage of the special properties of WordNet glosses and relations, we developed a system that used the list of human-annotated adjectives from (**?**) as a seed list and learned additional unigrams from WordNet synsets and glosses. The list was then expanded by adding to it all the words annotated with "Positive" or "Negative" tags in 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 our earlier work (**?**). 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 for our experiments is 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, that are used by special handling rules that enable our 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, we introduced into the system the parse tree analysis using MiniPar (**?**).

As a result of this processing, every headline received a system score assigned based on the combined fuzzy NOS of its constituents. We then mapped this score into the [-100 to 100] scale as required by the competition organizers.

**CLaC-NB:** In order to assess the performance of basic Machine Learning techniques on headlines we also implemented a second system ClaC-NB. This system uses Naïve Bayes classifier in order to assign valence to headlines. It was trained on small corpus composed of the development corpus of 250 headlines provided for this competition, 200 headlines manually annotated by us 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 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) is 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 is estimated and used to weight the final PMI score. The obtained results are normalized in the range 0-100.

**SWAT** Description not received on time; to be added later.

| | 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 2: System results for valence 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 |

Table 3: System results for emotion annotations

## 4 Results and Discussion

Tables 2 and 3 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.

While further analysis is still needed, the results indicate that the task of emotion annotations is difficult. Although the inter-annotator agreement is not particularly high either, the gap between the results obtained by the systems and the upper bound represented by the agreement between annotators suggests that there is room for future improvements.

## References


A. Ortony, G. L. Clore, and M. A. Foss. 1987. The referential structure of the affective lexicon. *Cognitive Science*, (11).

B. Pang 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, July.

C. Strapparava and A. Valitutti. 2004. Wordnet-affect: an affective extension of wordnet. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, Lisbon.

C. Strapparava, A. Valitutti, and O. Stock. 2006. The affective weight of lexicon. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*.

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