# Sentiment classification in English from sentence-level annotations of emotions regarding models of affect

*Alexandre Trilla, Francesc Alías*

GTM – Grup de Recerca en Tecnologies Mèdia
LA SALLE – UNIVERSITAT RAMON LLULL
Quatre Camins 2, 08022 Barcelona (Spain)
atrilla@salle.url.edu, falias@salle.url.edu

## Abstract

This paper presents a text classifier for automatically tagging the sentiment of input text according to the emotion that is being conveyed. This system has a pipelined framework composed of Natural Language Processing modules for feature extraction and a hard binary classifier for decision making between positive and negative categories. To do so, the Semeval 2007 dataset composed of sentences emotionally annotated is used for training purposes after being mapped into a model of affect. The resulting scheme stands a first step towards a complete emotion classifier for a future automatic expressive text-to-speech synthesizer.

**Index Terms**: natural language processing, text categorization, emotion tagging, sentiment classification

## 1. Introduction

Emotion recognition is a topic that has gained interest and popularity in time. Presently, with the increasing demand of a more natural Human-Computer Interaction (HCI), the emotional space is one of the key aspects to understand the implicit channel of communication [1], which transmits non-verbal messages along with the explicit verbal messages, say the objective information. This "reading between the lines" has traditionally been tackled by psychology, trying to build an emotional knowledge base to deal with these recognition/classification aspects. Such typology has been made possible with the adoption of emotional dimensions.

Emotional dimensions are though a simplified description of basic properties of emotional states [2]. While they do not capture all the relevant aspects of an emotional state, they provide a taxonomy allowing simple distance measures between emotion categories to be used to contrast these basic properties. This approach has historically been embraced for data-driven research activities [1] and recently it has been adopted by the W3C with the EmotionML specification [3].

In the literature, one of the most popular emotion evaluation spaces is the circumplex: a bidimensional space that represents the valence (positive/negative evaluation) and the activation (stimulation of activity) of emotions. This approach has though some slight differences according to the considerations taken by their authors. For instance: Russell's affective model [4], Scherer's model [5], Plutchik's model [6] and Whissell's dictionary of affect [7]. Some later works [8] also intend to measure a complementary emotional feature/dimension for a given environment, the control or power (dominant/submissive), in order to grasp the finest distinctions between emotions.

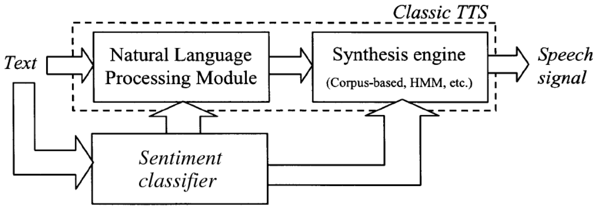

Figure 1: *Block diagram of a TTS synthesis system including a sentiment text classifier following the approach introduced in [14] for conducting multidomain TTS synthesis.*

Tagging affectively an incoming plain text is the goal aimed at many publications, from Knowledge Engineering based systems like EmoTag [9] and EmoLib [10] to data-driven approaches like [11] and [12]. The latter also introduces our final objective, the emotional/affective Text-To-Speech (TTS) synthesis, along with the works of [13] and [14] (see Figure 1).

This work is based on EmoLib [10], a library built entirely upon vocabulary expert knowledge to tag the emotion of input text. In this paper we discuss how this system can be enhanced through the knowledge acquired from more complex linguistic structures: sentence-level annotations of emotions considering models of affect. This approach may increase the effectiveness of the system compared to positive/negative valence annotation alone [15]. The resulting scheme should stand a first step towards automatic emotional sounding TTS synthesis in contrast to including explicit text tagging.

## 2. EmoLib: emotion identification from text

The original EmoLib architecture, thoroughly described in [10], is based on a set of expert decisions in order to assign emotional labels to the input text. As is shown in Figure 2, it firstly extracts the most relevant features from text aiming to spot the emotional keywords for the classification purpose and then it applies a rule-based classifier to assign the most appropriate emotional tag to the text being analyzed. The different modules that build EmoLib are described hereunder.

**Lexical analyzer:** Converts the plain input text into an output token stream. Spots the possible emotion containers (nouns, verbs, etc.) from the rest of emotionally irrelevant particles (prepositions, articles, etc.), also known as "stop words". This module is produced with a parser generator.

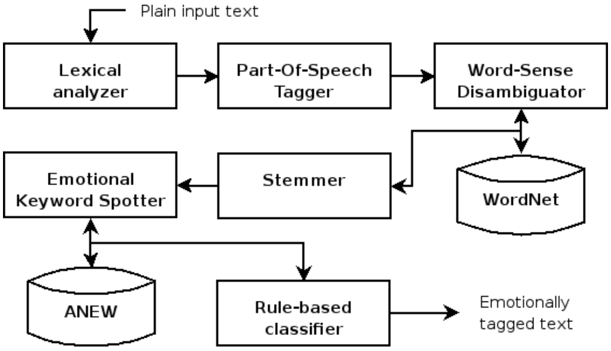

Figure 2: *EmoLib processing framework diagram.*

- **Part-Of-Speech (POS) Tagger:** Determines the function of nouns, verbs and adjectives (emotion containers) in the sentence using the Stanford log-linear POS tagger [16].
- **Word-Sense Disambiguator:** Determines the meaning of nouns according to the context. Additionally provides a set of synonyms for the resulting sense. In this module the WordNet database [17] is of use.
- **Stemmer:** Removes the inflection of words for indexing purposes using the Porter stemming algorithm [18]. Related words should map to the same stem, base or root form.
- **Emotional Keyword Spotter:** Provides the emotional dimensions to the emotional words using the ANEW dictionary of affect [8].
- **Rule-based classifier:** Computes the averaged emotional dimensions for the text of analysis and determines the correspondent emotional tag.

The ANEW dictionary contains 1035 words scored for valence, activation and control, with the Self Assessment Manikin graphical tool [8]. In this paper, we suggest that the emotional dimensions that ANEW provides may have some relation with some of the aforementioned models of affect. This statement is analyzed and determined in the following sections.

# 3. Mapping the dataset into a model of affect

In [10], the Semeval 2007 training dataset, consisting of 250 headlines, was used for evaluation purposes. These headlines were appraised in six different emotions by different evaluators, as reported by the emotion labeling task described in [15] (also see [11]). The six emotions considered were weighed according to their individual contribution to each headline. Since EmoLib considers five out of these six different emotions, like in [9], apart from the additional neutral affective state, there is a conversion process required to treat the dataset accordingly. As a first step, this issue was tackled by somewhat heuristic decisions experimentally validated [10]. However, there was room for further improvements.

As a next step, this paper proposes considering formal emotional theories to map the Semeval 2007 headlines. Since there is no unified emotional theory, three different emotion representations are considered to find the best one for mapping the Semeval 2007 dataset: Russell's model, Whissell's dictionary and Scherer's theory of affect.

Due to the availability of the dataset, both the training and test sets (1250 headlines in total) are considered. Taking each annotation of emotion (out of the six annotations for each headline) for a weighed vector, we can compute the vector sum in order to obtain the resulting projection of the headline in the given emotional space (circumplex). A similar approach was followed in [13]. Then the closest basic emotion to this resulting point is assigned to the headline. Circumplex models provide an explicit notion of the degree of similarity between emotion categories: adjacent categories in the space are very similar while opposite categories are maximally different from each other [19].

Finally, in order to score the adequacy of the affective model to classify the dataset, a 10-fold cross-validation procedure with 7-Nearest Neighbor (7-NN) should yield the effectiveness measure of the dataset w.r.t. a model of affect. After dealing with the three proposed models, the one which results in a highest effectiveness rate for mapping the dataset at hand, and thus building the ground truth, will be taken for further analysis. Notice that we compute the classification performance by means of macroaveraging [20] so as to prevent the results from being biased due to the balance of the data distribution.

## 3.1. Distribution of emotions

Russell's model of affect appears in [21] as a reference circumplex through a figure with a setting of points representing the emotions. The numerical data has been obtained from the relative position of the points in the canvas.

Whissell's model of affect, used in [13], appears in [1] contrasted with the completely different approach to emotional dimensions that Plutchik proposed [6], arranged in an "emotion wheel" instead of a circumplex. As it can be seen in the extensive table provided in the article, the emotional values have some significant differences with Russell's. These differences imply a different location of the basic emotions in the space of evaluation, which in its turn it is sensible to believe that the distance-based approach of retrieval will be more or less biased.

Finally, an adapted Scherer's model of affect appears in [22] as a reference circumplex for the binary classification experiments presented. Note that not all the basic emotions in the Semeval 2007 dataset can be directly mapped into the emotional representations proposed. In order to surpass this mismatch we make use of the existing similarity between two emotions close together in the circumplex model [19] accounting for the synonyms for each emotion given by WordNet [17].

Table 1: *Distribution of emotions in the Semeval 2007 dataset according to the considered models of affect.*

| Emotion | Russell | Whissell | Scherer |
|---|---|---|---|
| anger | 21.55 % | 14.02 % | 12.58 % |
| fear | 6.09 % | 37.42 % | 8.89 % |
| sorrow | 5.69 % | 0.32 % | 1.92 % |
| neutral | 53.93 % | 20.03 % | 54.41 % |
| happiness | 9.21 % | 25.64 % | 15.30 % |
| surprise | 3.53 % | 2.56 % | 6.89 % |

Labeling each headline in the dataset with the nearest emotion given by a determined model of affect, Table 1 can be produced showing the resulting balance of the Semeval 2007 dataset regarding each model. For Whissell's model and Scherer's model some emotions are barely represented, while for Russell's model all emotions have a reasonable amount of instances.

### 3.2. Grouped approach

Following the intention to predict the sentiment of the input text, as a first attempt we had *a priori* grouped all the anger, fear and sorrow instances into a "negative" class and all the neutral, happiness and surprise instances into the complementary "positive" class. Nevertheless, the resulting efficiency rates (around 98% for the three models) prevented us from discriminating the best model of affect. Thus, according to these results and letting the door open for emotion classification, we conduct a new approach considering each emotion separately in order to score the best model for a finer-level classification.

### 3.3. Separate approach

As it can be seen in Table 1 the representation of the Semeval 2007 dataset in the different emotional spaces shows unbalanced distributions. In the 10-fold cross-validation procedure for Whissell's dictionary and Scherer's model of affect, the 7-NN classifier is unable to predict the categories (emotions) with the lowest generality [20] (i.e. scarcely populated) due to the lack of examples. On the contrary, Russell's model performs successfully. The effectiveness rates (mean ± std) of this classifier are:

$$\hat{\pi}^M = 97.48\% \pm 1.26, \quad \hat{\rho}^M = 91.27\% \pm 5.57$$
$$\hat{F_1}^M = 94.20\% \pm 3.26$$

Hence, nine out of ten times it evaluates some emotional dimensions it successfully retrieves the correct instances from the dataset and every time it does so, almost always the model is accurate in its prediction. Russell's model of affect has resulted to be the best affective model to represent the emotions of the dataset at hand considering each emotion separately. Thus, it is chosen to label the emotions of the Semeval 2007 dataset to conduct the following experiments.

## 4. Sentiment text classifier

The next question is determining how EmoLib (which is based on the ANEW dictionary of affect) will perform in sentiment classification once the Semeval 2007 dataset is mapped into Russell's model of affect (used to build the ground truth).

The original rule-based classifier described in [10] was adapted to the features and the dataset by a heuristic procedure. Its effectiveness rates would be taken for baseline if the precision rate was computable: some emotions cannot be predicted by this classifier. This fact is reflected in [10] when the confusion matrix of the system shows 0% for the "surprise" and "sorrow" labels.

If we let EmoLib set its own predictions, which are the arithmetic mean of the emotional dimensions provided by ANEW at sentence-level, the resulting distribution in the circumplex is a mess. Note that in this environment the "control dimension" has been discarded because the circumplex only represents the valence and the activation of emotions. The sentences that pertain to the same emotional category are scattered all over the emotional plane. The protruding aspect observed is that the sentences with the lowest valence are placed on the lowest part of this dimension while the sentences with the highest valence are set in an inverse manner. We profit from this facet to build a sentiment classifier. Assuming that EmoLib may not be able to differentiate emotions within the same positive or negative gross evaluation, it could still successfully discern these two sentiment classes. If the most negatively evaluated emotions (anger, fear and sorrow) are grouped under the "negative" class ($N$) and the most positively evaluated ones (neutral, happiness and surprise) are grouped under the "positive" class ($P$), Figure 3 can be produced.

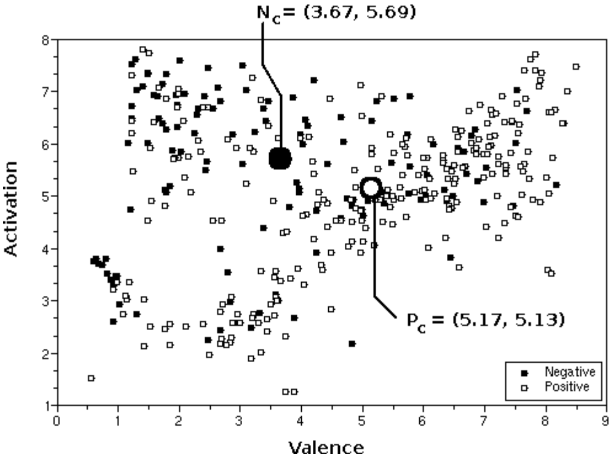

Figure 3: *Distribution of the Semeval 2007 dataset in the sentiment space.*

Considering the data distribution, we aim to build a hard binary classifier with the centroids of the two sentiments involved ($N_C$ and $P_C$). In order to test the adequacy of this approach to match the predictions given by EmoLib to the ground truth, a 10-fold cross-validation process with the hard centroid-based classifier is conducted, obtaining the following effectiveness rates (mean ± std):

$$\hat{\pi}^M = 62.11\% \pm 6.82, \quad \hat{\rho}^M = 60.42\% \pm 5.49$$
$$\hat{F_1}^M = 61.25\% \pm 6.12$$

The balance of the resulting rates denotes the consistency of the classifier. Then given an unknown input sentence we can state whether if it is positively or negatively evaluated with an accuracy of 61% approximately.

An interesting notice appointed in [15] is the belief that a fine-grained emotion annotation, in contrast to positive/negative valence annotation, would increase the effectiveness of sentiment classifiers. We have computed these rates with the valence annotation approach since the Semeval 2007 dataset also delivers these figures. Taking the negative valences for negative sentiment examples and the positive valences for positive sentiment examples, a 10-fold cross-validation procedure with a new hard centroid-based classifier yields $\hat{F_1}^M = 57.34\% \pm 5.95$, which is about 4% lower than the sentiment classifier built with fine-grained emotion annotations. Hence the impression appointed in [15] is experimentally confirmed.

## 5. Discussion and conclusions

Emotional dimensions are acknowledged to be of use in the scientific community to characterize emotions, the literature is extensive towards this belief, but when different approaches are brought together several mismatches arise. In any case, the circumplex model of affect provides a tractable typology to deal with these classification issues. In this sense, this paper presents an evolution of the EmoLib library [10] enabling it to tag the sentiment of affective texts. This development replaces the original heuristic rule-based classifier for a data-driven hard binary classifier. Its training has been accomplished by mapping the Semeval 2007 dataset, which is emotionally annotated at sentence level, into Russell's model of affect. As a result, a sentiment text classifier is obtained providing a higher classification grade compared to the previous approach. In addition, it avoids its main drawback: it could not guarantee the prediction of certain emotions. The resulting module acts like the first layer of the reported flow chart in [13]. However the latter reference, on one hand, only takes into account a particular model of affect, and, on the other hand, it makes use of isolated words for determining the emotion of the input text. In contrast, we describe a methodology for including a sentence-level emotionally annotated dataset into the classifier having evaluated the best model of affect for mapping.

EmoLib always predicts a single emotional label given an input sentence because it is aimed at emotional TTS synthesis, where each emotion corresponds to a single prosodic model. This intended goal prevents our results from being contrasted with the results obtained by the systems participating in the Semeval task for emotion recognition, despite using the same corpus, because those systems may allow multiple labels per sentence (following the evaluation rules of the Semeval 2007 task [15] [11]).

Note that the available dataset is somewhat small to draw reasonable conclusions with confidence, but since no databases with these annotations are accessible we believe it can still yield an intuitive performance score of the system. In any case, the resulting effectiveness rates, which are close to 61%, show that there's still room for improvement. For instance, the default neutral label that the system delivers in the case that no keywords are found. Although this neutral bias approach for unclear decisions is also followed in [13], we have observed that considering only the predictions bound to the emotional words spotted in the sentences (about half the dataset), and thus strictly related to the ANEW dictionary, the F-measure rate presents a slight increase $\hat{F}_1^{\,M} = 64.50\% \pm 3.97$. This rate shows that the neutral bias approach worsens the performance of the system a little. Then, in order to reduce the amount of default neutral sentences a more extensive knowledge of emotional words would definitely help. Also, the addressing of the neutral state as an independent class, instead of simple positive/negative sentiment classification, may increase the overall performance.

Finally, we believe that the emotions we perceive in text are not only restricted to lexical features, the only available features so far, but to more complex linguistic structures. Therefore, for future work we plan to delve into these linguistic topics to infer finer emotional predictions. Bearing in mind that our final goal is the generation of expressive synthetic speech, we plan to apply the resulting scheme to a complete emotion hierarchical classifier, following the work described in [13] and [14].

## 6. References

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, Jan 2001.

[2] M. Schröder, R. Cowie, E. Douglas-Cowie, M. Westerdijk, and S. Gielen, "Acoustic correlates of emotion dimensions in view of speech synthesis," in *Proc. of the 7th European Conference on Speech Communication and Technology (EUROSPEECH'01)*. Aalborg: Kommunik Grafiske Losninger A/S, 2001, pp. 87–90.

[3] P. Baggia, F. Burkhardt, J.-C. Martin, C. Pelachaud, C. Peter, B. Schuller, I. Wilson, and E. Zovato, "Elements of an emotionml 1.0," W3C, Tech. Rep., November 2008.

[4] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, pp. 1161–1178, 1980.

[5] K. R. Scherer, "Emotion as a multicomponent process: A model and some cross-cultural data," *Review of Personality and Social Psychology*, vol. 5, pp. 37–63, 1984.

[6] R. Plutchik, *Emotion: A Psychoevolutionary Synthesis*. New York, NY, USA: Harper & Row, 1980.

[7] C. M. Whissell, "The dictionary of affect in language," *Emotion: Theory, Research, and Experience*, pp. 13–131, 1989.

[8] M. M. Bradley and P. J. Lang, "Affective norms for english words (ANEW): Stimuli, instruction manual, and affective ratings," Center for Research in Psychophysiology, University of Florida, Gainesville, Florida, Tech. Rep., 1999.

[9] V. Francisco and R. Hervás, "EmoTag: Automated Mark Up of Affective Information in Texts," in *Proceedings of the Doctoral Consortium in EUROLAN 2007 Summer School*, C. Forascu, O. Postolache, G. Puscasu, and C. Vertan, Eds., Iasi, Romania, July–August 2007, pp. 5–12.

[10] D. García and F. Alías, "Emotion identification from text using semantic disambiguation," in *Procesamiento del Lenguaje Natural*, no. 40, March 2008, pp. 75–82 *(in Spanish)*.

[11] C. Strapparava and R. Mihalcea, "Learning to identify emotions in text," in *SAC'08: Proc. of the 2008 ACM symposium on Applied computing*. New York, NY, USA: ACM, 2008, pp. 1556–1560.

[12] C. O. Alm, D. Roth, and R. Sproat, "Emotions from text: machine learning for text-based emotion prediction," in *HLT '05: Proc. of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Morristown, NJ, USA: Association for Computational Linguistics, 2005, pp. 579–586.

[13] G. Hofer, K. Richmond, and R. Clark, "Informed blending of databases for emotional speech synthesis," in *Proc. Interspeech*, Sep. 2005.

[14] F. Alías, X. Sevillano, J. Socoró, and X. Gonzalvo, "Towards high-quality next-generation text-to-speech synthesis: A multidomain approach by automatic domain classification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 7, pp. 1340–1354, Sept. 2008.

[15] C. Strapparava and R. Mihalcea, "Semeval-2007 task 14: Affective text," in *Proc. of the 4th International Workshop on the Semantic Evaluations (SemEval 2007)*, Prague, Czech Republic, June 2007.

[16] K. Toutanova and C. D. Manning, "Enriching the knowledge sources used in a maximum entropy part-of-speech tagger," in *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora*. Morristown, NJ, USA: Association for Computational Linguistics, 2000, pp. 63–70.

[17] C. Fellbaum, Ed., *WordNet: An Electronic Lexical Database*, 1st ed. MIT Press, 1998.

[18] M. F. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, no. 3, pp. 130–137, 1980.

[19] M. Schröder, "Speech and emotion research: An overview of research frameworks and a dimensional approach to emotional speech synthesis," Ph.D. dissertation, Research Report of the Institute of Phonetics, Saarland University, 2004.

[20] F. Sebastiani and C. N. D. Ricerche, "Machine learning in automated text categorization," *ACM Computing Surveys*, vol. 34, pp. 1–47, 2002.

[21] J. A. Russell, M. Lewicka, and T. Niit, "A cross-cultural study of a circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 57, no. 5, 1989.

[22] M. Généreux and R. Evans, "Towards a validated model for affective classification of texts," in *Sentiment and Subjectivity in Text, Workshop at the Annual Meeting of the Association of Computational Linguistics (ACL 2006)*, Sydney, Australia, July 2006.