Intensional Learning to Efficiently Build up Automatically Annotated Emotion Corpora

Lea Canales, Carlo Strapparava, Ester Boldrini, and Patricio Martínez-Barco

Abstract—Textual emotion detection has a high impact on business, society, politics or education with applications such as, detecting depression or personality traits, suicide prevention or identifying cases of cyber-bulling. Given this context, the objective of our research is to contribute to the improvement of emotion recognition task through an automatic technique focused on reducing both the time and cost needed to develop emotion corpora. Our proposal is to exploit a bootstrapping approach based on intensional learning for automatic annotations with two main steps: 1) an initial similarity-based categorization where a set of seed sentences is created and extended by distributional semantic similarity (*word vectors* or *word embeddings*); 2) train a supervised classifier on the initially categorized set. The technique proposed allows us an efficient annotation of a large amount of emotion data with standards of reliability according to the evaluation results.

Index Terms—Affective Computing, Corpora Annotation, Sentiment Analysis, Textual Emotion Recognition

1 Introduction

AUTOMATIC detection of affective states in text is becoming more and more important due to the fact that it has the potential of bringing substantial benefits for different sectors. Example of this can be for instance the applications in e-learning environment [1], [2]; suicide prevention [3], [4]; depression detection [5]; identification cases of cyber-bullying [6]; or tracking well-being [7].

This paper is focused on textual emotion detection because it is one of the main media employed to interact with humans through chats room, public reviews, emails, social networks or web blogs. This produces an abundance volume of information and hence requires a computer processing to classify automatically the text in accordance with its emotional degree and orientation.

The creation of an accurate emotion detection system would allow evaluating and representing people's emotions through automatically analyzing online content, such as the comments on the Social Web. This system would be composed of one module of emotion detection, where the comments written on channels of Web 2.0 by users would be analyzed. Taking into account that these commentaries also contain geographic and temporal information, the system would allow us to determine the welfare state of a social group in a specific place and temporal range. Therefore considering the above, this research area could be very beneficial from a social point of view.

The majority emotion recognition systems developed so far consists in supervised machine-learning approaches; systems first infer a function from a set of examples labeled with the correct emotion (this set of examples is called the training data or labelled corpus). After this, the model is able to predict the emotion of new examples. Hence, the training dataset in supervised machine learning algorithms is crucial to building accurate emotion detection systems with reliable results.

Despite its importance and need, the creation of a labelled corpus is not trivial; detecting emotion in text can be difficult even for humans due to the influence of each own background that can affect emotion interpretation. Generally, corpora have been annotated manually since, in this way, machine learning systems learned from human annotation. Although, as Mohammed [8] analyzed, manual annotations can be significantly influenced by a set of different factors, such as clarity of instructions, the difficulty of the task, training of the annotators, and even by the annotation scheme whose creation requires a hard and time-consuming work. More importantly, due to the subjectivity of the task most relevant research carried out so far has shown that the agreement between annotations when associating emotion to instances is significantly lower compared to other Natural Language Processing (NLP) tasks. This produce that the availability of emotion corpus which can be used in supervised machine learning algorithms is low. Consequently, in this paper, an automatic technique is proposed to create emotion resources reducing the cost and time-consuming to build, with aim of improving the emotion detection task.

Our proposal is to exploit a bootstrapping approach based on intensional learning from an emotional point of view. This algorithm consist of two main steps: 1) an initial similarity-based categorization where a set of seed sentences is created and this seed is extended by distributional semantic similarity (word vectors or *embeddings*); and 2) train a supervised classifier on the initially categorized set.

The proposed technique will allow us the annotation of a large amount of emotion data with efficiently and standards of reliability proven by the evaluation performed. In partic-

- *L. Canales is with the Department of Software and Computing System, University of Alicante, Alicante, Spain. E-mail: lcanales@dlsi.ua.es.*
- *C. Strapparava is with the Human Language Technology, Fondazione Bruno Kessler, Trento, Italy. E-mail: strappa@fbk.eu.*
- *E. Boldrini is with the Department of Software and Computing System, University of Alicante, Alicante, Spain. E-mail: eboldrini@dlsi.ua.es.*
- *P. Martínez-Barco is with the Department of Software and Computing System, University of Alicante, Alicante, Spain. E-mail: patricio@dlsi.ua.es.*

Manuscript received April 19, 2005; revised August 26, 2015.

ular, our approach has been assessed in two corpora: Aman corpus [9] and Affective Text corpus [10] where an emotional model built from the corpus annotated automatically and the agreement between corpus annotated automatically and the gold standard of both corpora is evaluated.

The rest of the paper is organized as follows. Section 2 presents the related work and some comments on the pending issues. After this, the proposed method is described in detail in Section 3. Then, Section 4 is aimed at showing the evaluation methodology, the results obtained and a discussion about these results. Finally, Section 5 details our conclusions and future works.

2 State of the Art

This section discusses the state of the art of different aspects related to our approach. On the one hand, bootstrapping technique and semantic similarity metrics are analyzed since both are the pillars of our approach. On the other hand, an exhaustive review of emotion lexicons and corpora is carried out with the aim of obtaining conclusions and determining the pending issues.

2.1 Bootstrapping technique

Bootstrapping is a strategy to automatically generate a number of sufficient instances from a small set of seed words, phrases or sentences. This technique was proposed to avoid, or at least considerably reducing, the need for manual corpora annotation. Hence, it has become an important topic in computational linguistics since for many language-processing tasks there is an abundance of unlabeled data.

There is a large variety of bootstrapping implementations, but one of the most used in NLP is the bootstrapping based on an iterative algorithm which it starts from a seed of words or sentences and in each iteration the most confident predictions of the algorithm are added to the initial seed. For instance, this bootstrapping method was used to extract automatically patterns to identify subjective words [11], [12]; or for the construction of English and Italian corpora from the domain in Psychiatry via automated Google queries with an iterative algorithm [13]. In this article, we follow an *Intensional Learning* (IL) proposal, an alternative bootstrapping approach proposed in [14]. They refer to the standard example-based supervision mode as *Extensional Learning* (EL) since classes are being specified by means of examples of their elements and feature-based supervision is referred to as IL, as features may often be perceived as describing the *intension* of a category.

In the literature, IL approach can be found as a technique for bootstrapping an extensional learning algorithm, as in word sense disambiguation [15]; or named entity classification [16].

According to [14], it is possible to recognize a common structure behind these works:

Step 1 *Initial similarity-based categorization*. This step is approached by applying some similarity criterion between the initial category seed and each unlabeled sentence. The result of this step is an initial categorization of (possibly a subset of) the unlabeled documents.

Step 2 *Train a supervised classifier on the initially categorized set*. The output of step 1 is exploited to train a supervised classifier. Different learning algorithms have been tested, as Support Vector Machines (SVMs) or Naive Bayes.

The core part of IL bootstrapping is step 1, that is, the initial unsupervised classification of the unlabeled dataset. This step has often been approached by simple method assuming that the supervised training algorithm would be robust enough to deal with noise from the initial set. Despite this assumption, the effectiveness of the first step is crucial for IL bootstrapping.

As mentioned previously, bootstrapping technique has been suitable to tackle many computational linguistics challenges as Word Sense Disambiguation [15]; Named Entity classification [16]; Information Extraction [17]; or Sentiment Analysis [18] since labelled data are lacking and too expensive to create in large quantities. This is also true for emotion detection task. Thus, the aim of our research is to tackle and resolve the lack of labelled emotion data by an automatic technique.

For that purpose, a bootstrapping process based on IL is proposed to create an emotion corpus annotated automatically with the aim of not to be dependent on the manual-labelled examples as in EL. In this way, the annotations are not influenced by the background of each annotator, since the personality or personal situation of each one can influence emotion interpretation. Specifically, the process has the common structure explained above 1) an *initial similarity-based categorization* where a set of seed sentences is created and then extended by the semantic similarity between sentences; 2) *train a supervised classifier on the initially categorized set*. Unlike EL, the IL approach is based on the classical rule-based classification method, where the user specifies exact classification rules that operate in the features space. In our case, this set of rules are specified in the algorithm explained in Section 3.1. This algorithm employs each seed keyword as features and thus, the initial seed of sentences is annotated based on the keywords contained in each sentence. Furthermore, in this step 1, the initial categorization is extended by semantic similarity (Section 3.2) with the aim to increase the instances to train a supervised classifier in step 2.

2.2 Semantic similarity of texts

The first step of the bootstrapping process is approached by applying similarity criterion. Our approach employs semantic similarity metrics to tackle this step due to the fact these measures have been used for a long time in a wide number of NLP applications such as word sense disambiguation [19]; text summarization [20]; or the evaluation of text coherence [21] with satisfactory performance. Thus, the employment of semantic similarity metrics for extending the initial seed sentences (step 1) is considered interesting and suitable.

Many approaches have been suggested to determine the semantic similarity between text such as approaches based on lexical matching, handcrafted patterns, syntactic parse trees, external sources of structured semantic knowledge and distributional semantics [22].

This research is focused on distributional semantics because we aim to employ a generic model that does not require lexical and linguistic analysis and does not use external sources of structured semantic knowledge.

Distributional Semantic Models (DSM) are based on the assumption that the meaning of a word can be inferred from its usage. Therefore, these models dynamically build semantic representations (high-dimensional semantic vector spaces) through a statistical analysis of the contexts in which the words occur[1]. Finally, each word is represented by a real-valued vector called *word vector* or *word embedding* and the geometric properties of high-dimensional semantic vector spaces prove to be semantically and syntactically meaningful ([23], [24]), thus words that are semantically or syntactically similar tend to be close in the semantic space.

Latent Semantic Analysis (LSA) and Word2Vec algorithms incorporate this intuition. On the one hand, LSA [25] builds a word-document co-occurrences matrix and performing a dimensional reduction by a Singular Value Decomposition (SVD) on it to get a lower-dimensional representation, whereas Word2Vec algorithm [23] learns a vector-space representation of the terms by exploiting a two-layer neural network. There are two architectures of Word2Vec: continuous bag-of-words (CBOW) that predicts the current word based on the context; and Skip-gram (SKIP) which predicts surrounding words given the current word. Both algorithms are employed to build DSMs in this research for several reasons: (i) both methods allow us to employ generic model to calculate the semantic similarity by measuring the distance between the word vectors in the extension of the seed; (ii) LSA and Word2vec algorithms have demonstrated their effectiveness to calculate the semantic similarity in many NLP task such as e-learning [26], text-categorization [14], [27]; emotion detection [28]; or sentiment analysis [29]; (iii) allow us to compare LSA, a consolidated and traditional technique, and Word2Vec, a recent technique based on neural networks, in emotion detection task.

Compositional Distributional Semantic Models (CDSMs) are employed to determine semantic similarity of sentences/phrases by word embedding. These models are an extension of DSMs that characterize the semantics of entire phrases or sentences. This is achieved by composing the distributional representations of the words that sentences contain [30].

Among these models, the approach employed in our research has been used in [31] and it is called *VectorSum*. This method consists of adding the vectors corresponding to non-stop words in bag of words (BOW) A and B, resulting in a vector V_A and V_B, respectively. The selection of this approach as CDSMs is due to its simplicity and because as [31] demonstrated, these vectors are able to capture the semantic meaning associated with the contexts, enabling us to gauge their relatedness using cosine similarity.

2.3 Emotion Lexicons

According to research in psychology, there is a number of theories about how to represent the emotions that humans perceive and express. Although, there are two main different models that represent emotions: the *categorical* and the *dimensional*.

Categorical emotion models assume that there are a discrete emotional categories or labels. According to this model, there is a set of basic emotions, which are often considered universals, employed to express and perceive emotions. Ekman's emotions [32] are the most popular set of basic emotions and determine *anger, disgust, fear, joy, sadness and surprise* as the set of basic emotions. However, nowadays Plutchick's emotions [33] are also another set of emotions commonly employed. Plutchik defines a set of eight basic bipolar emotions, considering of a superset of Ekman and with two additions: *trust* and *anticipation*. These eight emotions are organized into four bipolar sets: joy vs. sadness, anger vs. fear, trust vs. disgust and surprise vs. anticipation.

Dimensional emotion models represent affects in a dimensional form where each emotion occupies a location in this space. One of the most representative models of these approaches is Rusell's Circumplex [34] who suggests a model of affect where emotions are distributed in a two-dimensional circular space: *valence dimension* and *arousal dimension*. The valence dimension indicates how *positive* and *negative* is an emotion whereas the arousal dimension differentiates *excited* and *calm* states. Within dimensional models, there are also models based on three dimensions as the Mehrabian's model with PAD (Pleasure - Arousal - Dominance) representation [35] where the *dominance dimension* indicates whether the subject feels in control of the situation or not.

Regardless the existence of two main different models, most emotion recognition works are focused on the categorical model since automatic classification text according to their emotional content is a complex task, and the use of a set of limited categories makes the task easier to deal with.

Emotion lexicons are a set of words labelled according to their emotional connotation. The label can be an emotional category when the categorical model is employed or can be a value of the strength of a given emotion dimension in a concrete dimensional model. Thus, in the literature, we can find lexicons labelled with one or another model.

ANEW [36]: the Affective Norms for English Words is a set of normative emotional ratings for a large number of words in the English language (1,034 English words including verbs, nouns, and adjectives). Each word is rated from 1 to 9 in terms of the three dimensions of *valence*, *arousal* and *dominance*.

DAL [37]: the Dictionary of Affective Language contains 8,742 words which have been evaluated by people for their *activation*, *evaluation* and *imagery*.

WordNet Affect (WNA) [38]: WordNet-Affect is an extension of WordNet Domains that includes a subset of synsets suitable to represent affective concepts correlated with affective words. The affective concepts representing emotional states are individuated by synsets marked with the a-label emotion. There are also other a-labels for those concepts representing moods, situations eliciting emotions or emotional responses.

NRC Word-Emotion Association Lexicon (Emolex) [39]: the NRC Word-Emotion Association Lexicon (also called

1. http://wordspace.collocations.de/doku.php/course:acl2010:start

Emolex) is a dataset of general domain consisting about 14,000 English unigrams (words) associate with the Plutchik's eight basic emotions [33] (*anger*, *fear*, *anticipation*, *trust*, *surprise*, *sadness*, *joy*, and *disgust*) and two sentiments (*negative* and *positive*), compiled by manual annotation.

NRC Hashtag Emotion Lexicon [40]: the NRC Hashtag Emotion Lexicon is a dataset of Twitter domain consisting about 16,000 words associated with the Plutchik's eight basic emotions [33] by automatic annotation from tweets with emotion word hashtags.

EmoSenticNet (ESN) [41]: EmoSenticNet is a lexical resource of 13,171 words that assigns qualitative emotions label and quantitative polarity scores to SenticNet concepts [42]. WordNet Affect emotion labels (Ekman's emotions: *anger*, *fear*, *disgust*, *sadness*, *surprise*, or *joy*) is the set of emotions employed for labelling the concepts.

DepecheMood [43]: DepecheMood is a lexicon of roughly 37 thousand terms associated with emotion scores of eight emotion categories (*happy*, *sad*, *angry*, *afraid*, *annoyed*, *inspired*, *amused*, and *don't care*). It was built exploiting the crowd-sourced affective annotation implicitly provided by readers of news articles from Rappler website[2].

TABLE 1
Emotion Lexicon

Dictionary	Size	Emotions	Annotation
ANEW [36]	1,034 words	Valence, Arousal and Dominance	Manual
DAL [37]	8,742 words	Activation, Evaluation and Imaginary	Manual
WordNet Affect (WNA) [38]	4,787 terms	Semantic levels, valence, arousal and affective labels	Manual
NRC Emotion Lexicon (Emolex) [39]	14,182 words	Plutchick's emotions	Manual
NRC Hashtag Emotion Lexicon [40]	16,862 words	Plutchick's emotions	Automatic
EmoSenticNet [41]	13,171 words	Ekman's Emotions	Automatic
DepecheMood (DPM) [43]	37,216 words	Eight emotion categories	Automatic

As the majority of works developed so far, our approach is based on categorical emotion model. Concretely, Ekman's basic emotions are employed since the corpora used to evaluate are annotated with this set of emotions. Although in our approach, the group of emotions is adaptable provided that the process employs an emotion lexicon annotated with the desired emotions.

In our research, NRC Word-Emotion Association Lexicon (Emolex) is adopted because: (i) it is general domain and it can be applied in different corpora; (ii) it is annotated with a superset of Ekman's six basic emotions; and (iii) the most relevant feature of this resource is that the terms in this lexicon are carefully chosen so that some of the most frequent nouns, verbs, adjectives and adverbs allowing us to create a general method for any genre are included.

2. www.rappler.com

2.4 Emotion Corpora

An emotion corpus is a large and structured set of sentences where each sentence is tagged with one or more emotional tags. Corpora are a fundamental part of supervised learning approaches, as they rely on a labelled training data, a set of examples. Supervised learning algorithm analyzes the training data and infers a function, which it use for mapping new examples [44].

As mentioned in the previous section, most automatic emotion detection systems are focused on a limited set of proposed basic emotions. Thus, this state-of-art is focused on the corpora annotated with categorical emotion model since the majority of corpora developed so far have been annotated with this model. Table 2 shows the corpora analyzed. For a more detailed overview of resources and corpora for sentiment analysis (subjectivity, polarity or psycholinguistic) and emotion detection in social media, see [45].

Emotion annotation task has been majority approached with a manual process, as Table 2 shows. In this way, machine learning systems learn from human annotations that are generally more accurate. Among these resources, we can find corpora labelled with the six basic emotions categories proposed by Ekman such as: [46] annotated a sentence-level corpus of approximately 185 children stories with emotion categories; [9] annotated blog posts collected directly from Web with emotion categories and intensity; or [10] annotated news headlines with emotion categories and valence.

Although there are corpora labelled with another small set of emotions by manual annotation like ISEAR corpus [47] contains reports on seven emotions each by close to 3,000 respondents in 37 countries on all 5 continents; [49] a corpus extracted 700 sentences from blog posts provided by BuzzMetrics annotated with one of nine emotion categories (a subset of emotional states defined by Izard) and a corresponding intensity value; [48] corpus extracted 1,000 sentences from various stories annotated with one of 14 categories of their annotation scheme (between them Izards emotions [57]) and a corresponding score (the strength or intensity value); Emotiblog-corpus that consists of a collection of blog posts annotated with three annotation levels: document, sentence and element using a group of 15 emotions [50]; or EmoTweet-28 corpus that consists of a collection of tweets annotated with 28 emotion categories. The corpus contains annotations for four facets of emotion: valence, arousal, emotion category and emotion cues [51].

Most recently, there is a new family of social media corpora which include multi-layered manual annotations, among which the annotations for basic emotions are also included. Examples of these corpora are: 1) the Twitter corpus developed by [52] where a set of 2012 US presidential election tweets were annotated by crowdsourcing for a number of attributes pertaining to sentiment, emotion, purpose

TABLE 2
Emotion Corpora

Corpus	Source	Size	Emotions	Annotation
Alm corpus [46]	Children stories	185 stories 1,200 sentences	Ekman's emotions	Manual
Aman corpus [9]	Blog posts	173 blog posts 4,000 sentences	Emotion category (Ekman's emotions) Emotion intensity (high medium, low)	Manual
Affective Text Semeval 2007 [10]	News headlines	1,250 headlines	Ekman's emotions and Valence	Manual
ISEAR [47]	Reports	7,667 sentences	7 Emotion categories	Manual
Nevioroskaya corpus (1) [48]	Stories	1,000 sentences	14 emotional categories (between them Izard's emotions)	Manual
Nevioroskaya corpus (2) [49]	Blog posts	700 sentences	Izard's emotions and Polarity categories	Manual
Emotiblog-corpus [50]	Blog posts	1,950 sentences	15 emotions	Manual
EmoTweet-28 corpus [51]	Tweets	15,553 tweets	28 Emotion Categories, Valence, Arousal and Emotion Cues	Manual
2012 US Electoral corpus [52]	Tweets	1,600 tweets	Sentiment, Emotions (Plutchick's emotions), Purpose, and Style	Manual
Multi-View Sentiment corpus [53]	Tweets	3,000 tweets	Subjectivity, Emotions (Plutchick's emotions), Sentiment, Explicit/Implicit, and Irony	Manual
Twitter Emotion Corpus (TEC) [54]	Tweets	21,000 tweets	Ekman's emotions	Automatic
Choudhurry corpus [55]	Tweets	6.8 million tweets	11 Emotion categories	Automatic
Wang corpus [56]	Tweets	2.5 million tweets	7 Emotion categories	Automatic

or intent behind the tweet, and style; and 2) [53] present a Multi-View Sentiment Corpus (MVSC), manually labelled with several aspects of the natural language text: subjective/objective, sentiment polarity, implicit/explicit, irony and emotion, which comprises 3,000 English microblogs posted related to the movie domain.

The main drawback of manual annotation is that the process is very time consuming and expensive due to different factors like the annotation scheme, the difficulty of the task or the training of the annotators. These aspects are even more complex to define in emotion annotation task because of its highly subjective. This produces the need to invest in many resources to annotate large scale emotion corpora.

Consequently and with the aim of overcoming the cost and time-consuming shortcoming of manual annotation, several emotion resources have recently been developed employing emotion word hashtags to create automatic emotion corpus from Twitter. Mohammad [54] describe how they created a corpus from Twitter post (Twitter Emotion Corpus - TEC) using this technique. In literature, several works can be found with the use emotion word hashtags to create emotion corpora from Twitter [55], [56]. Aiming at sparing annotation effort, distant supervision using Facebook reactions has also been explored as an alternative way to obtain training data in emotion detection. For instance, [58] exploit the Facebook reaction feature in a distant supervised fashion to train a support vector machine classifier for emotion detection.

Thus, in textual emotion recognition research community, the interest in developing amounts of emotion corpora has increased because that would allow improving supervised machine learning systems. Example of this is the new shared task on Emotion Intensity (EmoInt) proposed at WASSA 2017, which proposes a novel related challenge for emotion detection[3]. Thus, despite that the use of emotion

3. https://www.aclweb.org/portal/content/wassa-2017-shared-task-emotion-intensity

word hashtags as a technique to label data is simple and efficient in terms of time and cost, this method can be exclusively applied to social networks and micro-blogging services. For this reason, our objective is to develop a bootstrapping technique for large-scale annotation in any genre with the aim to reduce the cost and time-consuming.

3 Automatic Association between Emotions and Sentences

The objective of emotion annotation task is to annotate unlabeled sentences with the emotions expressed in each sentence. For this purpose, a bootstrapping technique based on *intensional learning* is presented. The common structure of this technique consists of 1) the initial similarity-based categorization and 2) training a supervised classifier on the initially categorized set.

Specifically, our bootstrapping approach consists in two unsupervised mechanisms within the initial categorization step: 1.1) the creation of an initial seed where NRC Emotion lexicon is employed to annotate the sentences by its emotional words (Section 3.1); and 1.2) the extension of the initial seed based on the measure of the semantic similarity between sentences (Section 3.2) ([59], [60]). The overview of the initial categorization step is described in Figure 1.

The process receives as input data a collection of unlabelled sentences/phrases, a set of emotional categories, specifically this paper works with the Ekman's six basic emotions [32], and the number of these categories to annotate (one or more categories).

The characteristics of the proposal to be adaptable to the set of emotional categories, as well as to be flexible to the number of emotional categories annotated (the predominant emotion or all of the emotions detected), are interesting and novel. This flexibility allows the use of this technique in different domains or applications. For instance, *boredom*, *anxiety* and *excitement* emotions are typically detected in education domain [61], whereas emotions like *amused* or *inspired* are analyzed in news domain[4]. Moreover, this adaptability can be useful in those applications where the detection of the emotion intensity is important such as recommender systems.

The section is divided into three subsections where the main tasks carried out by the process are explained.

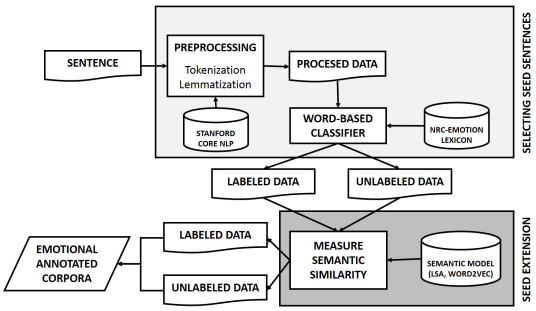

Fig. 1. Overview of the initial categorization step (Step 1 of bootstrapping).

4. http://www.rappler.com/

3.1 Selection of Seed Sentences by Emolex

In this section, the process of creating the initial seed by exploring NRC Word-Emotion Association Lexicon (Emolex) [39] is presented.

Emolex is a lexicon of general domain consisting of 14,000 English unigrams (words) associated with an emotional vector of Plutchik's eight basic emotions [33] (*anger*, *fear*, *anticipation*, *trust*, *surprise*, *sadness*, *joy* and *disgust*) and two sentiments (*negative* and *positive*), compiled by manual annotation. Our approach only employs the Ekman's basic emotions and for this reason, the lexicon is reduced to 3,462 English unigrams. The coverage of this reduced version of Emolex is shown in Table 3.

TABLE 3
The Coverage of Emotions in the reduced version of Emolex

	Emolex (Ekman's emotions) # of Words
Anger	1,247
Disgust	1,058
Fear	1,476
Joy	689
Sadness	1,191
Surprise	534

Due to this reduction, the improvement of Emolex with synonyms can be considered relevant to test a different set of seeds. For this reason, Emolex is extended automatically with Wordnet (WN) [62] and Oxford [63] synonyms. Hence, three approaches are presented where each one employs different versions of Emolex (original approach, enriched approach by WN synonyms and enriched approach by Oxford synonyms). The process of the seed creation is the same for all of the approaches. The extension process of Emolex is completely automatic and is explained in detail in subsection 3.1.1 and subsection 3.1.2.

The algorithm of the creation of the seed consists of:

- Step 1: each sentence has an emotional vector associated with a value of each emotion ([`ANGER`, `DISGUST`, `FEAR`, `JOY`, `SADNESS`, `SURPRISE`]) initialized to zero.
- Step 2: each sentence is tokenized and lemmatized using Stanford Core NLP [64].
- Step 3: each word of the sentence is looked up in Emolex. If a word is in Emolex, its emotional values are added to the emotional vector of the sentence.
- Step 4: if the process annotates the predominant emotion, each sentence is annotated with the emotion whose value is the highest in the emotional vector of the sentence. Instead, if the process annotates all of the emotions expressed in the sentence, each sentence is annotated with all of the emotions detected.

Figure 2 shows two examples of the creation of the seed. Sentence 1: *"We played fun baby games and caught up on some old time"*, whose emotional vector is initialized to zero, contains three emotional words: 'fun', 'baby' and 'catch'. The values of these three words are added and the sentence has finally this vector: [0, 0, 0, 2, 0, 1] associated.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2017.2764470, IEEE Transactions on Affective Computing

This sentence will have JOY emotion associated because this emotion has the highest value associated when the process is detecting the predominant emotion and will have JOY and SURPRISE emotions associated when all of the emotions are detected. Sentence 2: *"My manager also went to throw a fake punch."*, whose emotional vector is initialized to zero, contains one emotional word: 'punch'. The sentence has finally this vector: [1, 0, 1, 0, 1, 1] associated. Hence, if the process is detecting the predominant emotion, this sentence will be not associated any emotion, whereas this sentence will have ANGER, FEAR, SADNESS and SURPRISE emotions associated when the objective is to detect all emotions.

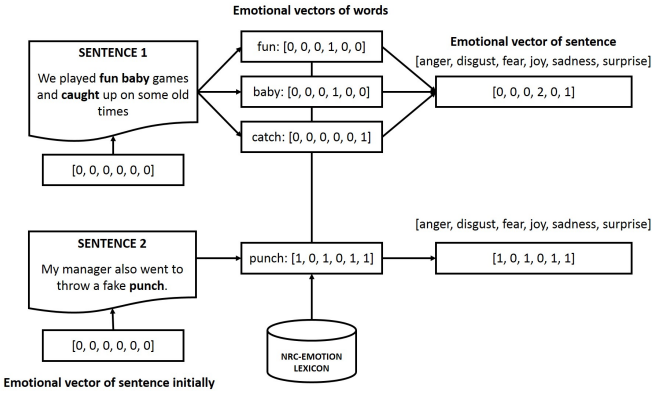

Fig. 2. Examples of the process of selecting seed sentences.

Linguistic phenomena such as negation or irony have not been addressed in this approach because the objective of our research is to propose a technique for large-scale annotation in any genre with the aim to reduce the cost and time-efforts. The management of these phenomena introduces a high level of complexity in the approach since the detection of these aspects requires an analysis in depth of each genre, thereby hampering the achievement of our purpose.

3.1.1 Enriched approach by WordNet synonyms

The extension of Emolex employing WordNet [62] synonyms is one of the enriched approaches evaluated.

In this process, each word contained in Emolex is looked up in WordNet, the synonyms of its more frequent sense are obtained and annotated with the emotions of the Emolex word. Figure 3 shows an example of the process. The word 'alarm' is contained in Emolex and has the emotions FEAR and SURPRISE associated. The process looks up 'alarm' in WordNet and obtains the synonyms of its more frequent sense: 'dismay' and 'consternation'. These synonyms are added to Emolex and annotated with the same emotions of 'alarm'.

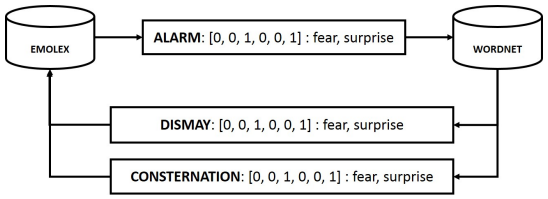

Fig. 3. Process of the extension of Emolex by WordNet synonyms.

After the process, Emolex has been extended with 4,029 words more resulting in a lexicon with 7,491 words.

3.1.2 Enriched approach by Oxford synonyms

The enriched approach by Oxford synonyms is carried out with the aim of analyzing the relevance of selecting a set of synonyms or other.

First, each word contained in Emolex is looked up in the Oxford American Writer Thesaurus [63] and all of the synonyms for all of its senses are collected. Then, each synonym of a word is associated with the emotions of the Emolex word and is added in Emolex. If a synonym is already in Emolex, their emotions associated will be the result of matching the emotional vector stored in Emolex and the new emotional vector.

Figure 4 shows an example of the process that corresponds the word 'sickness'. The first step gets their Oxford synonyms and for each synonym (in this example the synonym 'vomiting'): 1) associate the emotions of 'sickness', this is, DISGUST, FEAR and SADNESS; and 2) check if 'vomiting' is already in Emolex. If it is not, their emotions associated will be the same as 'sickness'. In another case, their emotional vector will contain the emotion in common between the vector saved in Emolex (old) and the new emotional vector (new). In this case, 'vomiting' will be associated with DISGUST emotion.

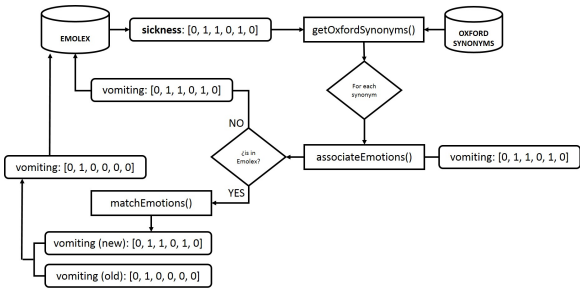

Fig. 4. Process of the extension of Emolex by Oxford synonyms.

After the process, Emolex has been extended with 6,789 words more, resulting in a lexicon with 10,251 words.

3.2 Seed Extension via Semantic Similarity

After obtaining the initial seed sentences, the next step will consist into increasing the number of annotated sentences with the help of Compositional Distributional Semantic Models (CDSMs).

As we mentioned previously, Compositional Distributional Semantic Models (CDSMs) are an extension of Distributional Semantic Models (DSMs) that characterise the semantics of entire phrases or sentences. With these models, the seed sentences will be extended based on the semantic similarity between annotated and non-annotated sentences.

The extension process of the seed consists of:

- Step 1: annotated and non-annotated sentences are represented by distributional vectors employing different DSMs and calculate the cosine similarity between them. The representation of each sentence is

achieved by adding the distributional vectors corresponding to non-stop words of each sentence.
- Step 2: when the similarity is higher than 80%, the non-annotated sentences are annotated with the emotions of the annotated one. The use of a strict similarity (80%) allows us to ensure that the seed is extended with high confidence. This value was empirically determined with different similarity thresholds. These experiments showed that employing thresholds lower than 80% added a lot of noise to the seed since the precision was too low.

Regarding the DSMs employed, a Latent Semantic Analysis (LSA) model and three Word2Vec (W2V) models are employed: a LSA model applied in [65], which is run on the lemmas of the British National Corpus (BNC)[5]; two Word2Vec models (CBOW and Skip-gram architecture) are run on the lemmas of New York Times Newswire Service from the Annotated English Gigaword[6]; and a Word2Vec model (CBOW) applied in [66], which is run on the words of the BNC and WackyPedia/ukWac.

In this process, non-annotated sentences could be matched to two or more annotated sentences. The process selects the annotated sentence whose similarity with non-annotated one is higher and annotates it.

Once the process is finished, we have labeled and unlabeled data that make up our emotion corpus annotated automatically.

3.3 Training supervised classifiers

In the second step of the bootstrapping technique, the annotated and the non-annotated sentences from the previous step are exploited to train a set of supervised classifiers. Concretely, a Support Vector Machines (SVM) with Sequential Minimal Optimization (SMO) [67] algorithm is applied where the sentences are represented as a vector of words weighted by their counts using Weka [68].

4 Evaluation

The objective of this research is to assess the viability of the use of IL bootstrapping technique to built emotion corpora reducing the cost and time-consuming. To achieve that, in this paper two evaluation, explained in Section 4.2, are carried out.

4.1 Datasets

Our approach is tested against two emotion corpora annotated at sentence level: (i) Aman corpus [9]; and (ii) Affective Text corpus [10].

These corpora are selected because of several reasons: (i) both corpora are manually annotated allowing us to compare automatic annotation to manual annotation; (ii) they are relevant to emotion detection task since they have been employed in many works to detect emotions [69], [70], [71]; and (iii) these corpora allow us to test our approach about corpora with different sources of information: news headlines and blog posts from Web. Thus, the usability and effectiveness of our approach can be checked.

Aman corpus. This dataset contains sentence-level annotation of 4,000 sentences from blog posts collected directly from Web. This resource was annotated manually with the six emotion categories proposed by Ekman and the emotion intensity (high, medium, or low).

Semeval-2007 Affective Text corpus. It contains sentence-level annotations of 1,250 short texts from news headlines, which were drawn from major newspapers such as New York Times, CNN and BBC News, as well as from the Google News. This corpus was annotated manually with Ekman's basic emotions and valence.

4.2 Evaluation Methodology

As we mentioned, the evaluation methodology is divided into two steps. On the one hand, an emotional model is built from the corpus annotated automatically to evaluate the usability of this corpus. On the other hand, the quality of automatic annotations is assessed through the measure of agreement between the corpus developed with our approach (automatic annotation) and the gold standard of Aman corpus and Affective Text corpus (manual annotation).

With regards to automatic emotion classification, a SMO multi-classifier is employed on Aman corpus because of it is annotated with one emotion. On Affective Text corpus, six binary classifiers SMO are applied since each sentence can be annotated with one or more emotions. For the evaluation, the versions of the corpora (Aman corpus and Affective Text corpus) annotated automatically with our approaches are performed with a 10-fold cross-validation. Specifically, precision, recall and F1-score are calculated in each model. This evaluation allows us to analyse the results obtained by machine learning algorithms when an emotion corpus automatically annotated is employed.

Concerning agreement evaluation, the annotation features of each corpus carried out in the original works, along with our evaluation are explained in the next subsections. This assessment is carried out because it indicates us how well our process annotates since the automatic annotations are directly compared to the gold standard of each corpus. If there is a disagreement between automatic and manual annotations, this indicates that it has been mistakes of the creation of the seed and thus there are incorrect associations between sentences and emotions.

4.2.1 Aman corpus

This corpus was manually developed by four annotators who received no training, though they were given samples of annotated sentences to illustrate the kind of annotations required. About the emotion categories, the Ekman's basic emotions were selected and two further categories were added: (i) mixed emotions and (ii) no emotion, resulting in eight categories to which a sentence could be assigned. As for annotation metric, Cohen's kappa [72] was employed.

Concerning our evaluation of agreement between the automatic annotation and the gold standard of Aman corpus, the Cohen's kappa measure is employed like in the original work.

5. http://www.natcorp.ox.ac.uk/
6. https://catalog.ldc.upenn.edu/LDC2012T21

4.2.2 Affective Text corpus

Concerning the emotion annotation task carried out on Affective Text corpus, they organized a manual annotation task constituted of six annotators who were instructed to select the appropriate emotions. The annotators assigned a value for each Ekman's basic emotion and a value for valence. Hence, each headline had associated a value for each emotion and a value for valence. About the inter-tagger agreement, it was conducted for each of the six emotions and for the valence annotations and was carried out using the Pearson correlation [73] measure.

In Task 14 of SemEval-2007, two evaluations were carried out: fine-grained and coarse-grained. The fine-grained evaluation was conducted using the Pearson correlation between the system and the gold standard scores. In the coarse-grained evaluation, each emotion of the gold standard was mapped to a 0/1 classification (0=[0,50), 1=[50,100]), and each valence was mapped to a -1/0/1 (-1=[-100, -50], 0=(-50,50), 1= [50,100]).

In our evaluation, the gold standard employed in coarse-grained evaluation is used to measure the agreement with our automatic annotations. To achieve that, Cohen's Kappa is employed because it is the most frequent metric used to compare the extent of consensus between annotators in classifying items.

4.3 Results

On the one hand, the results obtained by each classifier employing the gold standard corpora annotated manually with same algorithms, set of features and evaluation (cross-validation) are shown in Tables 4 and 5.

On the other hand, the results obtained by each classifier in all of our approaches of corpora annotated automatically are shown in the tables below. Tables 6 and 7 detail results obtained with all of the DSMs on Aman corpus and Tables 8 and 9 show the results on Affective Text corpus. Precision (P), recall (R) and F1-values (F1) are shown for each emotion employing the original approach and the enriched approaches.

TABLE 4
Precision, Recall and F1-values Obtained by the SMO Multi-Classifier on the gold standard of Aman Corpus

	Aman Corpus		
	P	R	F1
Anger	0.538	0.274	0.363
Disgust	0.714	0.32	0.442
Fear	0.672	0.357	0.466
Joy	0.720	0.513	0.599
Sadness	0.577	0.260	0.359
Surprise	0.553	0.226	0.321
Neutral	0.798	0.955	0.869
Macro Avg.	0.653	0.415	**0.488**

Regarding the comparison between automatic and manual annotations, Cohen's kappa values obtained by each one of our approaches when they are compared to the gold standard of both corpora are shown in Tables 10-11.

TABLE 5
Precision, Recall and F1-values Obtained by the Six Binary-Classifiers SMO on the gold standard of Affective Text corpus

	Affective Text Corpus		
	P	R	F1
Anger	0.946	0.962	0.953
Disgust	0.986	0.988	0.985
Fear	0.876	0.902	0.881
Joy	0.843	0.879	0.850
Sadness	0.904	0.913	0.897
Surprise	0.967	0.970	0.962
Macro Avg.	0.920	0.936	**0.921**

4.4 Discussion

4.4.1 Aman corpus

As for the emotion model evaluation, the macro average of F1-values obtained for all of the approaches are encouraging since most of them are near 40%, obtaining the best value of 41.2% in Gigaword W2V (CBOW). Although these values do not improve the results obtained with the original Aman corpus (48.8%), these results are interesting taking into account the corpus has been annotated automatically with great benefits in terms of cost and time.

About the agreement evaluation that allows us to evaluate the effectiveness of the first step of bootstrapping and the quality of annotations, the results show the agreement between automatic and manual annotation since most of the results are near 80% for each emotion except JOY emotion which obtains worse values. This may be due to the fact that the process of the creation of the seed introduces false JOY sentences and this error generates noisy to the second part of the bootstrapping process. As we mentioned previously, linguistic phenomena such as negation or irony have not been addressed in this approach and this may be the cause of this problem. Moreover, the presence of JOY words in the sentences is frequent since Emolex contains words like *child*, *found*, *clean*, etc. associated with JOY, thus this could be another cause. For this reason, the number of sentences annotated with JOY is higher than they should be and the agreement is worse. Nevertheless, these results are hopeful and demonstrate the viability of this technique and the possibility of creating an emotion corpus reducing the cost and time-consuming.

With regards to original and enriched approach in both evaluations, the comparative must be done per emotion since there are different situations. Emotions like ANGER, FEAR, and SADNESS obtain improvements in enriched approaches in F1-values. Respect to agreement values in these emotions, the improvements are showed in original approach though the values obtained by enriched ones are higher 80% in the majority of the emotions, thus the quality of the annotations is high. About DISGUST emotion, it is an emotion where the improvements in F1-value in enriched

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2017.2764470, IEEE Transactions on Affective Computing

TABLE 6
Precision, Recall and F1-values Obtained by the SMO Multi-Classifier on the Corpus Developed Applying LSA and ukWak W2V (CBOW) Models as Semantic Metric in the Extension of the Seed on Aman Corpus.

	LSA model (Aman corpus)									ukWak W2V (CBOW) (Aman corpus)								
	Original approach			Enriched approach WN			Enriched approach Oxford			Original approach			Enriched approach WN			Enriched approach Oxford		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Anger	0.198	0.137	0.162	0.444	0.348	**0.391**	0.338	0.330	0.334	0.184	0.152	0.167	0.413	0.330	**0.367**	0.360	0.356	0.358
Disgust	0.250	0.068	0.107	0.308	0.178	**0.225**	0.353	0.120	0.179	0.121	0.047	0.067	0.350	0.286	**0.315**	0.200	0.098	0.132
Fear	0.401	0.236	0.297	0.392	0.303	**0.342**	0.412	0.251	0.312	0.289	0.179	0.221	0.409	0.282	**0.334**	0.336	0.219	0.265
Joy	0.574	0.571	0.572	0.677	0.702	**0.689**	0.565	0.604	0.584	0.507	0.586	0.544	0.680	0.796	**0.733**	0.520	0.600	0.557
Sadness	0.247	0.107	0.149	0.467	0.269	0.341	0.591	0.462	**0.519**	0.307	0.226	0.260	0.406	0.241	0.303	0.552	0.586	**0.568**
Surprise	0.459	0.224	**0.301**	0.366	0.152	0.214	0.359	0.192	0.250	0.345	0.185	0.241	0.294	0.103	0.153	0.376	0.229	**0.285**
Neutral	0.706	0.846	**0.770**	0.559	0.676	0.612	0.551	0.668	0.604	0.608	0.702	**0.652**	0.587	0.573	0.580	0.596	0.554	0.574
Macro Avg.	0.405	0.313	0.337	0.459	0.375	**0.402**	0.453	0.375	0.397	0.337	0.297	0.307	0.448	0.373	**0.398**	0.420	0.377	0.391

TABLE 7
Precision, Recall and F1-values Obtained by the SMO Multi-Classifier on the Corpus Developed Applying Gigaword W2V (CBOW & SKIP) Models as Semantic Metric in the Extension of the Seed on Aman Corpus.

	Gigaword W2V CBOW) (Aman corpus)									Gigaword W2V (SKIP) (Aman corpus)								
	Original approach			Enriched approach WN			Enriched approach Oxford			Original approach			Enriched approach WN			Enriched approach Oxford		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Anger	0.113	0.074	0.089	0.541	0.400	**0.460**	0.399	0.385	0.392	0.139	0.094	0.112	0.465	0.354	**0.402**	0.383	0.385	0.384
Disgust	0.250	0.052	0.086	0.262	0.129	**0.173**	0.235	0.080	0.119	0.176	0.037	0.061	0.365	0.256	**0.301**	0.160	0.073	0.100
Fear	0.419	0.233	0.300	0.387	0.287	**0.329**	0.300	0.172	0.219	0.336	0.223	0.268	0.388	0.286	**0.329**	0.257	0.167	0.203
Joy	0.554	0.423	0.480	0.674	0.706	**0.690**	0.550	0.556	0.553	0.528	0.597	0.560	0.688	0.748	**0.717**	0.557	0.609	0.582
Sadness	0.305	0.105	0.157	0.496	0.298	0.372	0.554	0.459	**0.502**	0.273	0.143	0.188	0.435	0.213	0.286	0.544	0.417	**0.472**
Surprise	0.407	0.222	**0.287**	0.406	0.160	0.230	0.338	0.150	0.208	0.353	0.156	0.217	0.250	0.092	0.134	0.359	0.168	**0.229**
Neutral	0.719	0.876	**0.790**	0.591	0.679	0.632	0.540	0.648	0.589	0.629	0.751	**0.685**	0.538	0.636	0.583	0.485	0.595	0.534
Macro Avg.	0.395	0.284	0.313	0.480	0.380	**0.412**	0.417	0.350	0.369	0.348	0.286	0.299	0.447	0.369	**0.393**	0.392	0.345	0.358

TABLE 8
Precision, Recall and F1-values Obtained by the SMO Six Binary-Classifiers on the Corpus Developed Applying LSA and ukWak W2V (CBOW) Models as Semantic Metric in the Extension of the Seed on Affective Text Corpus.

	LSA model (Affective Text corpus)									ukWak W2V (CBOW) (Affective Text corpus)								
	Original approach			Enriched approach WN			Enriched approach Oxford			Original approach			Enriched approach WN			Enriched approach Oxford		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Anger	0.722	0.737	0.720	0.705	0.705	0.701	0.769	0.763	**0.760**	0.754	0.765	0.746	0.722	0.720	0.716	0.772	0.766	**0.763**
Disgust	0.870	0.883	**0.855**	0.830	0.831	0.808	0.803	0.824	0.797	0.875	0.884	**0.853**	0.827	0.829	0.806	0.799	0.822	0.792
Fear	0.764	0.762	**0.754**	0.672	0.671	0.671	0.726	0.724	0.722	0.768	0.768	**0.761**	0.697	0.696	0.696	0.753	0.749	0.747
Joy	0.815	0.826	**0.803**	0.769	0.771	0.759	0.717	0.722	0.712	0.823	0.835	**0.813**	0.769	0.772	0.758	0.745	0.749	0.739
Sadness	0.778	0.789	**0.774**	0.730	0.737	0.727	0.755	0.757	0.748	0.805	0.812	**0.793**	0.738	0.744	0.732	0.756	0.756	0.745
Surprise	0.851	0.852	**0.819**	0.822	0.823	0.801	0.799	0.811	0.790	0.853	0.857	**0.826**	0.817	0.823	0.805	0.802	0.813	0.791
Macro Avg.	0.800	0.808	**0.788**	0.755	0.756	0.745	0.762	0.767	0.755	0.813	0.820	**0.799**	0.762	0.764	0.752	0.771	0.776	0.763

TABLE 9
Precision, Recall and F1-values Obtained by the SMO Six Binary-Classifiers on the Corpus Developed Applying Gigaword W2V (CBOW & SKIP) Models as Semantic Metric in the Extension of the Seed on Affective Text Corpus.

	Gigaword W2V (CBOW) (Affective Text corpus)									Gigaword W2V (SKIP) (Affective Text corpus)								
	Original approach			Enriched approach WN			Enriched approach Oxford			Original approach			Enriched approach WN			Enriched approach Oxford		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Anger	0.765	0.775	0.755	0.714	0.713	0.708	0.777	0.772	**0.769**	0.738	0.748	0.730	0.697	0.696	0.693	0.767	0.761	**0.759**
Disgust	0.884	0.889	**0.862**	0.832	0.834	0.813	0.804	0.825	0.797	0.856	0.868	**0.834**	0.818	0.822	0.801	0.792	0.815	0.784
Fear	0.782	0.782	**0.775**	0.692	0.691	0.690	0.726	0.724	0.722	0.770	0.768	**0.762**	0.694	0.693	0.693	0.719	0.716	0.715
Joy	0.824	0.837	**0.815**	0.772	0.774	0.761	0.724	0.729	0.718	0.813	0.826	**0.804**	0.771	0.771	0.757	0.728	0.732	0.722
Sadness	0.805	0.812	**0.795**	0.729	0.737	0.725	0.757	0.758	0.748	0.788	0.796	**0.780**	0.723	0.729	0.717	0.737	0.740	0.732
Surprise	0.868	0.868	**0.838**	0.812	0.819	0.799	0.814	0.822	0.799	0.854	0.857	**0.824**	0.824	0.828	0.813	0.804	0.815	0.792
Macro Avg.	0.821	0.827	**0.807**	0.759	0.761	0.749	0.767	0.772	0.759	0.803	0.811	**0.789**	0.755	0.757	0.746	0.758	0.763	0.751

TABLE 10
Cohen's kappa values Obtained by All Approaches (the Original Approach and the Enriched Approaches) in the Comparison of their Annotations to the Gold of Aman Corpus.

	Cohen's kappa values (Aman corpus)											
	LSA			ukWak W2V (CBOW)			Gigaword W2V (CBOW)			Gigaword W2V (SKIP)		
	Original	Enriched WN	Enriched Oxford	Original	Enriched WN	Enriched Oxford	Original	Enriched WN	Enriched Oxford	Original	Enriched WN	Enriched Oxford
Anger	**0.9368**	0.9051	0.8882	**0.9193**	0.9004	0.8713	**0.9430**	0.9089	0.8875	**0.9328**	0.9044	0.8675
Disgust	0.9495	0.9417	**0.9537**	0.9452	0.9392	**0.9529**	0.9507	0.9430	**0.9527**	0.9460	0.9412	**0.9514**
Fear	0.9226	0.8919	**0.9323**	0.9315	0.9099	**0.9328**	**0.9380**	0.9136	0.9343	0.9106	0.9009	**0.9223**
Joy	**0.7719**	0.6041	0.7241	0.6987	0.5359	**0.7219**	**0.8053**	0.6414	0.7443	**0.7281**	0.5752	0.6942
Sadness	**0.9285**	0.9193	0.8033	0.8750	**0.9119**	0.7425	**0.9340**	0.9173	0.8396	**0.9131**	0.9066	0.8230
Surprise	0.9186	**0.9512**	0.9345	0.9014	**0.9522**	0.9338	0.9368	**0.9557**	0.9325	0.9146	**0.9509**	0.9295

TABLE 11
Cohen's kappa values Obtained by All Approaches (the Original Approach and the Enriched Approaches) in the Comparison of their Annotations to the Gold of Affective Text Corpus.

	Cohen's kappa values (Affective Text corpus)											
	LSA			ukWak W2V (CBOW)			Gigaword W2V (CBOW)			Gigaword W2V (SKIP)		
	Original	Enriched WN	Enriched Oxford	Original	Enriched WN	Enriched Oxford	Original	Enriched WN	Enriched Oxford	Original	Enriched WN	Enriched Oxford
Anger	**0.6896**	0.5544	0.5312	**0.6976**	0.5600	0.5448	**0.7024**	0.5632	0.5496	**0.6800**	0.5480	0.5344
Disgust	**0.8552**	0.7422	0.7888	**0.8560**	0.7448	0.7904	**0.8568**	0.7456	0.7896	**0.8416**	0.7336	0.7832
Fear	**0.6576**	0.5476	0.5792	**0.6696**	0.5504	0.5832	**0.6712**	0.5520	0.5848	**0.6520**	0.5464	0.5752
Joy	**0.7704**	0.6902	0.6456	**0.7752**	0.6928	0.6488	**0.7792**	0.6928	0.6480	**0.7688**	0.6856	0.6392
Sadness	**0.7576**	0.6693	0.6576	**0.7712**	0.6752	0.6616	**0.7720**	0.6768	0.6624	**0.7544**	0.6664	0.6520
Surprise	**0.7856**	0.7182	0.7328	**0.7912**	0.7208	0.7352	**0.7976**	0.7224	0.7392	**0.7904**	0.7168	0.7352

approaches is also shown in agreement values. Respect to SURPRISE emotion, it is one of the emotions more complicated to classify but the enriched approach Oxford of ukWak W2V(CBOW) and Gigaword W2V (SKIP) models improve the F1 results and moreover, the agreement values in these approaches remain higher than 90%. About JOY emotion, as we mentioned before, there is an excess of sentences annotated with this emotion, then the agreement is worse in enriched approaches but F1-values are better since the algorithm tend to classify by the most frequent class. Considering all emotion, the results demonstrate the benefits of enriched approaches and the usability of extending the seed in Aman corpus since the improvements in F1-values obtained by enriched approaches also gets high agreement values. However, the use of Wordnet or Oxford synonyms should be analyzed in depth since the results vary depending on each emotion.

Finally, regarding the DSMs employed, the F1-values, as well as the agreement, do not obtain significant results for considering that one model is better than the rest.

4.4.2 Affective Text corpus

Regarding the F1-values obtained on Affective Text corpus, the results are really interesting because most of them are near 80%, obtaining the best value of 80.7% in Gigaword W2V (CBOW). As in Aman corpus, these results do not improve the value obtained with the original Affective Text corpus (92.1%), however, they are considered encouraging since the corpus employed have been automatically developed.

About the agreement values obtained, the values are between 65% and 85%, values near to get a good reliability and hence they are encouraging results. In this corpus, the worst values are obtained in FEAR emotion. This values can be due to the fact that the fear words saved on Emolex are not representative of the vocabulary employed in headlines, since these results are obtained even though the recall of fear word in Emolex is the highest respect to other emotions.

With reference to the comparative between original and enriched approaches in both evaluations, the situation is different to Aman corpus. In this case, the improvements F1-value are only observed in enriched approach Oxford for ANGER emotion regardless the model employed. However, these improvements are not reflected in agreement evaluation. Therefore, in this case, the extension is not recommended, since the original approach results are encouraging in both evaluations: classification and agreement. In the rest of the emotions, the best results are obtained by the original approach in the classification and the agreement. With all this in mind, we can conclude that the use of the resources to extend Emolex and the election of these resources would depend on the genre of text to annotate. Despite this, the results of the original approach demonstrate the usability of the technique for Affective Text corpus.

Finally, as on Aman corpus, the results obtained by the different DSMs do not report significant results for concluding which is the best model.

5 Conclusion

As presented in the introductory section of this paper, the rationale beyond our research is the need to tackle the annotation task of emotions automatically due to the cost and time associated with the manual annotation process.

The paper presents a bootstrapping technique based on IL for automatic annotations with two main steps: 1) an initial similarity-based categorization where a set of seed sentences is created and this seed is extended by the semantic similarity between sentences; 2) train a supervised classifier on the initially categorised set.

According to the evaluation performed, the appropriateness and reliability of our approach are demonstrated. Hence, our main conclusions are as follows: 1) the viability of IL bootstrapping process as technique to automatically label emotion corpora reducing the cost and time-consuming is demonstrated, since both evaluations carried out obtains encouraging results taking into account the automatic process used to create corpora; 2) the results do not allow us to conclude which DSMs is better for extending the seed since all of them obtains similar results. Thus, we can conclude that the step 1.2 of the process is independent of the DSMs employed, providing flexibility to our proposal; 3) about the use of NRC Word-Emotion Association Lexicon, the results have been satisfactory taking into account it is a general domain resource and it has been applied in two different genres: headlines and blog posts; and 4) the improvement of enriched approaches has been demonstrated for several emotions in Aman corpus, thus the process of extension could be beneficial depending on the genre of text analyzed. Hence, the usability of these approaches will be analysed in depth in future works.

Our future research will deal with 1) exploiting larger emotion lexicons than Emolex for creating the seed such as EmoSenticNet or DepecheMood since the core of IL is the first step; 2) analysing in depth the use of DSMs and testing the approach with domain specific embedding; 3) testing the technique in more corpora with another group of emotions, since the adaptation of the process is really simple provided that the emotion lexicon was annotated with the desired emotions; and 4) analysing the usability of other resources to extend the seed in the enriched approaches.

ACKNOWLEDGMENTS

This research has been supported by the FPI grant (BES-2013-065950) and the research stay grants (EEBB-I-15-10108 and EEBB-I-16-11174) from the Spanish Ministry of Science and Innovation. It has also funded by the Spanish Government (DIGITY ref. TIN2015-65136-C02-2-R and RESCATA ref. TIN2015-65100-R), the Valencian Government (grant no. PROMETEOII/ 2014/001), the University of Alicante (ref. GRE16-01) and BBVA Foundation (Análisis de Sentimientos Aplicado a la Prevención del Suicidio en las Redes Sociales (ASAP) project).

REFERENCES

[1] P. Rodríguez, A. Ortigosa, and R. M. Carro, "Extracting Emotions from Texts in E-Learning Environments." in *Complex, Intelligent and Software Intensive Systems (CISIS)*, L. Barolli, F. Xhafa, S. Vitabile, and M. Uehara, Eds. IEEE Computer Society, 2012, pp. 887–892.

[2] C. S. Montero and J. Suhonen, "Emotion analysis meets learning analytics: online learner profiling beyond numerical data," in *Proceedings of the 14th Koli Calling International Conference on Computing Education Research*, 2014, pp. 165–169.

[3] B. Desmet and V. Hoste, "Emotion Detection in Suicide Notes," *Expert Syst. Appl.*, vol. 40, no. 16, pp. 6351–6358, 2013.

[4] F. Vaassen, "Measuring emotion," Ph.D. dissertation, Universiteit Antwerpen, 2014.

[5] C. Cherry, S. M. Mohammad, and B. De Bruijn, "Binary Classifiers and Latent Sequence Models for Emotion Detection in Suicide Notes," *Biomedical informatics insights*, vol. 5(Suppl 1), pp. 147–154, 2012.

[6] M. Dadvar, D. Trieschnigg, R. Ordelman, and F. de Jong, "Improving cyberbullying detection with user context," in *Advances in Information Retrieval*, 2013, pp. 693–696.

[7] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, R. E. Lucas, M. Agrawal, G. J. Park, S. K. Lakshmikanth, S. Jha, M. E. P. Seligman, and L. Ungar, "Characterizing Geographic Variation in Well-Being Using Tweets," in *Proceedings of the International AAAI Conference on Weblogs and Social Media*, 2013.

[8] S. M. Mohammad, "Sentiment Analysis: Detecting Valence, Emotions, and Other Affectual States from Text," in *Emotion Measurement*, H. Meiselman, Ed. Elsevier, 2016.

[9] S. Aman and S. Szpakowicz, "Identifying Expressions of Emotion in Text," in *Text, Speech and Dialogue*, 2007, pp. 196–205.

[10] C. Strapparava and R. Mihalcea, "Semeval-2007 task 14: Affective text," in *Proceedings of the 4th International Workshop on Semantic Evaluations*, 2007, pp. 70–74.

[11] E. Riloff, J. Wiebe, and T. Wilson, "Learning subjective nouns using extraction pattern bootstrapping," in *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, ser. CONLL '03. Stroudsburg, PA, USA: Association for Computational Linguistics, 2003, pp. 25–32.

[12] C. Banea, R. Mihalcea, and J. Wiebe, "A bootstrapping method for building subjectivity lexicons for languages with scarce resources," in *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC-08)*. Marrakech, Morocco: European Language Resources Association (ELRA), May 2008, aCL Anthology Identifier: L08-1086.

[13] M. Baroni and S. Bernardini, "Bootcat: Bootstrapping corpora and terms from the web." in *Proceedings of the Fourth Conference on Language Resources and Evaluation (LREC-04)*. European Language Resources Association, 2004.

[14] A. Gliozzo, C. Strapparava, and I. D. O. Dagan, "Improving Text Categorization Bootstrapping via Unsupervised Learning," *ACM Transactions on Speech and Language Processing*, vol. 6, no. 1, 2009.

[15] D. Yarowsky, "Unsupervised Word Sense Disambiguation Rivaling Supervised Methods," in *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics*, ser. ACL '95. Stroudsburg, PA, USA: Association for Computational Linguistics, 1995, pp. 189–196.

[16] M. Collins and Y. Singer, "Unsupervised Models for Named Entity Classification," in *In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 1999, pp. 100–110.

[17] E. Riloff and R. Jones, "Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping," *Evaluation*, no. 1032, pp. 474–479, 1999.

[18] S. Chowdhury and W. Chowdhury, "Performing Sentiment Analysis in Bangla Microblog Posts," in *International Conference on Informatics, Electronics & Vision (ICIEV)*. IEEE, 2014.

[19] H. Schütze, "Automatic Word Sense Discrimination," *Computational Linguistics*, vol. 24, no. 1, pp. 97–123, 1998.

[20] C.-Y. Lin and E. Hovy, "Automatic evaluation of summaries using N-gram co-occurrence statistics," in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - NAACL '03*, vol. 2003, no. June, 2003, pp. 71–78.

[21] R. Barzilay and M. Lapata, "Modeling Local Coherence: An Entity-Based Approach," *Computational Linguistics*, vol. 34, no. 1, pp. 1–34, 2008.

[22] T. Kenter and M. de Rijke, "Short Text Similarity with Word Embeddings," in *International on Conference on Information and Knowledge Management (CIKM'15)*, 2015, pp. 1411–1420.

[23] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," *CoRR*, pp. 1–12, 2013.

[24] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global Vectors for Word Representation," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

[25] S. Deerwester, S. T. Dumais, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American society for information science*, vol. 41, no. 6, pp. 391–407, 1990.

[26] J. Villalón, P. Kearney, R. A. Calvo, and P. Reimann, "Glosser: Enhanced feedback for student writing tasks," in *Proceedings - The 8th IEEE International Conference on Advanced Learning Technologies, ICALT 2008*, no. 1, 2008, pp. 454–458.

[27] L. Li, M. Wang, L. Zhang, and H. Wang, "Learning Semantic Similarity for Multi-label Text Categorization," in *Chinese Lexical Semantics (CLSW 2014)*, vol. 8922, 2014, pp. 260–269.

[28] C. Predoiu, M. Dascalu, and S. Trausan-Matu, "Trust and user profiling for refining the prediction of reader's emotional state induced by news articles," in *RoEduNet Conference 13th Edition: Networking in Education and Research Joint Event RENAM 8th Conference, 2014*, 2014.

[29] A. García-Pablos, M. Cuadros, and G. Rigau, "Unsupervised word polarity tagging by exploiting continuous word representations," *Procesamiento de Lenguaje Natural*, vol. 55, pp. 127–134, 2015.

[30] M. Marelli, L. Bentivogli, M. Baroni, R. Bernardi, S. Menini, and R. Zamparelli, "SemEval-2014 Task 1: Evaluation of Compositional Distributional Semantic Models on Full Sentences through Semantic Relatedness and Textual Entailment," in *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, no. 1, 2014, pp. 1–8.

[31] C. Banea, D. Chen, R. Mihalcea, C. Cardie, and J. Wiebe, "SimCompass: Using Deep Learning Word Embeddings to Assess Cross-level Similarity," in *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, no. SemEval, 2014, pp. 560–565.

[32] P. Ekman, "Basic emotions," in *Handbook of cognition and emotion*, 1999, pp. 45–60.

[33] R. Plutchik, "A general psychoevolutionary theory of emotion," in *Theories of Emotion*, 1980, pp. 3–33.

[34] J. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39(6), pp. 1161–1178, 1980.

[35] A. Mehrabian, "Pleasure-arousal-dominance: A general framework for describing and measuring individual," *Current Psychology*, vol. 15(4), pp. 505–525, 1996.

[36] M. M. Bradley and P. P. J. Lang, "Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings," *Psychology*, vol. Technical, no. C-1, p. 0, 1999.

[37] C. Whissell, M. Fournier, R. Pelland, D. Weir, and K. Makarec, "a Dictionary of Affect in Language: Iv. Reliability, Validity, and Applications," *Perceptual and Motor Skills*, vol. 62, no. 3, pp. 875–888, 1986.

[38] C. Strapparava and A. Valitutti, "WordNet-Affect: an affective extension of WordNet," *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pp. 1083–1086, 2004.

[39] S. M. Mohammad and P. D. Turney, "Crowdsourcing a word-emotion association lexicon," in *Computational Intelligence*, vol. 29, no. 3, 2013, pp. 436–465.

[40] S. M. Mohammad and S. Kiritchenko, "Using hashtags to capture fine emotion categories from tweets," *Computational Intelligence*, vol. 31, no. 2, pp. 301–326, 2015.

[41] S. Poria, A. Gelbukh, A. Hussain, N. Howard, D. Das, and S. Bandyopadhyay, "Enhanced senticnet with affective labels for concept-based opinion mining," *IEEE Intelligent Systems*, vol. 28, no. 2, pp. 2–9, 2013.

[42] E. Cambria, S. Poria, and R. Bajpai, "SenticNet 4 : A Semantic Resource for Sentiment Analysis Based on Conceptual Primitives," *Sentic.Net*, 2016.

[43] J. Staiano and M. Guerini, "Depechemood: a lexicon for emotion analysis from crowd-annotated news," *CoRR*, vol. abs/1405.1605, 2014.

[44] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*. MIT Press, 2012.

[45] M. Nissim and V. Patti, "Chapter 3 - semantic aspects in sentiment analysis," in *Sentiment Analysis in Social Networks*, F. A. Pozzi, E. Fersini, E. Messina, and B. Liu, Eds. Boston: Morgan Kaufmann, 2017, pp. 31 – 48.

[46] C. O. Alm, D. Roth, and R. Sproat, "Emotions from text: Machine learning for text-based emotion prediction," in *Proceedings of the conference on HLT-EMNLP*, 2005, pp. 579–586.

[47] K. R. Scherer, "What are emotions? And how can they be measured?" *Social Science Information*, vol. 44, no. 4, pp. 695–729, 2005.

[48] A. Neviarouskaya, H. Prendinger, and M. Ishizuka, "Recognition of Affect, Judgment, and Appreciation in Text," in *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, 2010, pp. 806–814.

[49] A. Neviarouskaya, H. Prendinger, M. Ishizuka, A. Neviarouskaya, and H. Prendinger, "Affect Analysis Model: novel rule-based approach to affect sensing from text," *Natural Language Engineering*, vol. 17, no. 01, pp. 95–135, 2011.

[50] E. Boldrini and P. Martínez-Barco, "EMOTIBLOG: A model to Learn Subjetive Information Detection in the New Textual Genres of the Web 2.0-Multilingual and Multi-Genre Approach-," Ph.D. dissertation, University of Alicante, 2012.

[51] J. S. Y. Liew, H. R. Turtle, and E. D. Liddy, "EmoTweet-28: A Fine-Grained Emotion Corpus for Sentiment Analysis," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 2016.

[52] S. M. Mohammad, X. Zhu, S. Kiritchenko, and J. Martin, "Sentiment, emotion, purpose, and style in electoral tweets," *Inf. Process. Manage.*, vol. 51, no. 4, pp. 480–499, Jul. 2015.

[53] D. Nozza, E. Fersini, and E. Messina, "A Multi-View Sentiment Corpus," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017)*, 2017, pp. 272–280.

[54] S. Mohammad, "#emotional tweets," in **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*. Montréal, Canada: Association for Computational Linguistics, 7-8 June 2012, pp. 246–255.

[55] M. D. Choudhury, M. Gamon, and S. Counts, "Happy, Nervous or Surprised? Classification of Human Affective States in Social Media," in *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media*, 2012.

[56] W. Wang, L. Chen, K. Thirunarayan, and A. P. Sheth, "Harnessing Twitter "Big Data" for Automatic Emotion Identification," in *International Confernece on Social Computing (SocialCom)*, 2012.

[57] C. E. Izard, *The face of emotion*, New York: Appleton-Century-Crofts., Ed., 1971.

[58] C. Pool and M. Nissim, "Distant supervision for emotion detection using facebook reactions," *CoRR*, vol. abs/1611.02988, 2016.

[59] L. Canales, C. Strapparava, E. Boldrini, and P. Martínez-Barco, "A Bootstrapping Technique to Annotate Emotional Corpora Automatically," in *Proceedings of the LREC 2016 Workshop Emotion and Sentiment Analysis*, 2016.

[60] ——, "Exploiting a Bootstrapping Approach for Automatic Annotation of Emotions in Texts," in *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 2016, pp. 726–734.

[61] S. M. Kim, "Recognising Emotions and Sentiments in Text," Ph.D. dissertation, University of Sydney, 2011.

[62] G. a. Miller, "WordNet: a lexical database for English," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[63] V. Reutter and D. Grabarek, "Oxford American Writer's Thesaurus." *School Library Journal*, vol. 51, no. 4, pp. 82–84, 2005.

[64] C. D. Manning, J. Bauer, J. Finkel, S. J. Bethard, M. Surdeanu, and D. McClosky, "The Stanford CoreNLP Natural Language Processing Toolkit," *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55–60, 2014.

[65] A. Gliozzo and C. Strapparava, *Semantic Domains in Computational Linguistics*. Springer-Verlag Berlin Heidelberg, 2009.

[66] G. Dinu and M. Baroni, "Improving zero-shot learning by mitigating the hubness problem," *CoRR*, vol. abs/1412.6, 2014.

[67] J. Platt, "Using Analytic QP and Sparseness to Speed Training of Support Vector Machines," in *Proc. Advances in Neural Information Processing Systems*, 1999, pp. 557–563.

[68] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA Data Mining Software: An Update," *SIGKDD Explorations*, vol. 11, no. 1, pp. 10–18, 2009.

[69] F. Keshtkar and D. Inkpen, "A Corpus-based Method for Extracting Paraphrases of Emotion Terms," in *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, ser. CAAGET '10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 35–44.

[70] S. Chaffar and D. Inkpen, "Using a Heterogeneous Dataset for Emotion Analysis in Text," in *Proceedings of the 24th Canadian Conference on Advances in Artificial Intelligence*, ser. Canadian AI'11. Berlin, Heidelberg: Springer-Verlag, 2011, pp. 62–67.

[71] S. Mohammad, "Portable Features for Classifying Emotional Text," in *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Montréal, Canada: Association for Computational Linguistics, 2012, pp. 587–591.

[72] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, pp. 37–46, 1960.
[73] N. L. L. Anthony J. Onwuegbuzie Larry Daniel, "Pearson Product-Moment Correlation Coefficient," *Encyclopedia of Measurement and Statistics*, vol. 2, no. 1, pp. 751–756, 2007.

Lea Canales received the MSc degree in computer technologies from University of Alicante, Spain, in 2013. Since then, she is working on Natural Language Processing. Currently, she is working as a PhD researcher in GPLSI Research Group of the University of Alicante, Spain. Her interests include the identification of affective states from text. She is especially interested in the development of emotion resources that allow to improve the automatic emotion detection from Web 2.0.

Carlo Strapparava is a senior researcher at Fondazione Bruno Kessler-Istituto per la ricerca scientifica e Tecnologica (FBK-irst) in the Human Language Technology - NLP unit. His research interests include AI, natural language processing, intelligent interfaces, cognitive science, knowledge-based systems, user models, adaptive hypermedia, lexical knowledge bases, word sense disambiguation, and computational humor. In these fields, he published more than 200 scientific reviewed publications.

Ester Boldrini has a PhD in Computational Linguistics and a European Master on English and Spanish for Institutions, Enterprises and Business from the University of Alicante in addition to the degree in Linguistic Mediation for Institution, Enterprises and Business from the University of Tuscia, Italy. In addition of being member of the evaluation committee of relevant international conference, she is author of many papers published in high rankings peer-reviewed journals results of her research work mainly focused on Sentiment Analysis and the creation of linguistic resources to improve its automatic detection. She is Deputy Director of OGPI. With more than eight years of experience, she is specialist in international cooperation in the field of Higher Education she has wide experience in drafting proposals for EU programmes for both research and international cooperation programmes such as FP7, H202020, but also EuropeAid, Tempus, Erasmus+ on different topics related to Higher Education and others. Coordinator of numerous initiatives from public and private donors worldwide (Africa, Latin America, Europe, Eastern Europe), especially in topics such as the Bologna reform of Higher Education, Quality Assurance & Accreditation, Curriculum Development, Research Management, or Staff and Students mobility.

Patricio Martínez-Barco obtained his PhD in Computer Science from the University of Alicante (2001). He is working since 1995 in the Department of Software and Computing Science (Language Processing and Information Systems research Group - GPLSI) at this University as Associate Professor, becoming Head of this department between 2009 and 2013. His research interests are focused on Computational Linguistics and Natural Language Processing. His last projects are related to Language Generation, Text and Opinion Mining, Information Extraction and Information Retrieval. He was the General Chair of the ESTAL04 (Alicante), SEPLN04 (Barcelona), and SEPLN15 (Alicante), and co-organized several workshops and conferences related to these topics. He has edited several books, and contributed with more than 80 papers to journals and conferences. Currently, he is Vice-President of the Spanish Society for Natural Language Processing (SEPLN).