Griol, D., Molina, J.M., Callejas, Z. (2019). Combining speech-based and linguistic classifiers to recognize emotion in user spoken utterances. *Neurocomputing*, 326-327, pp. 132-140.

DOI:https://doi.org/10.1016/j.neucom.2017.01.120

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 by different classifiers. This approach allows to employ already existing 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 a database with users' descriptions of images and the results show that the proposed fusion substantially outperforms the operation of the individual recognizers.

Keywords: Sentiment Analysis, Emotion Recognition, Paralinguistics Fusion, Affective Computing, Spoken Interaction, Dialog Systems

1. Introduction

Human communication involves transmitting contents, and thus has an objective component but also includes other elements which express subjective characteristics, such as opinions, attitudes, and emotions, which are the focus of many

related tasks in the field of Sentiment Analysis (SA), such as opinion mining, subjectivity analysis, or emotion detection.

Sentiment Analysis [1, 2, 3, 4, 5, 6] is a research topic at the intersection of different areas, including natural language processing, computational linguistics, information retrieval, and data mining. The growing interest in Sentiment Analysis is related to the rapid increase of available text data containing opinions and emotions, e.g. critics and recommendations on the social media, blogs, forums and other media on the web. Similarly, emotion recognition is currently at the core of the most advanced conversational interfaces [7, 8, 9] to operate in scenarios that are colored with affect and provide personalized services fostering acceptance and trust.

This way, advances in the field of conversational interfaces have provided an excellent opportunity to build richer user models and adapt the system's behavior accordingly. 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 different sciences including Computer Science, Linguistics, Psychology and Sociology.

However, as described in [10], sentiment/opinion detection and emotional interaction techniques for conversational systems have rarely benefited from each other. These systems must usually confront social, emotional and relational issues in order to enhance users satisfaction [11, 12, 13]. 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, a comprehensive and updated review can be found in [14, 15]. While few works [16, 17] have tackled the challenge of identifying sentiment in spoken conversations by using features extracted from the text itself. The fusions of these information sources for emotion recognition still must be addressed more consistently. For example, Poria et al. [18] 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 by developing affective dialog models for conversational systems. Our approach merges two classifiers developed using 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. The proposal is focused on recognizing negative emotions, so that it can be used by conversational systems to tackle situations that may have a detrimental effect on the system's usability and acceptance.

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

As it has been described previously, Sentiment Analysis offers many opportunities to develop new applications due to the huge growth of available information on the Internet [3]. Some applications are product and movie reviews [19, 20, 21], health systems [22], disaster management [23], stock markets [24], medical informatics [25], sentiment recognition of educative course reviews [26], Aspect based Sentiment Analysis (ABSA) [27], and social network and microblogging sites [28, 29].

Many disciplines and tasks have arisen linked to SA. Usually they are closely related as they are based on concepts such as opinion, subjectivity or emotion [10]. Some relevant areas are: sentiment classification (classify opinions into positive, negative or neutral) [30], subjectivity classification (detect whether a given sentence is subjective or not) [31], opinion summarization (extract the main features of an entity shared within one or several documents and the sentiments regarding them) [32], opinion retrieval (retrieve documents which express an opinion about a given query) [33], sarcasm and irony detection (detect statements which contain ironic and sarcastic content) [34], genre or authorship detection (determine the genre or the person who has written a text or opinion) [35], opinion spam (detect opinions or reviews which contain untrusted contents) [36].

Three main classification levels have been defined for SA: document-level, sentence-level, and aspect-level SA. Document-level SA aims to classify an opinion document as expressing a positive or negative opinion or sentiment [21, 37], so it considers the whole document a basic information unit. Sentence-level SA aims to classify sentiment expressed in each sentence [38], while Entity/Aspect-level SA classifies sentiment with respect to the specific aspects of relevant entities [39].

Sentiment analysis is a multi-faceted problem that is completed by means of different steps. Data acquisition and data preprocessing are the most common initial steps and usually require: 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), part of speech (POS) tagging (recognize different parts of speech in the text), and feature extraction and representation. Aside from feature extraction, feature selection is also critical to the success of the analysis. After

preprocessing, the identification and classification steps are performed [4, 40].

Firstly, identifying subjective features usually requires semantic thesauri created by human annotators that compile sentiment terms, phrases and even idioms (e.g., WordNet and SentiWordNet, MPQA subjectivity lexicon, Opinion Finder lexicon, General Inquirer). They are used to expand the polarity lexicon from a small set of seed words with known polarity that are usually collected and annotated in a manual way. In the proposals presented in [41, 42], two positive and negative verb and adjective seed lists were bootstrapped using WordNet. In the proposal described in [43], a lexical network was built by linking synonyms provided by the thesaurus. An inverse and bidirectional model of random walking algorithm was proposed in [44]. The main drawback of this kind of approaches is the incapability to deal with domain and context specific orientations, although it can be an interesting solution depending on the application domain [45].

Domain corpus-based techniques have the main objective of providing dictionaries related to a specific domain. These dictionaries are generated from a set of seed opinion terms that is expanded by means of the search of related words using either statistical or semantic techniques (e.g., Latent Semantic Analysis (LSA), frequency of occurrence of the words within a collection of documents, etc.). In [46], the frequencies of words are considered to set their polarity according to the word occurred more frequently among positive or negative texts. The use of dependency rules between opinion words and opinionated targets is proposed in [47]. The use of label propagation graphs is proposed in [48, 49].

More recently, many studies have also tried to exploit prior sentiment knowledge in source domains to assist sentiment lexicon construction in a different target domain. Several techniques have been proposed: random walking processes [50], word alignment models [51], prior web knowledge [52], co-occurrence patterns of words in different contexts [29], two phases models to detect key terms and analyze the context in which they appear [17], segments of words and their dependency relation pairs [53].

Secondly, sentiment classification techniques follow a machine-learning approach, lexicon-based approach, or hybrid approach [54]. Machine-learning approaches apply these kinds of algorithms and uses linguistic features. Lexicon-based approaches rely on a sentiment lexicon, a collection of known and precompiled sentiment terms. It is divided into dictionary-based approach and corpus-based approach which use statistical or semantic methods to find sentiment polarity. Hybrid approaches combine both approaches and is very common with sentiment lexicons playing a key role in the majority of methods.

Supervised machine learning approaches are based on classifiers built from linguistic features that use two sets of documents: a labeled training set to learn the differentiating characteristics of texts and a test set to check classifier performance.

The text is normally represented by a bag-of-words (BOW) mode [41, 55], mapped into a feature vector, which disrupts word order, breaks the syntactic structures and discards some semantic information of the text.

Some of the most important features are (1) terms (words or n-grams) and their frequency; (2) Part-Of-Speech (POS) information, adjectives play an important role but nouns can be significant; (3) negations can change the meaning of any sentence [56]; and (4) syntactic dependencies (tree parsing) can determine the meaning of sentence; among others [57, 29].

Cui et al. [58] have very recently proposed the use of distributed semantic features of word sequence as a solution to the insensitiveness of n-gram features to the order of the n-gram. The proposed features are able to automatically capture local and global contexts automatically.

The most common techniques are Naive Bayes (NB), neural networks (NN), Maximum Entropy (ME), Stochastic Gradient Descent (SGD), and Support Vector Machine (SVM) [3, 21]. The accuracy of different methods is usually examined in order to access their performance on the basis of parameters such as precision, recall, f-measure, and accuracy. Vinodhini and Chandrasekaran [59] have very recently compared neural network based sentiment classification methods (back propagation neural network (BPN), probabilistic neural network (PNN) and homogeneous ensemble of PNN (HEN)) using varying levels of word granularity as features for feature level sentiment classification.

The major disadvantages of this kind of approaches are that training data is difficult to obtain, the automatic labeling of training data introduces errors that may affect the performance of the classifiers, and domain dependence (classifiers trained on data from one domain produce unsatisfactory performance when applied to data from a different domain)

Hajmohammadi et al. [60] have recently proposed the combination of active learning and semi-supervised self-training to reduce the human labeling effort and increase the classification performance in cross-lingual sentiment classification. [61] have also very recently proposed the use of both labeled and unlabeled data in the training process for building classification models.

Khan et al. [62] have also very recently proposed the Semi-supervised feature Weighting and Intelligent Model Selection (SWIMS) to cope with the problem of unavailability of labeled data for corpus-based sentiment analysis and improve the performance of sentiment orientation detection using the SentiWord-Net lexicon.

Semi-supervised and unsupervised techniques are proposed when it is not possible to have an initial set of labeled documents/opinions to classify the rest of items. A traditional way to perform unsupervised SA is that of lexicon-based approaches, which rely on sentiment lexicons (i.e.pre-built dictionaries of words with associated sentiment orientations) [63]. 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.

Kamps et al. [43] used the WordNet database to estimate the minimum path distance between a word and pivot words. different proposals apply polarity scores directly, aggregating them from a sentence or a document and computing the resulting sentiment on a continuous scale [64, 65, 66]. Li et al. [67] proposed a constrained non-negative matrix trifactorization approach to SA, with a domain-independent sentiment lexicon as prior knowledge. More sophisticated methods employ strategies involving lexis, syntax and semantics and then aggregate their values [68, 69].

As described in [18], available data sets and resources for sentiment analysis are restricted to text-based sentiment analysis only. However, the advent of social media platforms have made people to increasingly express their opinions using videos, images, and audios. Thus, it is very important to mine opinions and identify sentiments from the diverse modalities. While most of the different emotion modeling proposals of the SA community are focused on determining the polarity of a text (positive, negative, or neutral), the community of spoken conversational interfaces usually considers more categories of emotion or subjectivity [70].

For conversational 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).

Numerous features such as prosodic and acoustic features of emotional speech signals have been discussed over the years [71, 72]. Usually, for each of these groups, different features are computed, including statistics such as minimum, maximum, variance, mean, and median [8]. Many acoustic features can be obtained from the speech signal, although there is no single approach for classifying them. Batliner et al. [73] 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. Features can be represented as raw data or they can be normalized, standardized, and presented as statistics (means, averages, etc.). 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 [72]. Morrison et al. [74] summarized the main correlations between prosodic features and emotions.

Speech features can be classified into two major categories including local (frame-level) and global (utterance-level) features [72]. Local features (e.g. MFCCs and Mel Filter Bank) represent the speech features extracted based on the unit of speech "frame". Global features (e.g., linear predictive coefficients) are calculated from the statistics of all speech features extracted from the entire "utterance".

Different open source software toolkits are available to extract these features, provide the algorithms to perform acoustic analysis and visualization, an also scripting and classification tasks. For instance, the openSMILE tool[1] was used in the 2009-2013 INTERSPEECH challenges to extract a standard range of commonly used features in audio signal analysis and emotion recognition; the Praat phonetics software[2] implements algorithms to perform the main phonetic measurement and analysis procedures, including working with waveforms and spectrograms, measuring pitch, pulses, harmonics, formants, intensity, and sound quality parameters; and the Open Social Signal Interpretation framework (OpenSSI)[3] offers tools to record, analyze, and recognize human behavior in real time, including gestures, mimics, head nods, and emotional speech.

The open source software OpenEAR[4] and the Jaudio tool have been recently employed in [18] to compute a total of 6373 audio features for emotion recognition. These features include the max and min values, standard deviation, and variance of the Mel frequency cepstral coefficients calculated based on the short time Fourier transform (STFT), the spectral centroid (center of gravity of the magnitude spectrum of the STFT), spectral flux (squared difference between the normalized magnitudes of successive windows), beat histogram (i.e., auto-correlation of the RMS), beat sum (sum of all entries in the beat histogram), strongest beat (strongest bin in the beat histogram), pause duration (percentage of time the speaker is silent in the audio segment), pitch, voice equality (harmonics to noise ratio in the audio signal), and Perceptual Linear Predictive Coefficients of the audio segment.

Pattern recognition methods such as Hidden Markov Model (HMM), Gaussian Mixture Model (GMM), support vector machine (SVM), etc. have been traditionally used to process these features and to decide the underlying emotion of the speech utterance.

Different fusion strategies have been developed in recent years to perform emotion recognition using different sources and/or input modalities [72, 18]. 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

[1]http://www.audeering.com/research/opensmile. Accessed May 2016.
[2]http://www.fon.hum.uva.nl/praat/. Accessed May 2016.
[3]http://hcm-lab.de/projects/ssi/. Accessed May 2016.
[4]https://sourceforge.net/projects/openart/. Accessed May 2016.

are concatenated to construct a joint feature vector then processed by a single classifier for emotion recognition [75].

Although fusion at feature level using simple concatenation of the audiovisual features 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 [72].

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 [18].

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) [76], the use of multi-algorithm fusion techniques [77], or combining databases and fusion of classifiers [78, 79].

3. Our proposal

We propose to combine two emotion recognizers that process the users' spoken input using relevant paralinguistic and textual features. Our goal is to create a framework in which different algorithms can be employed and their hypothesis are 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.

As a sample configuration 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 develop 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 negative emotions, mainly to detect scenarios when the user experience is negative once the emotion recognizer is integrated into a conversational interface. The emotions considered are angry, bored and doubtful. 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

Our proposal to develop an emotion recognizer is based solely in acoustic and dialog information because in most application domains the user utterances are not long enough for the linguistic parameters to be significant for the detection of emotions. Our recognition method, is described in [80]. 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.

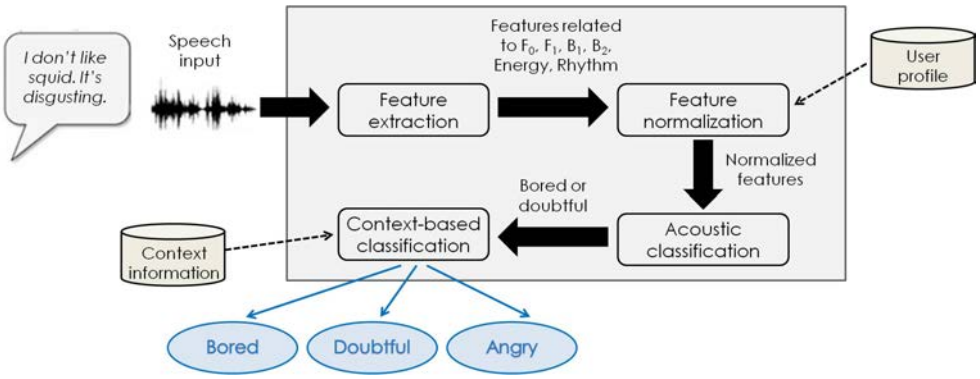

Figure 1: Schema of the proposed emotion recognizer

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 using the default values of their respective parameters provided by the Weka toolkit.

The second step of the emotion recognition process is feature normalization, with which the features extracted in the previous phase are normalized around

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 [81, 82, 74]

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 [80], 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*, 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

The proposed model for Sentiment Analysis aims to extend common sentiment classification of text, which is usually focused on polarity, to the classification of more categories. Thus, the main goal is to recognize a specific set of human emotions instead of whether a piece of text is negative, neutral or positive.

The Knowledge Base (KB) contains the main information sources used by the Analysis Module to extract sentiment values from words. The Analysis Module completes the words analysis. By splitting texts in sentences an tokenizing words, this module can query the Knowledge Base to extract emotional information or

know whether words are modifiers or carry an associated negation. Moreover, this module identify entities in the input text and track the number of occurrences of each one of them in a similar way bag-of-words models do this using occurrences vectors.

Once the entities have been identified and words are annotated with values from the KB, the Scoring Module computes the overall relevance of the entities and assigns a weighting factor for each of the words carrying emotional information, which are also known as concepts. A weight for each of the four independent emotional categories is then computed to classify the input text.

The last stage of the model deals with knowledge learning. To do this, the Learning Module takes as input the provided analysis from users when they disagree with the results of the Sentiment Analysis, and computes a learning factor to modify sentiment values of involved concepts.

3.2.1. Knowledge Base

As previously described, the Knowledge Base contains the main information sources used by the Analysis Module to extract sentiment values from words. In our proposal, this information has been classified into the following categories:

- **Concepts**: A concept refers to the emotions associated to a specific pair of $(word-PoS)$, where PoS (part of Speech) 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").

The NRC[5] and SenticNet[6] emotion lexicons have been used to complete the KB. Both are publicly available semantic resources for concept-level Sentiment Analysis. A total of 12,297 concepts are currently stored in the KB.

3.2.2. Parser Module

The parsing process of a sentence generates its semantically analysis containing part-of-speech tags organized in a tree of predicates. Between the set of general-purpose libraries currently available, we have selected OpenNLP[7]. This library supports the most common NLP tasks, such as tokenization, sentence segmentation, part-of-speech tagging, named entity extraction, chunking, parsing, and coreference resolution.

OpenNLP uses the Penn Treebank notation[8], which consider 36 sort of part of speech defined on the basis of their syntactic distribution rather than their semantic function. As a consequence nouns used in the function of modifiers are tagged as nouns instead of adjectives. Before parsing a text, it should be split into sentences by using the OpenNLP probabilistic *Sentence Detector*, which offers a precision of 94% and a 90% recall.

3.2.3. Emotion Classification Model

We have considered both rule-based approaches and machine learning techniques to perform the classification task. The rule-based algorithm developed for computing sentiment values of concepts is based both on distances of concepts to the selected representative nodes of all four categories and the weights assigned to each of them. The weights that are associated to each of the terms representing emotional categories are not trivial. They correspond to the maximum values of the different intensity levels of the original Hourglass model. The approach followed by the designed algorithm is to maximize the emotional intensities of concepts. Therefore, instead of using the returned distances of all the nodes of each

[5]http://www.saifmohammad.com/WebPages/lexicons.html
[6]http://sentic.net/
[7]https://opennlp.apache.org/
[8]http://www.cis.upenn.edu/ treebank/

category, only the most significant are considered. Figure 2 shows the designed algorithm.

Require: Term *concept*, Category *category*.
Ensure: Sentiment value for the specified category of the input term
$finalValue \leftarrow 0$
$maxDistance1 \leftarrow 0$
$maxDistance2 \leftarrow 0$
$weight1 \leftarrow 0$
$weight2 \leftarrow 0$
$auxiliaryMaxValue \leftarrow 0$ {Will store the max allowed value based on weights of maximum distances}
$targetNodes \leftarrow$ Nodes of the passed *category*
$distances \leftarrow$ Distances to $targetNodes$ {Array of distances preserving target nodes order}
$maxDistance2 \leftarrow 0$
for all *distances* **do**
    $nodeWeight \leftarrow$ Weight associated to node whose distance is being considered
    if $nodeDistance > maxDistance1$ **then**
        $maxDistance1 \leftarrow nodeDistance$
        $weight1 \leftarrow nodeWeight$
    else if $nodeDistance = maxDistance1$ **then**
        if $|nodeWeight| = |weight1|$ **and** $(nodeWeight - weight1) \neq 0$ **then**
            $maxDistance2 \leftarrow nodeDistance$
            $weight2 \leftarrow MAX(|nodeWeight|, |weight2|)$ with corresponding sign
        else if $(nodeWeight - weight1) \neq 0$ **then**
            $maxDistance2 \leftarrow maxDistance1$
            $weight2 \leftarrow weight1$
            $weight1 \leftarrow MAX(|nodeWeight|, |weight1|)$ with corresponding sign
        end if
    end if
end for
return $finalValue$

Figure 2: Rule-based algorithm for computing sentiment value for an affection category of a concept

We have considered different supervised learning techniques applied for classification purposes, such as Naive Bayes (NB), Maximum Entropy (ME), Support Vector Machines (SVM), Probabilistic Neural Networks (PNN), and Extreme Learning Machines (ELM).

- **The Naive Bayes (NB)** 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' [83]. 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 [83, 21]. 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 [83]. 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)** [84] 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 [18], ELMs offer fast learning speed, ease of implementation,and minimal human intervention.

3.2.4. Text Scoring Scheme and Adaptive Learning

Once the parsing process has finished and all the concepts, modifiers and negators have been properly tagged, it is possible to begin with the computation of the sentiment values of the text. The scoring process follows a bottom-up approach based on a fixed algorithm that relies on the Knowledge Base accuracy, a proximity based approach for modifiers, and a topic detection module to detect the most relevant topics of a text.

The way sentences are weighted is based on entities occurrences. Let w_i be the weight of a predicate and n the total number of sibling predicates that are being combined, the sentiment value of a category for weighted predicates can be defined as:

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

Our proposal also integrates an adaptive learning process for improving the Knowledge Base used for Sentiment Analysis. This process uses Eq. 2 to consider the difference between the Sentiment Analysis output proposed by the SA algorithm and the feedback provided by the user. Let U be the set of sentiments of a text corrected by the user, M be the sentiments calculated by the SA algorithm, W_{C_s} be the weight of concept C for sentiment s, and A_c be the number of accumulated adjustments of concept C. Therefore the new value of each sentiment s for a concept C is defined as:

$$C_s = C_s + \frac{(U_s - M_s) * W_{C_s}}{1 + (A_C/1000)} \tag{2}$$

3.3. Decision-level fusion

The main objective of the 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 for pattern recognition.

- In the simple voting approach, the phrase 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 rule-based approach:

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

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 [85], 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

We have selected a dataset for the experiments consisting of users' descriptions of images. The dataset contains 260 utterances by 35 users, which were manually-labeled to train the different classifiers.

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.

Figure 3 shows the process that we have followed to complete the evaluation. As it can be observed, we have assessed 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).

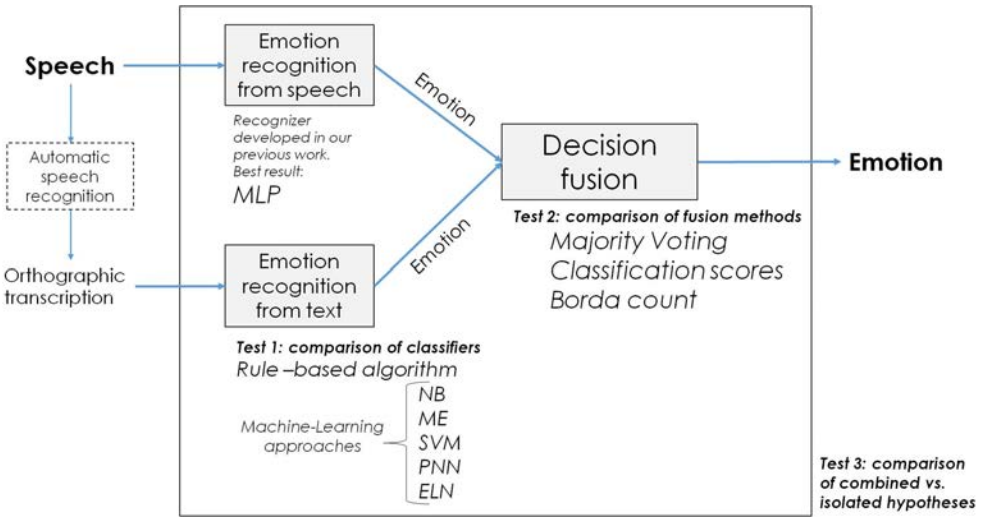

Figure 3: Experimental set-up showing the two evaluation processes

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. It 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 3.2.3, a designed rule-based algorithm and several 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. The 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. 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 we preferred the ELM as a classifier which provided the best performance in terms of both accuracy and training time.

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. In particular, SenticNet-based features mainly helped detect associated sentiments in text in an unsupervised way.

Classification technique	**Precision**	**Recall**	**F-Measure**	**Accuracy**
Rule-based Algorithm	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 (ELN)	0.79	0.81	0.79	0.79

Table 2: Results of the Test 1

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 gives the best results.

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

Table 3: Results of the Test 2

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

Table 4 shows the experimental results obtained if only the speech or the text classifier is used. It is clear from the table that the accuracy improves substantially when the two classifiers are combined.

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

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. The results show that the combined results outperformed the individual hypotheses and provide insight on the features, classifiers and combination approaches that can be employed for emotion 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] K. Ravi, V. Ravi, A survey on opinion mining and sentiment analysis: Tasks, approaches and applications, Knowledge-Based Systems 89 (2015) 14–46.

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

[3] 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.

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

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

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

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

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

[9] 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.

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

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

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

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

[14] 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.

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

[16] 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.

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

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

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

[25] 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.

[26] 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.

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

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

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

[30] 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.

[31] 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.

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

[33] 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.

[34] 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.

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

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

[37] C. Zhang, D. Zeng, J. Li, F.-Y. Wang, W. Zuo, Sentiment analysis of chinese documents: from sentence to document level, Journal of the American Society of Information Science Technology 60 (12) (2009) 2474–2487.

[38] T. Wilson, J. Wiebe, P. Hoffmann, Recognizing contextual polarity: an exploration of features for phrase-level sentiment analysis, Computational Linguistics 35 (3) (2009) 399–433.

[39] B. Ojokoh, O. Kayode, A feature-opinion extraction approach to opinion mining, Journal of Web Engineering 11 (1) (2012) 51–63.

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

[41] M. Hu, B. Liu, Mining and summarizing customer reviews, in: Proc. of KDD'04, 2004, pp. 168–177.

[42] S. Kim, E. Hovy, Determining the sentiment of opinions, in: Proc. of COLING'04, 2004, pp. 1367–1373.

[43] J. Kamps, M. Marx, R. Mokken, M. de Rijke, Using wordnet to measure semantic orientation of adjectives, in: Proc. of LREC'04, 2004, pp. 1115–1118.

[44] A. Esuli, F. Sebastiani, Random-walk models of term semantics: An application to opinion-related properties, in: Proc. of LTC'07, 2007, pp. 221–225.

[45] A. Montejo-Raez, L. U.-L. E. Martínez-Cámara, M.T. Martín-Valdivia, Ranked WordNet graph for sentiment polarity classification in twitter, Computur Speech and Language 28 (1) (2014) 93–107.

[46] J. Read, J. Carroll, Weakly supervised techniques for domain-independent sentiment classification, in: Proc. of TSA'09, 2009, pp. 45–52.

[47] G. Qiu, B. Liu, J. Bu, C. Chen, Opinion word expansion and target extraction through double propagation, Computational Linguistics 31 (1) (2011) 9–27.

[48] D. Rao, D. Ravichandran, Semi-supervised polarity lexicon induction, in: Proc. of EACL'09, 2009, pp. 675–682.

[49] S. Huang, Z. Niu, C. Shi, Automatic construction of domain-specific sentiment lexicon based on constrained label propagation, Knowledge-Based Systems 56 (2014) 191–200.

[50] S. Tan, Q. Wu, A random walk algorithm for automatic construction of domain-oriented sentiment lexicon, Expert Systems with Application 38 (10) (2011) 12094–12100.

[51] K. Liu, L. Xu, J. Zhao, Co-extracting opinion targets and opinion words from online reviews based on the word alignment model, IEEE Transactions on Knowledge & Data Engineering 27 (3) (2015) 636–650.

[52] D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, B. Qin, Learning sentiment-specific word embedding for twitter sentiment classification, in: Proc. of the 52nd Annual ,eeting of the ACL, 2014, pp. 1555–1565.

[53] Z. Zhang, M. Singh, Renew: A semi-supervised framework for generating domain-specific lexicons and sentiment analysis, in: Proc. of ACL'14, 2014, pp. 542–551.

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

[55] A. Pak, P. Paroubek, Twitter as a corpus for sentiment analysis and opinion mining, in: Proc. of LREC'10, 2010, pp. 1320–1326.

[56] A. Konig, E. Brill, Reducing the human overhead in text categorization, in: Proc. of KDD'06, 2006, pp. 598–603.

[57] J. Chenlo, D. Losada, An empirical study of sentence features for subjectivity and polarity classification, Information Science 280 (2014) 275–288.

[58] Z. Cui, X. Shi, Y. Chen, Sentiment analysis via integrating distributed representations of variable-length word sequence, Neurocomputing 187 (2016) 126–132.

[59] G. Vinodhini, R. Chandrasekaran, A comparative performance evaluation of neural-network based approach for sentiment classification of online reviews, Journal of King Saud University 28 (2016) 2–12.

[60] M. Hajmohammadi, R. Ibrahim, A. Selamat, H. Fujita, Combination of active learning and self-training for cross-lingual sentiment classification with density analysis of unlabelled samples, Information Sciences 317 (2015) 67–77.

[61] N. F. F. da Silva, L. F. Coletta, E. R. Hruschka, E. R. H. Jr., Using unsupervised information to improve semi-supervised tweet sentiment classification, Information Sciences (2016) 1–18.

[62] F. Khan, U. Qamar, S. Bashir, SWIMS: Semi-supervise d subjective feature weighting and intelligent model selection for sentiment analysis, Knowledge-Based Systems 100 (2016) 97–111.

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

[64] A. Fahrni, M. Klenner, Old wine or warm beer: target-specific sentiment analysis of adjectives, in: Proc. of AISB'08, Vol. 1, 2008, pp. 1–27.

[65] M. Missen, M. Boughanem, Using wordnet's semantic relations for opinion detection in blogs, in: Proc. of of ECIR'09, Vol. 1, 2009, pp. 729–733.

[66] M. Tsytsarau, T. Palpanas, K. Denecke, Scalable discovery of contradictions on the web, in: Proc. of the WWW'10, Vol. 1, 2010, pp. 1195–1196.

[67] T. Li, Y. Zhang, V. Sindhwani, A non-negative matrix tri-factorization approach to sentiment classification with lexical prior knowledge, in: Proc. of ACL'09, 2009, pp. 244–252.

[68] K. Quinn, B. Monroe, M. Colaresi, M. Crespin, D. Radev, How to Analyze Political Attention with Minimal Assumptions and Costs, American Journal of Political Science 54 (1) (2010) 209–228.

[69] G. Zhou, J. Zhao, D. Zeng, Sentiment classification with graph co-regularization, in: Proc. of COLING'14, 2014, pp. 1331–1340.

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

[71] 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.

[72] 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.

[73] 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.

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

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

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

[77] G. Verma, U. Tiwary, S. Agrawal, Error weighted semi-couple d hidden Markov model for audio-visual emotion recognition, Advances in Computing and Communications 192 (2011) 452–459.

[78] 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.

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

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

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

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

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

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

[85] 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.

Dr. David Griol obtained his Ph.D. degree in Computer Science from the Technical University of Valencia (Spain) in 2007. He has also a B.S. in Telecommunication Science from this University. He is currently visiting lecturer at the Department of Computer Science in the Carlos III University of Madrid (Spain). He has participated in several European and Spanish projects related to natural language processing and dialog systems. His research activities at the Applied Artificial Intelligence Group are mostly related to the development of statistical methodologies for the design of spoken dialog systems. His research interests include dialog management and optimization, corpus-based methodologies, user modeling and simulation, adaptation and evaluation of spoken dialog systems and machine learning approaches.

Dr. José Manuel Molina is Full Professor at Universidad Carlos III de Madrid, Spain. He joined the Computer Science Department of the same university in 1993. Currently, he coordinates the Applied Artificial Intelligence Group (GIAA). His current research focuses on the application of soft computing techniques (NN, evolutionary computation, fuzzy logic, and multiagent systems) to radar data processing, air traffic management, and e-commerce. He (co)authored up to 50 journal papers and 200 conference papers. He received a degree in telecommunications engineering from the Technical University of Madrid in 1993 and the Ph.D. degree from the same university in 1997.

Dra. Zoraida Callejas is Assistant Professor in the Department of Languages and Computer Systems at the Technical School of Computer Science and Telecommunications of the University of Granada (Spain). She completed a PhD in Computer Science at University of Granada in 2008 and has been a visiting researcher in University of Ulster (Belfast, UK), Technical University of Liberec (Liberec, Czech Republic), University of Trento (Trento, Italy), University of Ulm (Ulm, Germany), Technical University of Berlin (Berlin, Germany) and Telecom ParisTech (Paris, France). Her research activities have been mostly related to speech technologies and in particular to the investigation of affective dialogue systems. She has participated in numerous research projects, and is a member of several research associations focused on speech processing and human-computer interaction.