Combining speech-based and linguistic classifiers to recognize emotion in user spoken utterances

Accepted manuscript

David Griol, José Manuel Molina, Zoraida Callejas

Software Engineering Department, University of Granada, Spain
`{dgriol, zoraida}@ugr.es`
Universidad Carlos III de Madrid, Computer Science Dept., Avda. de la Universidad 30, 28911 Leganés, Spain
`molina@ia.uc3m.es`

Combining speech-based and linguistic classifiers to recognize emotion in user spoken utterances

David Griol[a], José Manuel Molina[a], Zoraida Callejas[b]

[a]*Applied Artificial Intelligence Group*
Dept. of Computer Science
Carlos III University of Madrid, Spain.
{david.griol,josemanuel.molina}@uc3m.es
[b]*Spoken and Multimodal Dialogue Systems Group*
Dept. of Languages and Computer Systems,
University of Granada, Spain.
zoraida@ugr.es

Abstract

In this paper we propose to combine speech-based and linguistic classification in order to obtain better emotion recognition results for user spoken utterances. Usually these approaches are considered in isolation and even developed by different communities working on emotion recognition and sentiment analysis. We propose modeling the users emotional state by means of the fusion of the outputs generated with both approaches, taking into account information that is usually neglected in the individual approaches such as the interaction context and errors, and the peculiarities of transcribed spoken utterances. The fusion approach allows to employ different recognizers and can be integrated as an additional module in the architecture of a spoken conversational agent, using the information generated as an additional input for the dialog manager to decide the next system response. We have evaluated our proposal using three emotionally-colored databases and obtained very positive results.

Keywords: Sentiment Analysis, Emotion Recognition, Paralinguistics Fusion, Affective Computing, Spoken Interaction, Context, Conversational Interfaces

1. Introduction

Emotion plays a key role in human interaction. Picard coined the term *affective computing* at a time when emotion was not considered a relevant aspect of the design of artificial systems [1]. Now there is a very active research community working on affective computing in many disciplines and tasks related to Sentiment Analysis (SA) [2, 3, 4, 5, 6, 7].

Sentiment Analysis is a research topic at the intersection of different areas, including Natural Language Processing, Computational Linguistics, Information Retrieval, and Data Mining. Usually these areas are closely related as they are based on concepts such as opinion, subjectivity or emotion. Some relevant tasks in these areas are: sentiment classification (usually classifying opinions into positive, negative or neutral) [8], subjectivity classification (detecting whether a given sentence is subjective or not) [9], opinion summarization (extracting the main features of an entity shared within one or several documents and the corresponding sentiments) [10], opinion retrieval (retrieve the opinion expressed in texts about a certain topic) [11], sarcasm and irony detection (detecting statements with ironic and sarcastic content) [12], genre or authorship detection (determining the genre or the person who has written a text or opinion) [13], and opinion spam (detecting opinions or reviews which contain untrusted contents) [14].

New applications developed in these important research areas include systems for product and movie reviews [15, 16, 17], health systems [18], disaster management [19], stock market prediction [20], business analytics [21], medical informatics [22], sentiment recognition of educative course reviews [23], aspect-based Sentiment Analysis [24], recommendation systems [25], and social network and micro-blogging processing [26, 27].

Similarly, emotion recognition is currently at the core of the most advanced conversational interfaces [28, 29, 30] to operate in scenarios that are colored with affect and provide personalized services fostering acceptance and trust. Advances in the development of these interfaces have provided an excellent opportunity to build richer user models and adapt the system's behavior accordingly [29]. Currently it is possible to obtain and manage a huge amount of information about the users, not only about what they say, but also about how they say it, where the say it and even predict why they said it and what they will say next, and these abilities will be increasingly more sophisticated in the future thanks to the multidisciplinary perspectives of the described areas.

Although sentiment/opinion detection and emotional interaction techniques for conversational systems must usually confront similar social, emotional and relational issues in order to enhance user satisfaction [31, 32, 33], as described in [34] they have rarely benefited from each other. Although emotion is receiving increasing attention from the dialog systems community, most research described in the literature is devoted exclusively to emotion recognition from paralinguistic cues. For example, comprehensive and updated reviews can be found in [35, 36]. Whereas less authors have tackled the challenge of identifying sentiment in spoken conversations by using features extracted from the text itself [37, 38]. The use of these information sources for emotion recognition still must be addressed more consistently. For example, Poria et al. [39] have very recently proposed to use both feature- and decision-level fusion methods to merge affective information extracted from multiple modalities.

In this paper, we describe a proposal that addresses these important issues. Our approach merges textual sentiment analysis and emotion recognition from paralinguistic features to respectively analyze the text transcription of the user's utterance and also consider input features extracted from the speech signal and its context.

Our main contributions are the following: Firstly, instead of focusing only on the polarity of the user's utterance (positive, negative, or neutral), our main interest is to recognize different negative emotions. These bad experiences may have a detrimental effect on the system's usability and acceptance, and may discourage users from finishing the interaction with the conversational interface or even from employing the system again. Concretely, we center on three negative emotions: doubtfulness, anger and boredom, as well as neutral. *Anger* is considered an active negative emotion, whereas *boredom* and *doubtful* passive negative emotions. We consider that the user is doubtful when he is uncertain about what to do next or what to say in that turn[1].

Secondly, the community of spoken conversational interfaces approaches emotion recognition using mainly paralinguistic information, given that speech is deeply affected by emotions (acoustic, contour, tone, voice quality, articulation changes, etc.). In our proposal, we consider features extracted from the speech signal and also features extracted from its orthographic representation. The latter account for the linguistic variability related to the emotional state of the user.

In addition, we evaluate different supervised machine learning proposals for sentiment analysis of the textual input. Unlike sentiment analysis in written text, the approach chosen for our research must provide a suitable input for the dialog management process, ideally producing a list of ranked hypotheses that can be used by a statistical dialog manager.

Besides, we consider the context of the dialog as a very important factor for emotion recognition in conversational interfaces. For the speech-based emotion recognition, we propose the inclusion of two context sources: user's neutral speaking style and dialog history. The former provides information about how users talk, which can lead to a better recognition of user's neutral emotional states. The latter involves using information about the current dialog state in terms of dialog

[1] These three emotions are part of the catalog of the HUMAINE project (http://emotion-research.net/projects/humaine. Accessed Nov 2016)

length and number of confirmations and repetitions, which gives a reliable indication of the user's emotional state at each moment (e.g., the user is likely to be angry if he has to repeated the same piece of information in numerous consecutive dialog turns). For the text-based emotion recognition, we also consider context information that mat be indicative of negative sentiments (e.g., long dialogs, customers repeating already provided information, ASR errors and misunderstoods, etc.).

The remainder of the paper is organized as follows. In Section 2 we describe the motivation of our proposal and related work. Section 3 describes our proposal, which is implemented in Section 4. This section also presents the results of the experimental set-up and results. Finally, Section 5 presents the conclusions and suggests some future work guidelines.

2. Related work

Sentiment analysis is a multi-faceted problem that is completed by means of different steps: data acquisition, data preprocessing, feature extraction, annotation, and learning [5, 40, 29].

As explained in [39], available data sets and resources for sentiment analysis are usually text-based. However, social media platforms have encouraged people to increasingly express their opinions using videos, images, and audios. Thus, it is very important to mine opinions and identify sentiments from such diverse modalities. In addition, while most emotion modeling proposals of the SA community are focused on determining the polarity of a text, the community of spoken conversational interfaces usually considers more categories of emotion or subjectivity [41].

As described in the introduction section, our proposal is focused on the integration of emotion recognition in spoken conversational interfaces. Traditionally the sentiment analysis community has not paid much attention to the internal aspects that are widespread in the conversational interfaces community [29, 38]. For this kind of interfaces, the user's spoken input is probably the most relevant source of emotional information in that it encodes the message being conveyed (the textual content) as well as how it is conveyed (paralinguistic features such as tone of voice).

On the one hand, the fact than spoken language tends to be less structured than written language, makes the task of sentiment analysis in spoken conversational interfaces especially challenging. For this reason, many of the studies in this area focus on non-textual aspects of the user's utterance related to prosodic and acoustic features of emotional speech signals [42, 43, 44]. Usually, for each of these groups, different values are computed, including statistics such as minimum, maximum, variance, mean, and median [29].

The choice of the appropriate speech signal features is still an open question. Many acoustic features can be obtained from the speech signal, although there is no single approach for classifying them. Batliner et al. [45] distinguish segmental and suprasegmental features. Segmental features are short-term spectral and derived features, including Mel-Frequency Cepstral Coefficients (MFCCs), Linear Prediction Cepstral Coefficients (LPCCs), formants, and wavelets. These features have been frequently used with other voice quality features such as Harmonics-to-Noise Ratio (HNR), jitter, or shimmer. Suprasegmental features model prosodic types such as pitch, intensity duration, and voice quality. As it is described in several related approaches, prosodic features have been found to represent the most significant characteristics of emotional content in verbal communication and were widely and successfully used for speech emotion recognition [43].

On the other hand, linguistic data acquisition and preprocessing usually requires: tokenization (break a sentence into words, phrases, symbols or other meaningful tokens by removing punctuation marks), stop word removal (remove words that do no contribute meaning or affect), stemming (bring a word into its root form), white space removal, expanding abbreviation, and feature extraction and representation.

A traditional way to perform unsupervised SA is by means of lexicon-based approaches, which rely on pre-built dictionaries of words with associated sentiment orientations [46]. The aim is to employ a sentiment lexicon composed of a collection of known and pre-compiled sentiment terms tagged with their semantic orientation to determine the overall sentiment of a given text. Lexicon-based approaches can be divided into dictionary-based techniques and corpus-based techniques, which use statistical or semantic methods to find sentiment polarity.

On the other hand, machine-learning approaches use algorithms to solve the sentiment analysis as a text classification problem. The main strength of learning-based approaches is their ability to analyze the text of any domain and produce classification models that are tailored to the problem at hand. These approaches can be classified into supervised and unsupervised learning techniques.

Supervised machine learning approaches are based on classifiers usually built from linguistic features that use a labeled training set to learn the differentiating characteristics of texts and a test set to check classifier performance. Most frequent definitions of the classification function include Naive Bayes (NB), neural networks (NN), Probabilistic Classifiers (PC), Maximum Entropy (ME), Stochastic Gradient Descent (SGD), and Support Vector Machines (SVM) [4, 17]. Hybrid approaches combine supervised and unsupervised techniques. Sentiment lexicons are usually used as unsupervised method in the majority of proposals in these approaches [2].

Our proposal is focused on coupling computational methods for emotion recognition and sentiment analysis considering linguistic, speech and conversational features. The results of the different analysis are merged by means of a fusion algorithm that provides a ranked list of emotions to be used by the dialog manager to control the interaction.

To the best of our knowledge, only the work by Park and Gates [37] has tackled the challenge of identifying sentiment in call-center conversations by using features extracted from the text itself. Text-based features included the existence of certain terms in the analyzed text as well as the identification of competitor names. The non-textual features included the analysis of the pause lengths in the conversation, the speed measured as the number of words divided by the speaking time of the customer, and the relative number of words spoken by each speaker (customer and agent) throughout the conversation.

Our proposal differs from this work in several key aspects. First, while theirs is only based on a lexicon, we evaluate different supervised machine learning approaches. In addition, some of the proposed features are domain-specific while ours are generic. We also include a richer repertoire of paralinguistic features.

The adaptation of computational methods to conversational speech also requires the integration of the dialog context, which is seldom addressed in the literature. The user's turn-taking behavior can be a powerful indicator of affect and can be integrated as an additional cue for sentiment detection (e.g., what is the user's dynamic in terms of turn-taking? does the user employ short turns? does the user try to frequently interrupt the agent?). Matt provides a state of the art of social signals in turn-taking and presents some studies analyzing the role of turn-taking as an indicator of speaker's attitude [47]. Lutfi et al. also demonstrate that conversational features can be used as a single source to reliably model user affect by predicting satisfaction ratings [48]. However, these features are not considered in the proposals for emotion recognition.

Two context sources are considered in our proposal: user's neutral speaking style and dialog history. Acoustic features are normalized around the neutral voice of the user. In addition, we consider the dialog length and width for emotion recognition, understood as the total number of turns and the number of turns required to provide a single information item (e.g., repetitions, asking for help, etc.) to improve emotion recognition accuracy.

In addition, the integration of emotion recognition in spoken conversational interfaces also requires considering automatic speech recognition (ASR) outputs and spontaneous speech features. The linguistic-based sentiment analysis carried out on automatic speech transcripts has to deal with speech variability inter-speaker and intra-speaker variability (emotion, speech style, linguistic variation, grammatical construction, badly pronounced words etc.). This variability causes two issues for oral data analysis, rarely addressed in the context of sentiment analysis.

Firstly, the performance of the ASR systems strongly depends on the background noise and quality of the recording systems. Thus, the confidence score and the various hypotheses of the ASR outputs have to be taken into account. This issue is scarcely handled in the sentiment analysis/opinion mining community [49] but is already handled by other communities. We have addressed both issues in our proposal as will be discussed in Section 3.

Secondly, even correctly transcribed, spontaneous speech features contained in the user utterance such as disfluencies, backchannels and interruptions (e.g., filled pauses, fillers, stuttering, laughter, breathing, sigh, etc.) introduce some noise into the text from the point of view of a text-based detection system.

Another important dimension that we have tackled is that the processing time of sentiment analysis is addressed in a quite different manner in the context of opinion mining. From the perspective of human-agent interaction, the question is how to provide a quick reaction to users' sentiment through affective dialog strategies. This implies reducing the time required in the prediction of the user's emotional state and to develop affective dialog strategies to be used in the dialog management task to select the best system response to continue the dialog.We have addressed this point when selecting the machine learning algorithms.

Recent work has also shown that fusion techniques to guide processes in opinion mining improve the obtained results [3, 50, 39]. Information Fusion is applied in Opinion Mining for the fusion of data sources (e.g., combine information coming from tweets and reviews from an e-commerce site) and in the fusion of resources or techniques in the Opinion Mining core process.

Different fusion strategies have been developed in recent years to perform emotion recognition using different sources and/or input modalities [43, 39]. These strategies can be classified into feature-level fusion, decision-level fusion, model-level fusion, and hybrid approaches. In feature-level fusion, the different features are concatenated to construct a joint feature vector then processed by a single classifier for emotion recognition [51].

Although this approach has been successfully used in several applications, high-dimensional feature sets may easily suffer from the problem of data sparseness, and this method does not take into account the interactions between features [43].

To avoid these disadvantages, in decision-level fusion multiple signals can be firstly modeled by the corresponding classifier and the recognition results from each classifier are fused. This method can thus combine several sources by exploring the contributions of different emotional expressions [39].

Model-level fusion strategies have been proposed to emphasize the information of correlation among multiple modalities (specially audio and visual inputs) and explore the temporal relationship between the signal streams. Finally, hybrid approaches have also been recently proposed to improve recognition results by means of the integration of different fusion approaches (e.g., feature-level and decision-level fusion strategies) [52], the use of multi-algorithm fusion techniques [53], or combining databases and fusion of classifiers [54, 55].

A number of studies favor decision-level fusion as the preferred method of data fusion in multimodal SA because errors from different classifiers tend to be uncorrelated and the methodology is feature-independent [39, 56, 57]. For this reason, in our work we analyze different decision-level fusion approaches to combine the results obtained for the speech-based and linguistic classification processes.

3. Our proposal

As introduced in the previous section, we propose to combine speech-based emotion recognition and sentiment analysis of linguistic transcripts. Our goal is to create a framework in which different algorithms can be employed and their hypothesis fused into a single recognized emotion. This way, it will be possible to use already existing approaches for emotion recognition and sentiment analysis that make the framework suitable for different application domains and emotion catalogs. In addition, it can be easily used as an extra module in the architecture of conversational systems to identify the user emotion from their utterances.

Nevertheless, we make our own proposal for the recognizers to be employed in order to cover the numerous challenges identified in Sections 1 and 2. Thus, we provide two recognizers: an emotion recognizer based on the spoken input, and a sentiment analysis approach based on its orthographic transcription. The former is based on a previously developed emotion recognizer to show how it can be easily integrated into our proposal. The latter has been created from the scratch comparing the performance of different sentiment analysis techniques and tailoring them to the task at hand. Both of them address the distinction of the emotions *angry*, *bored* and *doubtful*, as well as *neutral*. The hypotheses of both recognizers are merged in a fusion module for which we have implemented and compared different fusion techniques.

3.1. Emotion Recognition from speech

The recognition method is described in [58, 59]. It employs acoustic information to distinguish anger from doubtfulness or boredom and dialog information to discriminate between doubtfulness and boredom, which are more difficult to discriminate only by using phonetic cues.

This process is shown in Figure 1. As can be observed, the emotion recognizer always chooses one of the three negative emotions, not taking neutral into account. This is due to the difficulty of distinguishing neutral from emotional speech in spontaneous utterances when the application domain is not highly affective. This is the case of most systems, in which a baseline algorithm which always chooses "neutral" would have a very high accuracy, which is difficult to improve by classifying the rest of emotions, that are very subtlety produced.

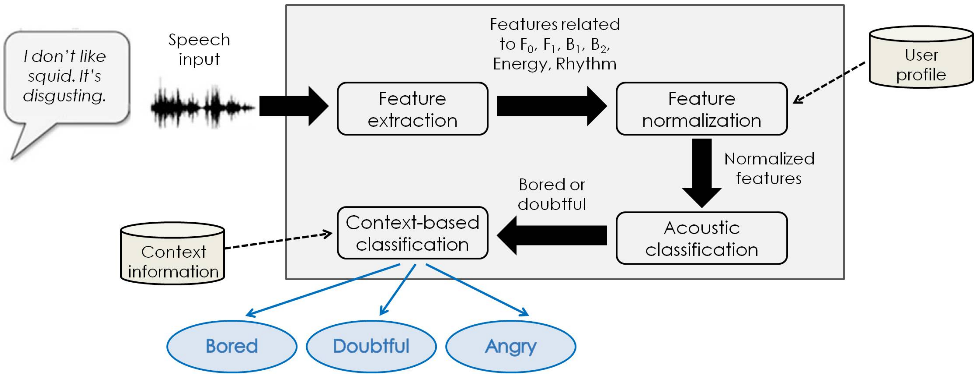
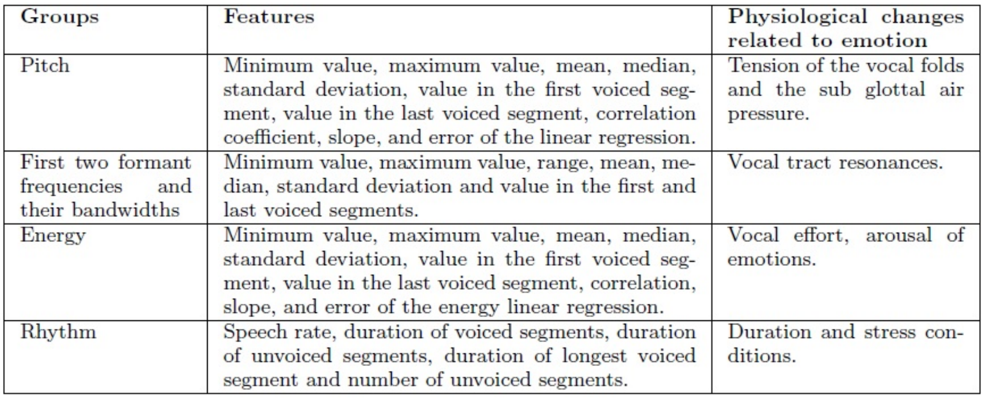


Figure 1: Schema of the speech-based emotion recognizer

The first step for emotion recognition is feature extraction. The aim is to compute features from the speech input which can be relevant for the detection of emotion in the users' voice. We extracted the most representative selection from the list of 60 features shown in Table 1. The feature selection process is carried out from a corpus of dialogs on demand, so that when new dialogs are available the selection algorithms can be executed again and the list of representative features can be updated. The features are selected by majority voting of a forward selection algorithm, a genetic search, and a ranking filter.

The second step of the emotion recognition process is feature normalization, with which the features extracted in the previous phase are normalized around the user neutral speaking style. This enables us to make more representative classifications, as it might happen that a user 'A' always speaks very fast and loudly, while a user 'B' always speaks in a very relaxed way. Then, some acoustic features may be the same for 'A' neutral as for 'B' angry, which would make the automatic classification fail for one of the users if the features are not normalized.

Once we have obtained the normalized features, we classify the corresponding utterance with a multilayer perceptron (MLP) into two categories: *angry* and *doubtful_or_bored*. The precision values obtained with the MLP are discussed in detail in [58], where we evaluated the accuracy of the initial version of this emotion recognizer. If an utterance is classified as angry, the emotional category is passed to the dialog manager of the system. If the utterance is classified as *doubtful_or_bored*,



Groups	Features	Physiological changes related to emotion
Pitch	Minimum value, maximum value, mean, median, standard deviation, value in the first voiced segment, value in the last voiced segment, correlation coefficient, slope, and error of the linear regression.	Tension of the vocal folds and the sub glottal air pressure.
First two formant frequencies and their bandwidths	Minimum value, maximum value, range, mean, median, standard deviation and value in the first and last voiced segments.	Vocal tract resonances.
Energy	Minimum value, maximum value, mean, median, standard deviation, value in the first voiced segment, value in the last voiced segment, correlation, slope, and error of the energy linear regression.	Vocal effort, arousal of emotions.
Rhythm	Speech rate, duration of voiced segments, duration of unvoiced segments, duration of longest voiced segment and number of unvoiced segments.	Duration and stress conditions.

Table 1: Features defined for emotion detection from the acoustic signal [60, 61, 62]

it is passed through an additional step in which it is classified according to two context parameters: depth and width. Depth represents the total number of sentences up to a particular point, whereas width represents the total number of extra turns needed to confirm or repeat information.

3.2. Emotion recognition from text

In conversational interfaces, the user's spoken input is translated into text by means of an automatic speech recognizer. The text is used to extract the semantics of the message conveyed and to compute the most adequate system response. However, the text also carries information about the user's emotional state. This is encoded in the words and grammatical structure.

The Apache OpenNLP[2] tools have been selected to carry out natural language processing tasks including tokenization, stemming, part-of-speech (PoS) tagging, and coreference resolution.

The NRC[3] and SenticNet[4] emotion lexicons have been used to complete a knowledge base of emotional keywords with the main information sources used to extract sentiment values from the obtained words. Both are publicly available semantic resources for concept-level Sentiment Analysis. The information in this database has been classified into the following categories:

- **Concepts**: A concept refers to the emotions associated to a specific pair

[2] https://opennlp.apache.org/

[3] http://www.saifmohammad.com/WebPages/lexicons.html

[4] http://sentic.net/

of $(word - PoS)$, where PoS denotes the grammatical function of a word inside a predicate. Only the primitive form of a word is considered and the rest of derivative words take the same set of emotional values. The different categories of words are:

  - **Nouns**: Only the singular form is considered, although they may have an irregular plural that could be harder to identify. Nouns containing prefixes and suffixes are the only exception to this rule.
  - **Adjectives**: The positive form is considered and both comparative and superlative forms are discarded.
  - **Verbs** The infinitive form is considered. Some exceptions are made for -ing forms acting as a noun (e.g., "The professor's reading about macro-economics was brilliant')'.
  - **Adverbs**: Only the positive form is considered, discarding comparative and superlative forms.

- **Modifiers**: Modifiers are denoted by an n-gram without associated sentiment states, which can increase, decrease or reverse the emotions of the associated concepts. They can be divided into two different categories:

  - **Intensity modifiers**: This category is composed by those modifiers than may increase or decrease emotions expressed by concepts (e.g., "as much" or "a bit").
  - **Negators**: These modifiers reverse the global emotion associated to a concept (e.g., "not" or "never").

Once the entities have been identified and words are annotated with values from the knowledge base, we compute the overall relevance of the entities and assign a weighting factor for each of the words carrying emotional information. A weight for each of the four independent emotional categories is then computed to classify the input text.

We have also defined a set of features to model the context of the interaction between the user and the conversational system. These features attempt to generate different aspects of the data that may be indicative of negative sentiment, including long dialogs, the customer restating the problem repeatedly, extremely long utterances, misunderstandings and errors in the ASR process. This set includes the position of the utterance in the conversation (how many utterances preceded it), the number of words in the utterance, the duration of the utterance, the number of word repetitions, the number of common words in the best two options selected by the ASR, and the values of the confidence scores for the concepts detected by the SLU module in the user's utterance.

We have studied 6 alternatives for the classification based on these cues, that are the ones that have obtained better results for the classification of multimodal sentiment analysis as discussed in Section 2. The approaches considered are: bag of words, Naive Bayes (NB), Maximum Entropy (ME), Support Vector Machines (SVM), Probabilistic Neural Networks (PNN), and Extreme Learning Machines (ELM).

- **Bag of words**. The first approach consists of a function developed for computing sentiment values represented using the *Hourglass model of Emotions* [63]. This model represents emotions using four affective dimensions (pleasantness, attention, sensitivity, and aptitude) and six activation or "sentic" levels that decompose the intensity level in the interval $[-1, +1]$ into six ranges.

  For this study we have restricted the set of emotions to the detection of the neutral state and the three negative emotions defined for the speech-based emotion recognizer. In addition, instead of using the complete set of activation levels for each affective dimension, only the most significant values are considered.

  The sentiment values assigned to each sentence depend on the weights of the different emotions (s_i) for each word in the sentence (w_i). The following equation is used for this computation:

$$S_w = \frac{\sum\limits_{i=0}^{n} w_i * s_i}{\sum\limits_{i=0}^{n} w_i}, \quad \begin{array}{c} \forall w_i > 0 \\ \forall s_i \neq 0 \\ s_i \in [-1, +1] \\ i = [0, n] \end{array} \tag{1}$$

- **The Naive Bayes** method is a probabilistic classifier method based on Bayes theorem. In this study, we propose the use of the multinomial Naive Bayes classification technique. This model considers word frequency information in document for analysis, where a document is considered to be an ordered sequence of words obtained from vocabulary 'V' [50]. The probability of a word event is independent of word context and its position in the document. Thus, each document d_i obtained from multinomial distribution of word is independent of the length of d_i. The probability of a document belonging to a class can be obtained using the following equation:

$$P(d_i|c_j;\theta) = P(|di|)|di! \prod_{t=1}^{|V|} \frac{P(w_t|c_j;\theta)^{N_{it}!}}{N_{it}!}$$

where N_{it} is the count of occurrence of w_t in document d_i; $P(d_i|c_j;\theta)$ refers to the probability of document d belonging to class c; $P(|di|)$ is the probability of document d and $P(w_t|c_j;\theta)$ is the probability of occurrence of a word w in a class c.

- In the **Maximum Entropy (ME)** method, the training data is used to define constraints, which express characteristics of training data on conditional distribution. The ME value can be expressed as

  $$P_{ME}(c|d) = \frac{1}{Z(d)} exp \sum_i \lambda_{i,c} f_{i,c}(d,c))$$

  where $P_{ME}(c|d)$ refers to probability of document d belonging to class c; $f_{i,c}(d,c)$ is the feature / class function for feature f_i and class c, $\lambda_{i,c}$ is the parameter to be estimated; and $Z(d)$ is the normalizing factor.

  The feature / class function can be instantiated as follows:

  $$f_{i,c'}(d,c) = \begin{cases} 0 & if\ c \neq c' \\ \frac{N(d,i)}{N(d)} & otherwise \end{cases}$$

  where $f_{i,c'}(d,c)$ refers to features in word-class combination in class c and document d, $N(d,i)$ represents the occurrence of feature i in document d, and $N(d)$ is the number of words in d.

- **Support Vector Machines** are a popular classifier that has proven to be efficient for various classification tasks in sentiment analysis and text classification [50, 17]. This method tries to find the optimal separating hyperplane between classes. The Sigmoid kernel function is used to implement SVM. It is given as follows:

  $$K(x_i, x_j) = tan\ h(\gamma \cdot x_i^t x_j + r)$$

  where γ and r are the kernel parameters. γ is given the value (1) and r is given the value (-100).

- **Probabilistic Neural Networks** are a versatile and efficient tool to classify high-dimensional data [50]. The probability distribution function (PDF) for a feature vector (X) to be of a certain category is given by

  $$f_a(X) = 1/(2\pi)^{(p/2)} \sigma^p (1/\eta_a) \sum_{i=1}^{\eta_a} exp(-(X - Y_{ai})^\tau (X - Y_{ai})/2\sigma^2)$$

where $f_a(X)$ is the value of the PDF for class a at point X; X is the test vector to be classified; i is the training vector number; p is the training vector size; η_a is the number of training vectors in class a; Y_{ai} is the $i-th$ training vector for class a; τ is the transpose; and σ is the standard deviation of the Gaussian curves used to construct the PDF.

Considering (n_a/n_{total}) to represent the relative number of trials in each category. Therefore, the $(1/n_a)$ term is canceled out as follows:

$$f_a(X) = 1/(2\pi)^{(p/2)}\sigma^p(1/\eta_{total})\sum_{i=1}^{\eta_a} exp(-(X - Y_{ai})^\tau(X - Y_{ai})/2\sigma^2)$$

Terms common to all classes such as $1/(2\pi)^{(p/2)}$, σ^p, and n_{total} could also be eliminated, leaving the following formula:

$$f_a(X)\alpha\sum_{i=1}^{\eta_a} exp(-(X - Y_{ai})^\tau(X - Y_{ai})/2\sigma^2)$$

For a feature parameter X to belong to a category(r); the following formula could be verified:

$$\sum_i exp(-(X - Y_{ri})^\tau(X - Y_{ri})/2\sigma^2) \geq \sum_i exp(-(X - Y_{si})^\tau(X - Y_{si})/2\sigma^2)$$

where (s) represents the other category. The expression allowing formula to be simplified as follows:

$$\sum_i exp((X^\tau Y_{ri} - 1)/\sigma^2) \geq \sum_i exp((X^\tau Y_{si} - 1)/\sigma^2)$$

- The **Extreme Learning Machine (ELM)** [64] is an emerging learning technique that provides efficient unified solutions to generalized feed-forward networks including single-/multi-hidden-layer neural networks, radial basis function networks, and kernel learning. As described in [39], ELMs offer fast learning speed, ease of implementation,and minimal human intervention.

3.3. Decision-level fusion

The main objective of decision-level fusion is to combine the separate classifiers used for the analysis of the speech signal and the transcribed text. We have evaluated three voting methods commonly used in multimodal SA [39, 56, 57].

- In the simple voting approach, the input is classified into a specific category based on the majority of individual classifier results.

- In the second approach, the output of each classifier was treated as a classification score. In particular, we obtained a probability score for each sentiment class, from each classifier. In our case, we obtained the same number of probability scores that sentiment classes from each classifier. We then calculated the final label of the classification using the following arg-max function:

  $$l' = \underset{i}{\text{argmax}}(q_1 s_i^s + q_2 s_i^t)$$

  where q_1 and q_2 represent weights for the two classifiers.

  We adopted an equal-weighted scheme, so in our case $q_1 = q_2 = 0.5$. i represents each class, and s_i^s and s_i^t denote the scores from each classifier (speech and text).

- Finally, using the Borda count [65], for every class, addition of the ranks in the n-best lists of each classifier with the first entry in the n-best list is accomplished. That means, the most likely class label, contributing the highest rank number and the last entry having the lowest rank number. Hence, the final output label for a given test pattern X is the class with highest overall rank sum. The following formula is used:

  $$r_i = \sum_{j=1}^{N} r_i^j$$

  where N is the number of classifiers (2), r_i^j is the rank of class i in the n-best list of the j-th classifier. Hence, the test pattern X is assigned the class i with the maximum overall rank count r_i.

4. Experimental results

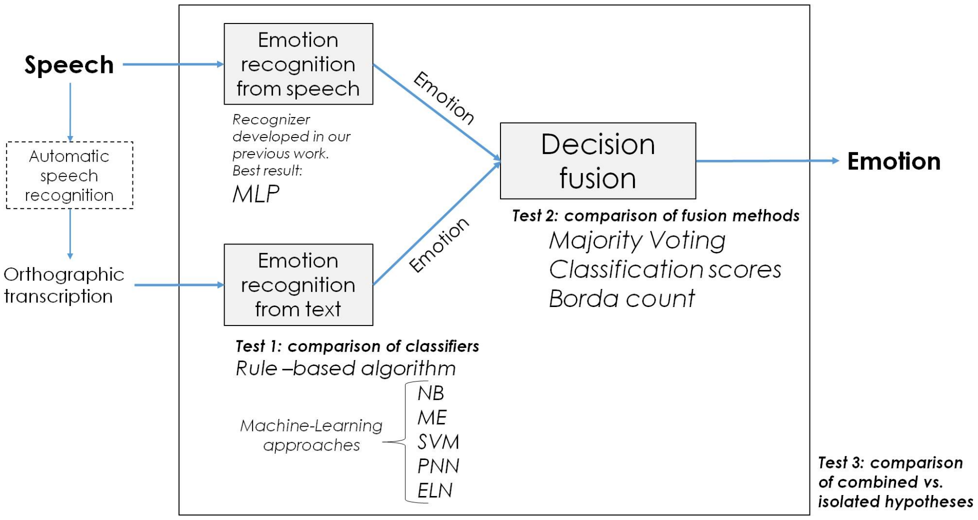
For the study, we have selected the following databases:

- **Descriptions of images**. This dataset contains 260 spoken utterances corresponding to users' descriptions of images in Spanish [66]. The corpus was acquired by 35 users. Stop words were removed and a stemmer was applied as preprocessing steps to prepare the data sets. Reviews texts sometimes contain some orthographic mistakes, abbreviations, colloquial expressions, idiomatic expressions, or ironic sentences. These bad portions of text could be filtered out (as a preprocessing step) using text summarization.

- **UAH**. Universidad Al Habla (UAH - University on the Line) is a spoken dialog system that provided academic information at the University of Granada, Spain. A corpus of 100 dialogs was acquired with this system from student telephone calls [67]. The total number of user turns was 422 and the recorded speech has a duration of 150 minutes. Nine annotators assigned an emotion category (neutral, doubtful, angry, or bored) to each user utterance twice and the final emotion for each utterance was assigned by majority voting. A detailed description of the annotation procedure and the intricacies of the calculation of inter-annotator reliability can be found in a previous study [58].

- **Let's Go task**. Let's Go is a spoken dialog system developed by the Carnegie Mellon University to provide bus schedule information in Pittsburgh at hours when the Port Authority phones are not carried out by operators. The information provided by the system covers a subset of 5 routes and 559 bus stops. This data corpus has been used as as a common testbed for several Spoken Dialog Challenges given the large large amount of data available and gathered from a real task in an operative dialog system that provided its service to real users [68]. A corpus containing 347 dialogs in American English with 9,083 system-user exchanges was annotated with quality, emotion, and task success labels by researchers of the University of Ulm [69].

Figure 2 shows the process that we have followed to complete the evaluation. As it can be observed, we have carried out three main tests, which are respectively related to the comparison of classifiers used in the emotion recognition from text (Test 1), the comparison of fusion methods to combine the classifiers used for the speech signal and the text transcription (Test 2), and the comparison of combined versus isolated hypotheses (Test 3).



Speech
Automatic speech recognition
Orthographic transcription
Emotion recognition from speech
Recognizer developed in our previous work. Best result:
MLP
Emotion recognition from text
Test 1: comparison of classifiers
Rule –based algorithm
Machine-Learning approaches
NB
ME
SVM
PNN
ELN
Emotion
Emotion
Decision fusion
Test 2: comparison of fusion methods
Majority Voting
Classification scores
Borda count
Emotion
Test 3: comparison of combined vs. isolated hypotheses

Figure 2: Experimental set-up showing the two evaluation processes

Accuracy, precision, recall and F-measure have been used as evaluation measures. Precision measures the exactness of the classifier result. Recall measures the completeness of the classifier result. F-measure is the harmonic mean of precision and recall. It is required to optimize the system towards either precision or recall. Accuracy is the most common measure of performance, and is preferred in many studies since the goal in sentiment classification is to achieve high separation between the different classes on a test set and low misclassification rates. The equations used for these performance measures are as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F - Measure = 2 * \frac{Recall * Precision}{Recall + Precision}$$

where ’TP’, ’FP’, ’TN’ and ’FN’ are true positives, false positives, true negatives and false negatives, respectively.

4.1. Test 1: Comparison of classifiers for emotion recognition from text

As described in Subsection 3.2, the bag of words and the supervised classifiers (Naive Bayes, Support Vector Machines, Maximum Entropy, Probabilistic Neural Networks, and the Extreme Learning Machine) have been employed to classify the feature vector for emotion recognition from text. 5-fold cross-validation was used for the evaluation. Each corpus was randomly split into five folds, each containing 20% of the corpus. The experiments were carried out in five trials, each using as a test set a different fold whereas the remaining folds were used as the training set. A validation subset (20%) was extracted from each training set.

Table 2 shows the results of Test 1. For the three corpora, the best accuracy was obtained using the ELM and PNN classifiers. However, we observed only a small difference in accuracy between the ELM and SVM classifiers. In terms of training time, the ELM outperformed the other classifiers by a huge margin. As our eventual goal is to develop a real-time sentiment analysis framework for spoken conversational interfaces, so the ELM provided the best performance in terms of both accuracy and training time.

Images Descriptions corpus				
Classification technique	**Precision**	**Recall**	**F-Measure**	**Accuracy**
Hourglass function	0.67	0.65	0.66	0.67
Naive Bayes (NB)	0.71	0.69	0.70	0.71
Maximum Entropy (ME)	0.75	0.73	0.74	0.74
Support Vector Machines (SVM)	0.77	0.78	0.77	0.78
Probabilistic Neural Networks (PNN)	0.78	0.77	0.78	0.79
Extreme Learning Machine (ELM)	0.79	0.81	0.79	0.79

UAH corpus				
Classification technique	**Precision**	**Recall**	**F-Measure**	**Accuracy**
Hourglass function	0.53	0.51	0.52	0.51
Naive Bayes (NB)	0.54	0.53	0.53	0.53
Maximum Entropy (ME)	0.57	0.55	0.56	0.55
Support Vector Machines (SVM)	0.59	x.xx	x.xx	x.xx
Probabilistic Neural Networks (PNN)	0.62	x.xx	x.xx	x.xx
Extreme Learning Machine (ELM)	0.64	0.61	0.62	0.63

Let's Go corpus				
Classification technique	**Precision**	**Recall**	**F-Measure**	**Accuracy**
Hourglass function	0.56	0.54	0.55	0.54
Naive Bayes (NB)	0.58	0.56	0.57	0.56
Maximum Entropy (ME)	0.63	0.61	0.62	0.62
Support Vector Machines (SVM)	0.64	0.63	0.64	0.63
Probabilistic Neural Networks (PNN)	0.66	0.64	0.65	0.64
Extreme Learning Machine (ELM)	0.67	0.66	0.67	0.67

Table 2: Results of the Test 1

We also analyzed the importance of each feature used in the classification task. The best accuracy was obtained when all features were used together. We found that concept-gram features play a major role compared to SenticNet-based features. It is also important to highlight that the features related to context awareness (e.g. duration of the dialog or ASR errors) have had a very positive impact on the results, as they are better than those obtained by other authors with the Let's Go corpus using speech features only [69].

4.2. Test 2: Comparison of fusion methods

Table 3 shows the results of the comparison of the three fusion methods described in Section 3.3. As it can be observed, the Borda count combination approach provided the best results.

Images Descriptions corpus				
Fusion method	**Precision**	**Recall**	**F-Measure**	**Accuracy**
Majority Voting	0.80	0.81	0.80	0.80
Classification scores	0.84	0.82	0.83	0.83
Borda count	0.85	0.83	0.84	0.85

UAH corpus				
Fusion method	**Precision**	**Recall**	**F-Measure**	**Accuracy**
Majority Voting	0.82	0.80	0.81	0.80
Classification scores	0.83	0.81	0.82	0.81
Borda count	0.87	0.85	0.86	0.86

Let's Go corpus				
Fusion method	**Precision**	**Recall**	**F-Measure**	**Accuracy**
Majority Voting	0.78	0.77	0.78	0.77
Classification scores	0.80	0.79	0.80	0.79
Borda count	0.82	0.81	0.81	0.82

Table 3: Results of the Test 2

4.3. Test 3: Comparison of combined vs. isolated hypotheses

Table 4 shows the experimental results obtained for each corpus if only the speech or the text classifier is used. The corpora used show a different relative importance of the text vs. the speech input for emotion recognition. On the one hand, the Images Descriptions corpus contains recordings in which the users dictate the emotion that different figures evoke. Thus, the text (the description of the figure) has more emotional content that the speech (the voice while dictating) and this is why the linguistic recognizer outperforms the speech-based recognizer in this case. On the other hand, the UAH and Let's Go corpora correspond to recordings of real user conversations with spoken dialog systems. In this case, the user speech was emotionally colored but the utterances were shorter, that is why the results obtained for the speech-based classifier are better compared to the linguistic sentiment analysis. In the three tasks the accuracy improves substantially when the two classifiers are combined.

Images Descriptions corpus				
Classifiers used	**Precision**	**Recall**	**F-Measure**	**Accuracy**
Experiment using only the speech classifier	0.67	0.68	0.68	0.67
Experiment using only the text classifier	0.79	0.81	0.79	0.79
Accuracy of decision-level fusion of the two classifiers	0.85	0.83	0.84	0.85

UAH corpus				
Classifiers used	**Precision**	**Recall**	**F-Measure**	**Accuracy**
Experiment using only the speech classifier	0.79	0.77	0.78	0.79
Experiment using only the text classifier	0.64	0.61	0.62	0.63
Accuracy of decision-level fusion of the two classifiers	0.87	0.85	0.86	0.86

Let's Go corpus				
Classifiers used	**Precision**	**Recall**	**F-Measure**	**Accuracy**
Experiment using only the speech classifier	0.76	0.75	0.76	0.76
Experiment using only the text classifier	0.67	0.66	0.67	0.67
Accuracy of decision-level fusion of the two classifiers	0.82	0.81	0.81	0.82

Table 4: Results of the Test 3

5. Conclusions and future work

Emotions are frequently mentioned in the literature as a relevant factor to select and adapt the responses of conversational systems. In this paper, we contribute a framework for recognizing the emotion conveyed in the user spoken utterances by means of a combination of Emotion Recognition and Sentiment Analysis methodologies.

We have evaluated our proposal with two recognizers: a speech based recognizer that employs acoustic and contextual features, and a linguistic recognizer that has been developed to account for the semantic and sentiment contained in the orthographic transcriptions. The results of both recognizers have been fused using different approaches that have been compared using three corpora. The results show that the combined results outperformed the individual hypotheses and provide insight on the features and classifiers that can be employed at each step, including recognition and fusion.

As future work, we would like to include our proposal as an additional module in a conversational system to assess the benefits derived from including the emotion detected as an additional parameter for dialog management.

Acknowledgements

Work partially supported by Projects MINECO TEC2012-37832-C02-01, CICYT TEC2011-28626-C02-02, CAM CONTEXTS (S2009/TIC-1485).

References

[1] R. Picard, Affective Computing, MIT Press, 2000.

[2] K. Ravi, V. Ravi, A survey on opinion mining and sentiment analysis: Tasks, approaches and applications, Knowledge-Based Systems 89 (2015) 14–46.

[3] J. A. Balazs, J. D. Velásquez, Opinion mining and information fusion: A survey, Information Fusion 27 (2016) 95–110.

[4] J. Serrano-Guerrero, J. A. Olivas, F. P. Romero, E. Herrera-Viedma, Sentiment analysis on social media for stock movement prediction, Information Sciences 311 (2015) 18–38.

[5] W. Medhat, A. Hassan, H. Korashy, Sentiment analysis algorithms and applications: A survey, Ain Shams Engineering Journal 5 (4) (2014) 1093–1113.

[6] R. Feldman, Techniques and applications for sentiment analysis, Communications of the ACM 56 (4) (2013) 82–89.

[7] B. Liu, Sentiment analysis and opinion mining. Synthesis digital library of engineering and Computer Science, Morgan & Claypool, 2012.

[8] L.-C. Yu, J.-L. Wu, P.-C. Chang, H.-S. Chu, Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news, Knowledge-Based Systems 41 (2013) 89–97.

[9] A. Montoyo, P. Martínez-Barco, A. Balahur, Subjectivity and sentiment analysis: an overview of the current state of the area and envisaged developments, Decision Support Systems 53 (4) (2012) 675–679.

[10] D. Wang, S. Zhu, T. Li, SumView: a Web-based engine for summarizing product reviews and customer opinions, Expert Systems with Applications 40 (1) (2013) 27–33.

[11] S.-W. Lee, Y.-I. Song, J.-T. Lee, K.-S. Han, H.-C. Rim, A new generative opinion retrieval model integrating multiple ranking factors, Journal of Intelligent Information Systems 38 (2) (2011) 487–505.

[12] A. Reyes, P. Rosso, D. Buscaldi, From humor recognition to irony detection: the figurative language of social media, Data Knowledge Engineering 74 (2012) 1–12.

[13] J. Savoy, Authorship attribution based on specific vocabulary, ACM Transactions on Information Systems 30 (2) (2012) 1–30.

[14] S. Xie, G. Wang, S. Lin, P. Yu, Review spam detection via temporal pattern discovery, in: Proc. of KDD'12, 2012, pp. 823–831.

[15] B. Jansen, M. Zhang, K. Sobel, A. Chowdury, Twitter power: Tweets as electronic word of mouth, Journal of the American Society for Information Science and Technology 60 (11) (2009) 2169–2188.

[16] B. Pang, L. Lee, A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts, in: Proc. of ACL'04, 2004, pp. 271–278.

[17] A. Tripathy, A. Agrawal, S. K. Rath, Classification of sentiment reviews using n-gram machine learning approach, Expert Systems with Applications 57 (2016) 117–126.

[18] M. Salathé, S. Khandelwal, Assessing vaccination sentiments with online social media: Implications for infectious disease dynamics and control, PLoS Computational Biology 7 (10) (2011) 1–7.

[19] B. Mandel, A. Culotta, J. Boulahanis, D. Stark, B. Lewis, J. Rodriguez, A demographic analysis of online sentiment during hurricane Irene, in: Proc. of LSM'12, 2012, pp. 27–36.

[20] T. Nguyen, K. Shirai, J. Velcin, Sentiment analysis on social media for stock movement prediction, Expert Systems with Applications 42 (24) (2015) 9603–9611.

[21] D. Kang, Y. Park, Review-based measurement of customer satisfaction in mobile service: sentiment analysis and VIKOR approach, Expert Systems with Applications 41 (4) (2013) 1041–1050.

[22] R. G. Rodrigues, R. M. das Dores, C. G. Camilo-Junior, T. C. Rosa, Sentihealth-cancer: A sentiment analysis tool to help detecting mood of patients in online social networks, International Journal of Medical Informatics 85 (1) (2016) 80–95.

[23] Z. Liu, S. Liu, L. Liu, J. Sun, X. Peng, T. Wang, Sentiment recognition of online course reviews using multi-swarm optimization-based selected features, Neurocomputing 185 (2016) 11–20.

[24] D. Ananda, D. Naorema, Semi-supervised Aspect Based Sentiment Analysis for Movies using Review Filtering, in: Proc. of IHCI'15, 2015, pp. 86–93.

[25] Y.-M. Li, Y.-L. Shiu, A diffusion mechanism for social advertising over microblogs, Decision Support Systems 54 (2012) 9–22.

[26] A. Balahur, J. Perea-Ortega, Sentiment analysis system adaptation for multilingual processing: The case of tweets, Information Processing & Management 51 (4) (2015) 547–556.

[27] H. Saif, Y. He, M. Fernandez, H. Alani, Contextual semantics for sentiment analysis of twitter, Information Processing & Management 52 (1) (2016) 5–19.

[28] R. Pieraccini, The Voice in the Machine: Building Computers That Understand Speech, MIT Press, 2012.

[29] M. F. McTear, Z. Callejas, D. Griol, The Conversational Interface, Springer, 2016.

[30] D. Griol, Z. Callejas, R. López-Cózar, G. Riccardi, A domain-independent statistical methodology for dialog management in spoken dialog systems, Computer, Speech and Language 28 (3) (2014) 743–768.

[31] M. Cohen, J. Giangola, J. Balogh, Voice User Interface Design, Addison-Wesley Professional, 2004.

[32] R. Harris, Voice Interaction Design: Crafting the New Conversational Speech Systems, Morgan Kaufmann, 2004.

[33] P. Kortum, HCI Beyond the GUI: Design for Haptic, Speech, Olfactory, and Other Nontraditional Interfaces, Morgan Kaufmann, 2008.

[34] C. Clavel, Z. Callejas, Sentiment Analysis: From Opinion Mining to Human-Agent Interaction, IEEE Transactions on Affective Computing 7 (1) (2016) 74–93.

[35] B. Schuller, A. Batliner, S. Steidl, D. Seppi, Recognising realistic emotions and affect in speech: state of the art and lessons learnt from the first challenge, Speech Communication 53 (9-10) (2011) 1062–1087.

[36] M. E. Ayadi, M. Kamel, F. Karray, Survey on speech emotio nrecognition: Features, classification schemes, and databases, Pattern Recognition 44 (2011) 572–587.

[37] Y. Park, S. Gates, Towards real-time measurement of customer satisfaction using automatically generated call transcripts, in: Proc. of CIKM'09, 2009, pp. 1387–1396.

[38] G. Katz, N. Ofek, B. Shapira, Consent: Context-based sentiment analysis, Knowledge-Based Systems 84 (2015) 162–178.

[39] S. Poria, E. Cambria, N. Howard, G.-B. Huang, A. Hussain, Fusing audio, visual and textual clues for sentiment analysis from multimodal content, Neurocomputing 174 (2016) 50–59.

[40] B. Pang, L. Lee, Opinion mining and sentiment analysis, Foundation and Trends in Information Retrieval 2 (1-2) (2008) 1–135.

[41] B. Schuller, A. Batliner, Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing, Wiley, 2013.

[42] C. Wu, W. Liang, Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels, IEEE Transactions on Affective Computing 2 (2011) 1–12.

[43] C.-H. Wu, J.-C. Lin, W.-L. Wei, Survey on audiovisual emotion recognition: databases, features, and data fusion strategies, APSIPA Transactions on Signal and Information Processing 3 (2014) 1–18.

[44] D. Ververidis, C. Kotropoulos, I. Pitas, Automatic emotional speech classification, in: Proc. of ICASSP'04, 2004, pp. 341–344.

[45] A. Batliner, B. Schuller, D. Seppi, S. Steidl, L. Devilliers, L. Vidrascu, T. Vogt, V. Aharonson, N. Amir, Emotion-oriented systems, Springer, 2011, Ch. The automatic recognition of emotions in speech, pp. 71–99.

[46] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, M.Stede, Lexicon-based methods for sentiment analysis, Computational Linguistics 37 (2) (2011) 267–307.

[47] M. Maat, Response Selection and Turn-Taking for a Sensitive Artificial Listening Agent, Ph.D. thesis, University of Twente (Netherlands) (2011).

[48] S. L. Lutfi, F. Fernńdez-Martínez, J. M. Lucas-Cuesta, L. López-Lebón, J. M. Montero, A satisfaction-based model for affect recognition from conversational features in spoken dialog systems, Speech Communication 55 (7-8) (2013) 825–840.

[49] E. Cambria, B. Schuller, Y. Xia, C. Havasi, New avenues in opinion mining and sentiment analysis, IEEE Intelligent Systems 28 (2) (2013) 15–21.

[50] M. A. Fattah, New term weighting schemes with combination of multiple classifiers for sentiment analysis, Neurocomputing 167 (2015) 434–442.

[51] A. Metallinou, M. Wollmer, A. Katsamanis, F. Eyben, B. Schuller, S. Narayanan, Context-sensitive learning for enhanced audiovisual emotion classification, IEEE Transactions on Affective Computing 3 (2012) 184–198.

[52] G. Chetty, M. Wagner, A Multilevel Fusion Approach for Audiovisual Emotion Recognition, in: Proc. of AVSP'08, 2008, pp. 26–29.

[53] G. Verma, U. Tiwary, S. Agrawal, Error weighted semi-coupled Hidden Markov Model for audio-visual emotion recognition, Advances in Computing and Communications 192 (2011) 452–459.

[54] I. Lefter, L. Rothkrantz, P. Wiggers, D. van Leeuwen, Emotion Recognition from Speech by Combining Databases and Fusion of Classifiers, in: Proc. of TSD'10, 2010, pp. 353–360.

[55] S. Scherer, F. Schwenker, G. Palm, Classifier fusion for emotion recognition from speech, in: Proc. of IE'07, 2007, pp. 152–55.

[56] Z. Zeng, J. Tu, M. Liu, T. Huang, B. Pianfetti, D. Roth, S. Levinson, Audio-visual affect recognition, IEEE Transactions on Multimedia 9 (2) (2007) 424–428.

[57] M.Paleari, B.Huet, Toward emotion indexing of multimedia excerpts, in: Proc. of CBMI'08, 2008, pp. 425–432.

[58] Z. Callejas, R. López-Cózar, Influence of contextual information in emotion annotation for spoken dialogue systems, Speech Communication 50 (5) (2008) 416–433.

[59] D. Griol, J. Molina, Modeling users emotional state for an enhanced human-machine interaction, in: Proc. of HAIS'15, 2015, pp. 357–368.

[60] J. Hansen, Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition, Speech Communication 20 (2) (1996) 151–170.

[61] D. Ververidis, C. Kotropoulos, Emotional speech recognition: resources, features and methods, Speech Communication 48 (2006) 1162–1181.

[62] D. Morrison, R. Wang, L. DeSilva, Ensemble methods for spoken emotion recognition in call-centres, Speech Communication 49 (2) (2007) 98–112.

[63] E. Cambria, A. Livingstone, A. Hussain, The hourglass of emotions, Cognitive Behavioural Systems: COST 2102 International Training School, Revised Selected Papers (2012) 144–157.

[64] G. Huang, G.-B.Huang, S.Song, K.You, Trends in extreme learning machines: a review, Neural Networks 61 (2015) 32–48.

[65] M. V. Erp, L. Vuurpijl, L.Schomaker, An overview and comparison of voting methods for pattern recognition, in: Proc. of IWFHR-8, 2002, pp. 195–200.

[66] D. Griol, J. Molina, A sentiment analysis classification approach to assess the emotional content of photographs, in: Proc. of ISAmI'15, 2015, pp. 105–113.

[67] Z. Callejas, D. Griol, R. López-Cózar, Predicting user mental states in spoken dialogue systems, EURASIP Journal of Advances in Signal Processing 6 (2011) 1–21.

[68] D. Griol, J. Iglesias, A. Ledezma, A. Sanchis, A Two-Stage Combining Classifier Model for the Development of Adaptive Dialog Systems, International Journal of Neural Systems 26 (1) (2016) 1–21.

[69] A. Schmitt, S. Ultes, W. Minker, A Parameterized and Annotated Spoken Dialog Corpus of the CMU Let's Go Bus Information System, in: Proc. of LREC'12, 2012, pp. 3369–3373.