# EmoLabel: Semi-Automatic Methodology for Emotion Annotation of Social Media Text

Lea Canales, Walter Daelemans, Ester Boldrini, and Patricio Martínez-Barco

**Abstract**—The exponential growth of the amount of subjective information on the Web 2.0. has caused an increasing interest from researchers willing to develop methods to extract emotion data from these new sources. One of the most important challenges in textual emotion detection is the gathering of data with emotion labels because of the subjectivity of assigning these labels. Basing on this rationale, the main objective of our research is to contribute to the resolution of this important challenge. This is tackled by proposing EmoLabel: a semi-automatic methodology based on pre-annotation, which consists of two main phases: (1) an automatic process to pre-annotate the unlabelled English sentences; and (2) a manual process of refinement where human annotators determine which is the dominant emotion. Our objective is to assess the influence of this automatic pre-annotation method on manual emotion annotation from two points of view: agreement and time needed for annotation. The evaluation performed demonstrates the benefits of pre-annotation processes since the results on annotation time show a gain of near 20% when the pre-annotation process is applied (Pre-ML) without reducing annotator performance. Moreover, the benefits of pre-annotation are higher in those contributors whose performance is low (inaccurate annotators).

**Index Terms**—Natural Language Processing, Sentiment Analysis, Textual Emotion Recognition, Corpora Annotation, Social Media Text

## 1 Introduction

Over the past few years, the social media phenomenon has expanded throughout the world and quickly attracted billions of users [1]. As a consequence, there has been an exponential growth in the amount of subjective information on the Web 2.0 due to the massive use of these social media services by users.

Parallel to the growth of this new textual genre, there has been an increasing interest from researchers to develop methods to extract data from the subjective information available in these new sources because of its high potential benefits for business, society, politics or education. Consequently, this phenomenon has led to the need for developing innovative Natural Language Processing (NLP) tools, resources and methods, able to properly analyze and manage such data.

The NLP task that deals with the treatment of subjective data is called Sentiment Analysis (SA). Within this field, it is possible to differentiate between two tasks: Opinion Mining (OM), which can be defined as the task that automatically detects opinion expressed in texts and classifies it depending on its polarity (positive, negative or neutral); and Emotion Recognition (ER) which is a task more specific than opinion analysis that looks at fine-grained types of emotion such as emotional categories (ANGER, DISGUST, FEAR, etc.) or emotional dimensions (valence, arousal, dominance, etc.).

The increasing interest in ER from the research community is due to the fact that it has the potential of bringing substantial benefits to different sectors: examples of this can be for instance suicide prevention [2], [3], identifying cases of cyber-bullying [4], or contributing towards the improvement of student motivation and performance [5].

There are different techniques applied by NLP researchers to tackle the textual ER task, including the use of Machine Learning (ML), rule-based methods and lexical approaches [6], [7]. However, the majority of such proposals has been performed with machine learning algorithms mainly due to their scalability, learning capacity and fast development. In particular, the common scenario in textual ER is the use of supervised learning since these algorithms lead to better results than the rest of alternatives.

The success of the predictions made by a supervised model directly depend on the quality and the size of our training data. Hence, the training dataset employed is crucial to building accurate emotion detection systems. Moreover, these requirements of quality and size of training data is even more important in the new discipline called Deep Learning (DL) [8].

The creation of a labelled corpus for textual ER is not trivial, since detecting emotion in text can be difficult even for humans, because everyone's personal context can influence emotion interpretation. Most relevant research carried out so far has shown the difficulties related to this task, such as obtaining a good Inter-Annotator Agreement (IAA) or the time required for its development. As a consequence, data gathering with emotion content has become one of the most biggest challenges tasks in ER.

In order to lessen and counteract the challenge of emotion annotation, this research proposes EmoLabel: a semi-automatic methodology for textual ER with the aim of providing large-scale annotation of English emotion corpora

- *L. Canales is with the Department of Software and Computing System, University of Alicante, Alicante, Spain. E-mail: lcanales@dlsi.ua.es.*
- *W. Daelemans is with CLiPS Research Center, University of Antwerp, Antwerpen, Belgium. E-mail: walter.daelemans@uantwerpen.be.*
- *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.*

in any genre and with robust standards of reliability.

EmoLabel consists of two main phases: (1) an automatic process of pre-annotation of the unlabelled sentences with a reduced number of emotional categories; and (2) a manual refinement process where human annotators determine which is the dominant emotion in the pre-defined set of possibilities.

As the number of coding categories can influence reliability estimation of the resultant corpus, this research proposes to automatically pre-annotate those emotion categories most related to each sentence. Hence, our hypothesis is that suggesting a reduced number of categories could help human annotators to decide which is the dominant emotion in the second phase of EmoLabel and thus improve its reliability.

With the aim of evaluating the usability of two processes in the second phase, an *unsupervised pre-annotation* process based on Distributional Semantic Models (DSM's) and a *supervised pre-annotation* method based on ML have been carried out. Regarding the manual annotation task, different experimental setups have been designed to evaluate the impact this pre-annotation has on the quality of the resulting corpus and on the time needed for annotation.

Concerning the results, they demonstrate the benefits of the pre-annotation processes in emotion labeling since the results on annotation time show a gain of near 20% when the pre-annotation process is applied (Pre-ML) compared to No-pre without reducing the IAA or annotator performance. Moreover, the benefits of pre-annotation are higher in those contributors whose performance is low (inaccurate annotators).

The rest of the paper is organized as follows. Section 2 presents related work. After this, the methodology proposed is described in detail in Sections 3 and 4. Then, in Section 5 we explain the evaluation methodology, the results obtained and a discussion about these results. Finally, Section 6 details our conclusions and future work.

## 2 Related Work

ER is part of the broader area of Affective Computing (AC) which aims to enable computers to recognize and express emotions [9]. Within this discipline, affect detection systems can be classified in terms of individual modalities or channels according to the source of information employed to tackle it (e.g. text [10], [11], speech [12], [13], or videos [14], [15]). The challenges and difficulties associated with emotion annotation differ depending on these modalities. Concretely, this work is framed within textual ER, a subfield of NLP that studies and treats subjectivity language focusing on its emotional connotation.

Even if it is a relatively recent area, the automatic detection of emotion in text is an active research field where a variety of tools and methods have been developed with the aim of tackling this task [6], [7]. In spite of this, the effective analysis and treatment of subjective data still represent an important challenge to overcome, since textual emotion detection task presents inherent problems. As previously mentioned, one of the most challenging is the building of emotion corpora since emotion detection is even difficult for humans.

Related to this, and given their utmost importance, in this section are presented on the one hand the most relevant textual emotion corpora developed for ER, their features and how they have been developed (Section 2.1); on the other hand, semi-automatic methodologies (Section 2.2) which shows works of other NLP areas where the productivity increase when automatic processes are employed in manual tasks.

### 2.1 Textual Emotion Corpora

Despite the fact that there is not a general consensus among psychologists on the definition of emotion or how many emotions there are, research in psychology outlines two main approaches to represent the emotions that humans perceive and express: the *categorical* model (the discrete emotions approach) and the *dimensional* one [16]. The categorical model conceptualizes emotion as a set of distinct categories, whereas the dimensional one represents affects in a dimensional form where each emotion occupies a location in this space.

Regardless of the psychological debate, the categorical approach is more popular than the dimensional one in Computation Linguistics (CL) [17], [18], [19]. Principally, because it is intuitive and easily interpretable from the human and computational point of view. For this reason, the categorical model is the approach chosen for this research and the background considered in this section is focused on textual emotion corpora annotated with emotion categories.

Most of the emotion resources developed so far have been annotated manually, since, in this way, machine learning systems learn from human annotations. Among these resources, we can find corpora labelled with the six basic emotions categories proposed by Ekman [20]. This research includes: Alm [21] corpus with sentence-level annotations of approximately 185 children stories with emotion categories; Aman [17], [22] corpus which contains blog posts collected directly from the Web annotated with emotion categories and emotion intensity (high, medium, or low) at sentence level; or Affective Text corpus developed for SemEval-2007, which consists of news headlines drawn from major newspapers such as New York Times, CNN, or BBC News, annotated with emotion categories and each category has a valence value associated between 0 and 100 [10].

Moreover, there are also corpora labelled with other inventories of emotional categories, such as: Neviarouskaya [23] corpus with 1,000 sentences extracted from various stories grouped by topics (Education, Family and friends, Health and wellness, etc...) within 14 different categories among which the Ekman basic emotions; the Emotiblog-corpus that consists of a collection of blog posts manually annotated with three annotation levels (15 emotions): document, sentence and element [24]; EmoTweet-28 corpus that consists of a collection of 15,553 tweets annotated manually four facets of emotion: valence, arousal, emotion category (28 categories) and emotion [25], or Affect [26] corpus recently developed for SemEval-2018 Task 1: Affect in Tweets (sub-task Emotion Classification), which consists of a set of tweets manually annotated with 12 emotion category by using Figure Eight (F8).

The cost of this, in terms of human effort, slows down the development of an accurate emotion recognition system.

Consequently, several emotion resources have recently been developed employing emotion word hashtags to create an automatic emotion corpus on Twitter data. Mohammad [11] describes how they created a corpus from Twitter posts (Twitter Emotion Corpus - TEC) using this technique.

Several studies found in the literature use emotion word hashtags to create emotion corpora from Twitter [27], [28]. Although the use of emotion word hashtags as a technique to label data is interesting because of its simplicity and efficiency in terms of time and cost, it can be exclusively applied to social networks and microblogging services because these tags are only used in these genres. Moreover, another drawback is that people do not use the hashtags always correctly or systematically.

The problems of creating a labelled corpus are shared by other NLP tasks and one usual way to improve this situation is to automatically suggest different annotation options, so that the work of the annotators is limited to the validation of these proposals.

The following section presents works from other NLP areas where semi-automatic methodologies are employed to tackle the problem of building an annotated corpus or improving the quality of machine-generated translation in Machine Translation (MT) task.

## 2.2 Semi-Automatic Methodologies

The usability and effectiveness of semi-automatic methodologies to improve manual tasks in NLP are widely demonstrated. Pre-annotation Technique (Section 2.2.1) and MT Post-editing (Section 2.2.2) are examples of these methodologies. Both employ automatic processes to help human annotators in manual tasks, such as building annotated corpora or improving the translation quality.

### *2.2.1 Pre-annotation Technique*

Pre-annotation has been widely studied in NLP tasks, such as Named Entity Recognition (NER), Part Of Speech (POS) tagging and Semantic Frame/Role Labelling, reporting a gain in time and quality in manual annotation tasks.

Marcus et al. [29] is one of the first approaches where the pre-annotation process is assessed for POS tagging. In this work, the model of annotation consists of two stages: 1) automatic POS assignment and 2) manual correction evaluation to determine how to maximize the speed, inter-annotator consistency, and accuracy of POS tagging. The experiment showed that manual tagging took about twice as long as correcting, with about twice the inter-annotator disagreement rate and an error rate that was about 50% higher. More recently, Fort et al. [30] evaluate the influence of automatic pre-annotation on the manual POS annotation of a corpus, both from the quality and the time points of views, with specific attention to biases. Their experiments confirmed and detailed the gain in quality and demonstrated that even if a not so accurate, the tagger can help improve annotation speed.

Rehbein et al. [31] performed quite thorough experiments in the field of semantic frame assignment annotation. Although in this case, the results of the experiments are a bit disappointing as they could not find a direct improvement of annotation time using pre-annotation, they found that noisy and low-quality pre-annotation does not overall corrupt human judgment.

Lingren et al. [32] evaluate the impact of pre-annotation on annotation speed and potential bias for clinical named entity recognition in clinical trial announcements. As in other NLP tasks, they concluded that the annotator with the pre-annotated text needed less time to annotate than the annotator with non-labeled text. Moreover, the pre-annotation did not reduce the IAA or annotator performance.

### *2.2.2 Machine Translation Post-Editing*

Post-editing in Machine Translation is another example of how automatic processes help humans in manual tasks. As in our annotation task, it can be interpreted as an automatic preprocessing stage to a manual process.

Post-editing in MT is the process where humans correct machine-generated translation output to achieve an acceptable translation. The objective is to ensure that the automatic translation meets the required level of quality while maximizing speed and reducing cost.

Many studies carried out so far have demonstrated a productivity increase of MT post-editing as compared to traditional translation, such as Plitt et al. [33], who evaluated this productivity in a two-day test involving twelve participants translating from English to French, Italian, German, and Spanish. The results showed that a productivity increase for each participant; or Green et al. [34] work where a rigorous and controlled analysis of post-editing is carried out and found that post-editing leads to reduced time and, surprisingly, improved quality for three diverse language pairs (English to Arabic, French, and German).

Consequently, considering the benefits of automatic processes in manual tasks and with the aim of overcoming the cost and time-consuming shortcoming of manual annotation in ER, the objective of this research is to propose EmoLabel: a semi-automatic methodology based on pre-annotation for large-scale annotation of English emotion corpora in any genre and with standards of reliability.

## 3 EmoLabel Phase 1: Pre-annotation Process

This section describes the first phase of the methodology proposed: the pre-annotation process where the number of emotional categories is automatically reduced. We have compared two pre-annotation processes: an unsupervised approach based on Distributional Semantic Models (DSM's) and a supervised method based on Machine Learning (ML), explained in Section 3.1 and 3.2, respectively.

As input data, both processes receive 1) a collection of unlabelled sentences; and 2) a set of emotional categories. In this research, Ekman's basic emotions [20] were chosen as the set of emotional categories because it is one of the most employed emotional models in textual emotion detection. Despite this, the methods can be easily adapted to another group of emotions provided an emotion lexicon or corpus annotated with the desired emotions.

This adaptability of EmoLabel allows the use of the processes proposed in different domains or application where the set of emotion categories is different. For instance, in the educational domain where the emotions typically detected

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.2019.2927564, IEEE Transactions on Affective Computing

are BOREDOM, ANXIETY, and EXCITEMENT [35], or in news domains where emotions such as AMUSED or INSPIRED[1] are frequently analyzed.

## 3.1 Unsupervised Pre-annotation

The *unsupervised pre-annotation* process proposed is based on the use of distributional representations of the emotions and the sentences that we want to annotate.

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 words occur [36]. Finally, each word is represented by a real-valued vector called *word vector* or *word embedding* whose geometric properties prove to be semantically and syntactically meaningful [37]. Thus, words that are semantically and syntactically similar tend to be close in the semantic space.

A big advantage of using these representations that encode semantic information is that they can be generated from large corpora of unlabelled text, and can be trained on very large corpora in a reasonable amount of time. These representations have shown to improve performance on a variety of tasks. Hence, we hypothesize that these kind of distributional representations are well-suited for the pre-annotation emotion data and is a simple way to filter the number of emotion categories that can be associated with each sentence, reducing the ambiguity of the second phase of EmoLabel.

The process consists of two main steps: the representation of emotion categories and sentences in a semantic space (Step 1) and the association between emotions and sentences (Step 2), explained in subsection 3.1.1 and subsection 3.1.2 respectively.

### *3.1.1 Step 1: Emotional Categories and Sentences in Semantic Space*

The first step towards data annotation consists in encoding the emotions and the sentences in a semantic space with the help of distributional representations. This step is split into two main sub-phases shown in Figure 1:

- **Step 1.1** *Pre-processing*: for emotions, it consists in building up a bag of words related to each emotion by exploring an emotion lexicon and adding those words associated with only one of the Ekman [20]'s basic emotions to create an accurate seed without ambiguous words. While the *pre-processing* of sentences consists of tokenizing and lemmatizing each sentence and build up a bag of words from the lemmas.
- **Step 1.2** *Representation*: it consists in creating emotion vectors and sentence vectors by replacing each word in every bag of words with its vector representation. Following this, for each emotion and sentence, a single vector is obtained by applying averaging as a compositional function.

In terms of the emotion lexicon employed in the *pre-processing*, the approach proposed uses a union of two emotion lexicons (*EmoSenticNet* + *Emolex*):

- *EmoSenticNet* [38]: is a lexical resource of 13,189 words that automatically assigns qualitative emotions label and quantitative polarity scores to SenticNet concepts [39]. Ekman [20]'s emotions: ANGER, FEAR, DISGUST, SADNESS, SURPRISE, or JOY is the set of emotions employed for labelling the concepts.
- *NRC Emotion Lexicon (Emolex)* [40]: is a lexicon of general domain consisting of 14,000 English words manually compiled and associated with [41]'s eight basic emotions and two sentiments: POSITIVE and NEGATIVE. The fact that our proposal employs Ekman's emotions implies that the lexicon is reduced to 3,462 English words.

In both resources, each word has associated an emotion vector associated where each position represents an emotion: [ANGER, DISGUST, FEAR, JOY, SADNESS, and SURPRISE]. For the union of the emotion lexicons, if a word is stored in both lexicons, the word is associated with the emotions they have in common. For instance, the word *'sterile'* has the vector [0, 1, 1, 0, 1, 0] associated in EmoSenticNet and the vector [0, 0, 0, 0, 1, 0] associated in EmoLex. Considering both vectors, the resultant vector is [0, 0, 0, 0, 1, 0]. The distribution of this resource is shown in Table 1.

TABLE 1
Distribution of the emotion words annotated with only one emotion in the resultant lexicon (union of *EmoSenticNet* and *EmoLex*)

| Anger | Disgust | Fear | Joy | Sadness | Surprise |
|---|---|---|---|---|---|
| 120 | 168 | 185 | 1357 | 386 | 65 |

The fact that the emotion lexicon was the result of combining two emotion resources allows us to employ a more precise lexicon since when a word is in the two lexicons, the emotion associated is verified in both lexica.

These lexicons were chosen for two main reasons: (i) both lexicons are widely employed in relevant emotion studies; and (ii) they are labeled with the Ekman's emotions, the set of emotions employed in this research. See [42] (Section 2.3 Emotion Lexicons) for a further revision of the emotion lexicons available in the research community.

### *3.1.2 Step 2: Associating Sentences with Emotions*

Once the emotions and the unlabelled sentences are represented by distributional vectors, the next step consists in associating the sentences with the emotions:

- **Step 1** *Associating Emotions-Sentences*: because all emotions and sentences are created using the same distributional vectors and compositional function, the vector space in which they are placed is also comparable. Hence, in this step, a first emotional ranking for each sentence is proposed by measuring the cosine distance between emotions and sentences.
- **Step 2** *Re-order*: the order of the emotions proposed by the system in the previous step is re-ordered

[1] http://www.rappler.com/

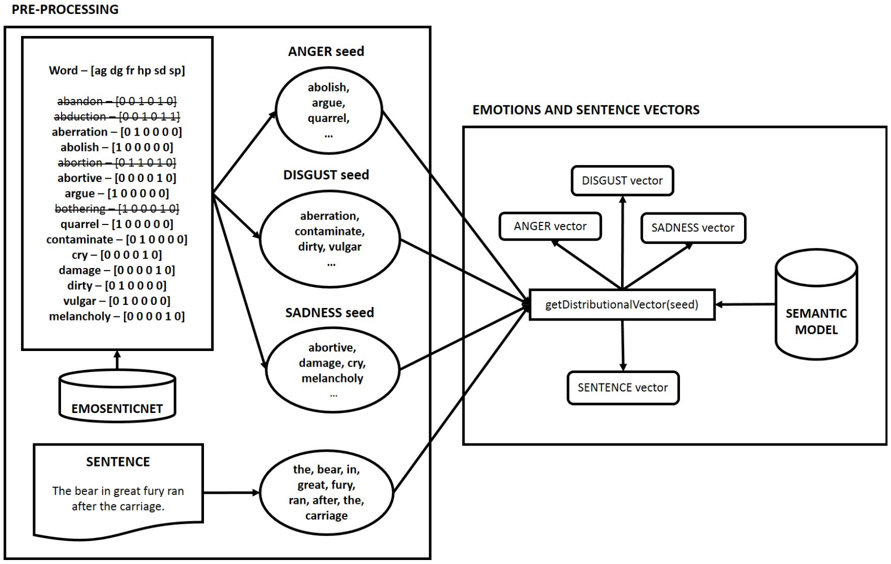

PRE-PROCESSING
Word – [ag dg fr hp sd sp]
~~abandon – [0 0 1 0 1 0]~~
~~abduction – [0 0 1 0 1 1]~~
aberration – [0 1 0 0 0 0]
abolish – [1 0 0 0 0 0]
~~abortion – [0 1 1 0 1 0]~~
abortive – [0 0 0 0 1 0]
argue – [1 0 0 0 0 0]
~~bothering – [1 0 0 0 1 0]~~
quarrel – [1 0 0 0 0 0]
contaminate – [0 1 0 0 0 0]
cry – [0 0 0 0 1 0]
damage – [0 0 0 0 1 0]
dirty – [0 1 0 0 0 0]
vulgar – [0 1 0 0 0 0]
melancholy – [0 0 0 0 1 0]
EMOSENTICNET
ANGER seed
abolish, argue, quarrel, ...
DISGUST seed
aberration, contaminate, dirty, vulgar ...
SADNESS seed
abortive, damage, cry, melancholy ...
SENTENCE
The bear in great fury ran after the carriage.
the, bear, in, great, fury, ran, after, the, carriage
EMOTIONS AND SENTENCE VECTORS
DISGUST vector
ANGER vector
SADNESS vector
getDistributionalVector(seed)
SEMANTIC MODEL
SENTENCE vector

Fig. 1. The process of representing the emotions and the sentences in the semantic space which consists of two main steps: pre-processing data and the creation of the distributional vectors.

according to the polarity (positive - negative) and subjectivity (objective - subjective) values of each sentence because, as we conclude in our preliminary work [43], this information is useful in the pre-annotation process. For that, the SA tool from `Pattern` [44] is employed, which returns an averaged (polarity, subjectivity) tuple for a given string.

- **Step 3** *Selection*: in this step, the pre-annotated emotions are finally chosen. The system selects the first three emotions resulting from the previous step results ranking. Concretely, the process pre-annotates with three emotions because it is half of the number of Ekman [20]'s basic emotions. This criterion was empirically determined, showing that the annotation of the half of emotion categories obtained a suitable balance between the reduction of the number of categories and the accuracy of the pre-annotation process. If the process would work with a greater or less group of emotion categories, the number of pre-annotated emotions would be increased or reduced respectively.

In the second sub-step (*re-order*), about the classification of Ekman [20]'s six basic emotions according to the polarity, we assume that `JOY` belongs to the positive class, while the other five emotions have negative polarity, except for `SURPRISE` since it can be employed from the positive and negative point of view. Hence, when `SURPRISE` is the first emotion proposed by the system and the subjective value is not zero, the polarity information is employed to re-order the rest of the emotions.

The re-ordering of the emotions is carried out considering the following conditions:

– If the subjective value is zero, the sentence is considered `NEUTRAL` and thus this category is proposed in the first place. Alternatively, the polarity value is evaluated it.

– If the polarity value is `POSITIVE` (higher than zero), the emotion considered positive (`JOY`) is proposed in the first position.

– If the polarity value is `NEGATIVE` (less than zero), the emotions considered negative (`ANGER`, `DISGUST`, `FEAR`, `SADNESS`) are proposed before the positive ones. The order between these emotions is determined by the semantic similarity obtained when emotion word vectors are compared to the sentence vector.

Table 2 shows examples of how the polarity and subjectivity information is employed in the pre-annotation process and the emotion proposed by the system for each sentence.

### 3.2 Supervised Pre-annotation

Supervised and unsupervised approaches have been used to automatically recognize expressions of emotion in text. In general, learning from annotated data (supervised learning) leads to better results than learning from raw information (unsupervised learning) [35]. Thus, the number of emotion recognition systems based on supervised approach is higher than unsupervised ones. The accuracy of these systems varies from 60%-70% when they try to determine the dominant emotion [17], [28], [45], which indicates that this task is unresolved. Nevertheless, these existing approaches and resources could be employed in the emotion annotation for reducing the number of emotion categories automatically. This is the intention of our method whose objective is to evaluate the usability of the supervised approach build up with existing resources in the pre-annotation task.

With this staring point in mind, three different experiments are performed:

- *Count-Emotion-Words-per-Emotion (CountWordEmo)*: the first experiment consists in the classification with a 8-feature array where the six first positions represent the number of words associated with

TABLE 2
Examples of the *unsupervised pre-annotation* process. The *1st ranking* column shows the order proposed by the system before employing the polarity and subjective information. The *Emotion proposed* column shows the pre-annotated emotions by the system after re-ordering the first ranking.

| Sentence | 1st ranking | Polarity | Subjectivity | Emotions proposed |
|---|---|---|---|---|
| This was the best summer I have ever experienced. | `joy, disgust, sadness, fear, surprise, anger` | 0.9 | 0.6 | `joy, disgust, sadness` |
| I hate fucking pills. | `anger, surprise, fear, disgust, sadness, joy` | -0.7 | 0.85 | `anger, surprise, fear` |
| Had a lovely birthday yesterday with Alex and Christine. | `sadness, joy, disgust, surprise, fear, anger` | 0.5 | 0.75 | `joy, sadness, disgust` |
| I'm becoming a broken toy and now that I have had twelve (I counted) vials of blood drawn, I just feel like I'm completely useless. | `joy, sadness, disgust, fear, surprise, anger` | -0.15 | 0.48 | `sadness, disgust, fear` |
| You don't know their middle name or the age of their sister. | `joy, disgust, sadness, fear, surprise, anger` | 0.0 | 0.0 | `neutral, joy, disgust` |

each emotion (anger, disgust, fear, joy, sadness, and surprise) and the other two contain the polarity and subjectivity values obtained with Sentiment Analysis Tool [44] for each sentence. Features: [anger, disgust, fear, joy, sadness, suprise, polarity, subjectivity].

- *Emotion-Lexicon-Words (EmoLexicon)*: the second experiment consists in the classification with features derived from the emotion lexicon. The features here are the tokens (words) of the emotion lexicon that are common with the chosen dataset. Features: [lexicon_word_1, lexicon_word_2,...].
- *Unigrams (1-grams)*: this last experiment is a corpus-based classification which uses unigrams. Unigram models have been extensively applied in text classification, and have shown good results in sentiment classification tasks [46]. Features: [corpus_word_1, corpus_word_2,...].

In *CountWordEmo* and *EmoLexicon* features, the emotion lexicon employed is the union of two emotion lexicons: EmoSenticNet and Emolex, the same resource that has been employed in the *unsupervised pre-annotation* process (Section 3.1.1).

As machine-learning algorithm, all experiments apply a Support Vector Machine (SVM) multi-class classifier using the scikit-learn [47] package throughout.

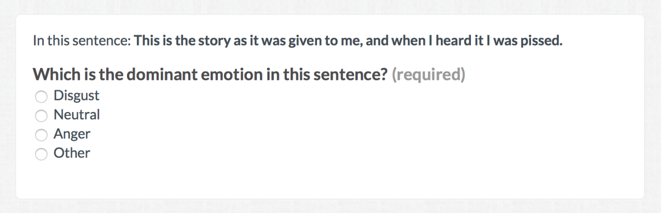
## 4 EmoLabel Phase 2: Manual Annotation

Once the unlabeled sentences have been pre-annotated, a manual labeling task is performed by humans annotators with the aim of determining which is the dominant emotion associated with each sentence.

In order to evaluate the impact of the pre-annotation on the quality of the resulting corpus and on the time needed to annotate, two experimentation have been designed. In both experiments, manually labeling are carried out with three different experimental setups:

- **Pre-ML**: in this setup, the best model of the *supervised* pre-annotation (machine learning approach) is used to select the pre-annotated emotions in each sentence.
- **Pre-WE**: in this setup, the best model of the *unsupervised* pre-annotation (*word embedding* approach) is used to select the pre-annotated emotions proposed to human annotators.
- **No-pre**: in this setup, no pre-annotation process is employed. Thus, all emotion categories employed are shown to human annotators, as Figure 3 shows.

When the pre-annotation process is employed (*Pre-ML* and *Pre-WE* tasks), the emotions proposed by the system are shown in first place to humans annotators, who also have the possibility of selecting another emotion (no automatically pre-selected). To do this, they have to choose the option 'Other' and the rest of emotions are displayed, as Figure 2 shows.

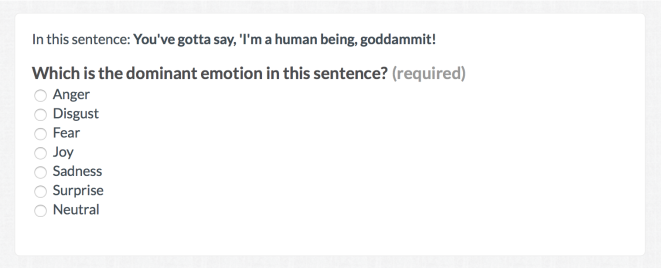


In this sentence: **This is the story as it was given to me, and when I heard it I was pissed.**

**Which is the dominant emotion in this sentence?** (required)

○ Disgust
○ Neutral
○ Anger
○ Other

Fig. 2. An example of how sentences are shown to the human annotators when the supervised pre-annotation (Pre-ML) process is employed.



In this sentence: **You've gotta say, 'I'm a human being, goddammit!**

**Which is the dominant emotion in this sentence?** (required)

○ Anger
○ Disgust
○ Fear
○ Joy
○ Sadness
○ Surprise
○ Neutral

Fig. 3. An example of how sentences are shown to the human annotators when the pre-annotation process is no employed.

All manual annotation tasks were carried out by three annotators with a good knowledge of English. They were in-

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.2019.2927564, IEEE Transactions on Affective Computing

structed to annotate from the *text perspective* since a previous work of emotion annotation task with emotion categories developed by Mohammad and Turney [40] demonstrated that the *text* perspective yields higher IAA than the *reader* perspective. *Reader* perspective is how someone feels after reading an utterance, whereas in *text* perspective, no actual person is specified as perceiving an emotion and emotion is an intrinsic property of a sentence.

In a previous experiment, this phase was designed using three different datasets ($D1$, $D2$, $D3$) for each setup (*Pre-ML*, *Pre-WE*, *NO-pre*). However, we detected that the random selection of the sentences that populates each corpus may negatively or positively affect the results achieved in each setup. Consequently, we decided to apply cross-validation so that the results are not affected by chance bias and annotators' learning curve. Hence, each annotator carried out three labeling tasks in the order described in Table 3. By doing this, each dataset was annotated with all setups and were labeled by different annotators.

TABLE 3
Cross-validation setup

| | $D1$ | $D2$ | $D3$ |
|---|---|---|---|
| Annotator 1 | Pre-ML | Pre-WE | *No-pre* |
| Annotator 2 | *No-pre* | Pre-ML | Pre-WE |
| Annotator 3 | Pre-WE | *No-pre* | Pre-ML |

Furthermore, due to the difficulty of manual emotion annotation, three training tasks were performed in order to ensure a correct understanding of the task. In each training, the three annotators labeled 21 sentences, three per emotion and three for the NEUTRAL category. After each training, we met for resolving doubts and clarifying aspects related to the annotation guide. Table 4 shows Fleiss [48]' kappa values reached between the three annotators in each training.

TABLE 4
IAA in terms of Fleiss' kappa between the three annotators in each training task

| Training 1 | Training 2 | Training 3 |
|---|---|---|
| 0.4512 | 0.655 | 0.610 |

Considering the benefits of crowdsourcing platforms in manual annotation [49], we consider Figure Eight[2] (F8) (earlier called CrowdFlower) as the most suitable tool to implement this second phase of EmoLabel.

F8 platform allows accessing an online workplace to clean, label and enrich data. A big advantage of this platform is that there are thousands of people available to read content and score it, with a relatively inexpensive cost. Moreover, F8 offers the possibility of sending the job/task exclusively to your team (using internal contributors option). In this research, we chose to use internal contributors, due to the need of controlling that all tasks were annotated by the same annotators. Although the external contributors were not used, this tool provides us the following advantages: (i) low level of complexity for the creation of the questionnaires and the tasks, (ii) user-friendliness of the application for annotators, and (iii) adaptability of the platform to different types of devices.

# 5 EVALUATION

The assessment of EmoLabel requires an intrinsic and an extrinsic evaluation. The intrinsic evaluation involves assessing the pre-annotation process whereas the extrinsic evaluation has as objective the evaluation of annotators' performance in the second phase of the methodology.

## 5.1 Data Description

In order to assess the usability for different genres, the approaches are evaluated against two emotion corpora: Aman corpus and EmoTweet-28 corpus.

*Aman corpus [17], [22]*. This corpus contains sentence-level annotations of 4,000 sentences from blogs posts collected directly from the Web. This corpus was annotated manually with the six emotion categories proposed by Ekman plus the emotion intensity (high, medium, or low). The distribution of the corpus is shown in Table 5

TABLE 5
Distribution of the sentences per emotion on Aman corpus, a corpus of blog posts annotated with Ekman's basic emotions.

| Anger | Disgust | Fear | Joy | Sadness | Surprise | Neutral | Total |
|---|---|---|---|---|---|---|---|
| 179 | 172 | 115 | 536 | 173 | 115 | 2,800 | 4,090 |

*EmoTweet-28 corpus [25]*. This dataset consists of a collection of 15,553 tweets annotated with 28 emotion categories. The corpus contains annotations for four facets of emotion: valence, arousal, emotion category and emotion cues. As this research works with the Ekman basic emotions, a reduced corpus of EmoTweet-28 is employed (EmoTweet-5). This corpus contains those tweets annotated with anger, fear, joy, sadness, surprise and the same proportion of neutral tweets as the original corpus. Finally, EmoTweet-5 comprises 5,931 tweets whose distribution per emotion is shown in Table 6.

TABLE 6
Distribution of the sentences per emotion on EmoTweet-5, a reduced version of EmoTweet-28 that contains tweets annotated with Ekman's basic emotions.

| Anger | Fear | Joy | Sadness | Surprise | Neutral | Total |
|---|---|---|---|---|---|---|
| 986 | 180 | 1306 | 350 | 179 | 2,930 | 5,931 |

These corpora were chosen for two main reasons: (i) both corpora have been employed in relevant emotion studies as a benchmark [25], [50], [51], [52]; and (ii) it is possible to assess the effectiveness of the pre-annotation process in different social media genres that allows people to post messages to share information, opinions, and emotions: blogs and tweets.

[2] https://www.figure-eight.com/

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.2019.2927564, IEEE Transactions on Affective Computing

## 5.2 Intrinsic Evaluation

The objective of the intrinsic evaluation is to assess which is the best pre-annotation process that is going to be employed in the second phase of EmoLabel. To achieve that, the evaluation is carried out comparing the emotions proposed by each method with the gold standard of the test corpora. Specifically, the intrinsic evaluation has been carried out in the two corpora: Aman corpus [17], [22] and a reduced version of EmoTweet-28 [25].

For the evaluation purpose, the 30% of data of each corpus is employed because the 70% of data is applied for training in the *supervised* approach. Moreover, the use of the same test data for all approaches allows the results comparability.

As far as the number of emotion categories pre-annotated is concerned, in Aman corpus the sentences are pre-annotated with three emotions since the corpus is labeled with six categories, while in EmoTweet-5 the sentences are pre-annotated with two emotions.

Concerning the evaluation methodology, the pre-annotation process is assessed measuring the precision (P), recall (R), and F1-score (F1) of the emotions proposed by our system against the gold standard of the test corpora, as well as the macro-average of each of these metrics for each model.

As the process pre-annotates the half of the number of emotion categories, if the correct emotion (the gold standard) is one of the pre-annotated emotions, the prediction is considered as correct. In the calculation of average scores, the `NEUTRAL` class is included since we consider important that the pre-annotation process is able to distinguish between emotional and non-emotional content.

### 5.2.1 Unsupervised Pre-annotation

Given that the pre-annotation process is based on distributional representations, different Distributional Semantic Models (DSM) have been evaluated. Concretely, our approaches have been evaluated using four semantic spaces.

- *Vector Space Model (VSM) (baseline)*: a simple semantic space is built by a VSM created with *EmoSenticNet+Emolex*, the emotion lexicon explained in Section 3.1.1. In this space, the emotions and sentences are represented by a vector that contains information about which *EmoSenticNet+Emolex* words occur in each sentence or emotion.
- *Affective Space* [53]: this set is the 100-dimensional vector space representation of AffectNet (a matrix of affective commonsense knowledge in which common-sense concepts are linked to semantic and affective features).
- *GloVe vectors* [37]: here, two set of vectors are employed depending on the test corpus. For Aman corpus, the 300-dimension vectors trained on 42 billion tokens of web data from Common Crawl[3] are applied. And for EmoTweet-5, the 200-dimension vectors trained on 2 billion tweets (27 billion tokens) is used.
- *Ultradense Sentiment Analysis Word Embeddings* [54]: these pre-trained embeddings are the results of learning an orthogonal transformation of the embedding space that focuses the information relevant for a task. For this evaluation, two sets of ultradense vectors are employed. Specifically, for Aman corpus, the 300-dimension Google News vectors are applied. And for EmoTweet-5, the 400-dimension embeddings on a Twitter corpus of size 5.4 billion of tweets. Both sets of vectors are focused on Sentiment Analysis.

The results of the *unsupervised pre-annotation* process for each DSM are shown in Table 7 for Aman corpus, and in Table 8 for EmoTweet-5.

The results of the *unsupervised* pre-annotation on Aman corpus show that considering the macro-average F1-score, all models outperform significantly the baseline. Although, the best result is obtained by *Common Crawl Glove* model due to its recall and precision values in `JOY` and `SADNESS` emotions. From these high values, we may draw that these emotions are frequently found between the emotions proposed by the system. With respect to the *Ultradense SA* model, it also obtains high recall values in `JOY` and `SADNESS` emotions and moreover, it reaches the best values for F1-measure for `ANGER` and `DISGUST`, two of the emotions hard to detect in text. And about *Affective Space*, it is interesting to highlight the results obtained by `FEAR` and `SURPRISE` considering that *Affective Space* is a set of 100-dimension vectors and the vocabulary represented in this space is smaller compared to the rest of the models.

Regarding the results of the *unsupervised* pre-annotation on EmoTweet-5, they show that *Twitter GloVe* and *Ultradense SA* outperform significantly the baseline whereas *Affective Space* does not improve it. This can be due to the fact that the vocabulary of *Affective Space* is formal and the language employed in Twitter is more informal, not carefully edited or with grammatical errors. Hence, the results emphasize the importance of using DSM's adapted to the genre when the process runs with social media texts. With respect to the *Ultradense SA* model, in this corpus, its recall improvements are confirmed for `JOY`, `SADNESS` and `SURPRISE`, and moreover, the interesting values obtained by `ANGER` and `DISGUST` continue to be noted when EmoTweet-5 is employed for the evaluation. As for the results of the *Twitter GloVe* model, the best performance is also achieved in `JOY` and `SADNESS` emotions for its high recall values.

Furthermore, the results reflect that another important factor in *unsupervised* approach is the coverage of the lexicon employed. For instance, the high coverage of `JOY` emotion allows to *Affective Space* model achieving good results in this emotion despite this space is not adapted to the genre. However, the low coverage in `SURPRISE` emotion could explain that the *unsupervised* pre-annotation approach is not able to detect this emotion in this genre since the best F1-value obtained is 13%.

Comparing both evaluations, the results show that the best models are *Glove vectors* and *Ultradense SA* and therefore the need of using word embeddings with the following features: (1) embeddings built from a large amount of data for representing a large vocabulary; (2) with high dimensionality to codify more semantic features because

[3] http://commoncrawl.org/

TABLE 7
Precision, Recall and F1-values obtained in the *unsupervised pre-annotation* using different distributional representations on Aman corpus.

| | *Unsupervised Pre-annotation - Aman corpus* | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Baseline | | | Affective Space | | | Common Crawl GloVe | | | Ultradense SA | | |
| | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 |
| Anger | 0.35 | 0.22 | 0.27 | 0.17 | 0.02 | 0.03 | 1.00 | 0.37 | 0.54 | 0.58 | 0.65 | **0.61** |
| Disgust | 0.49 | 0.38 | 0.43 | 0.21 | 0.85 | 0.33 | 1.00 | 0.23 | 0.38 | 0.61 | 0.54 | **0.57** |
| Fear | 0.34 | 0.32 | 0.33 | 0.84 | 0.76 | **0.80** | 0.96 | 0.68 | 0.79 | 0.55 | 0.35 | 0.43 |
| Joy | 0.30 | 0.83 | 0.44 | 0.31 | 0.94 | 0.47 | 0.71 | 0.94 | **0.81** | 0.60 | 0.94 | 0.73 |
| Sadness | 0.40 | 0.60 | 0.48 | 0.65 | 0.25 | 0.36 | 0.87 | 0.75 | **0.80** | 0.62 | 0.87 | 0.72 |
| Surprise | 0.16 | 0.18 | 0.17 | 0.55 | 0.62 | **0.58** | 0.07 | 1.00 | 0.13 | 0.10 | 0.97 | 0.17 |
| Neutral | 0.86 | 0.58 | **0.69** | 0.92 | 0.48 | 0.63 | 0.93 | 0.48 | 0.63 | 0.95 | 0.48 | 0.63 |
| Macro Avg. | 0.42 | 0.44 | 0.40 | 0.52 | 0.56 | 0.46 | 0.79 | 0.64 | **0.58** | 0.57 | 0.69 | 0.55 |
| Micro Avg. | 0.57 | 0.57 | 0.57 | 0.54 | 0.54 | 0.54 | 0.56 | 0.56 | 0.56 | 0.58 | 0.58 | **0.58** |

TABLE 8
Precision, Recall and F1-values obtained in the *unsupervised pre-annotation* using different distributional representations on EmoTweet-5 corpus.

| | *Unsupervised Pre-annotation - EmoTweet-5* | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Baseline | | | Affective Space | | | Twitter GloVe | | | Ultradense SA | | |
| | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 |
| Anger | 0.41 | 0.11 | 0.18 | 1.00 | 0.01 | 0.01 | 1.00 | 0.07 | 0.14 | 0.77 | 0.30 | **0.43** |
| Fear | 0.29 | 0.31 | 0.30 | 0.16 | 0.59 | 0.25 | 0.52 | 0.22 | 0.31 | 0.54 | 0.35 | **0.43** |
| Joy | 0.40 | 0.92 | 0.56 | 0.57 | 0.94 | 0.71 | 0.75 | 0.96 | **0.84** | 0.58 | 0.96 | 0.73 |
| Sadness | 0.23 | 0.11 | 0.15 | 0.92 | 0.11 | 0.20 | 0.67 | 0.48 | 0.56 | 0.51 | 0.69 | **0.59** |
| Surprise | 0.09 | 0.20 | **0.13** | 0.06 | 0.44 | 0.11 | 0.05 | 0.63 | 0.10 | 0.07 | 0.46 | **0.13** |
| Neutral | 0.71 | 0.45 | **0.55** | 0.73 | 0.44 | **0.55** | 0.74 | 0.44 | **0.55** | 0.76 | 0.44 | **0.55** |
| Macro Avg. | 0.36 | 0.35 | 0.31 | 0.57 | 0.42 | 0.31 | 0.62 | 0.47 | 0.42 | 0.54 | 0.53 | **0.48** |
| Micro Avg. | 0.47 | 0.47 | 0.47 | 0.46 | 0.46 | 0.46 | 0.49 | 0.49 | **0.49** | 0.54 | 0.54 | 0.54 |

the better performances; and (3) adapted to the genre of the text that we want to annotate in order to the semantic space be representative. Moreover, the results highlight the importance of the lexicon coverage of the lexicon employed in the *unsupervised* pre-annotation. Finally, it is interesting to mention the improvements obtained by *Ultradense SA* in ANGER and DISGUST emotions since this enhancement is shown regardless the genre employed.

### 5.2.2 Supervised Pre-annotation

As mentioned previously, a multi-classifier Support Vector Machine (SVM) is applied using three set of features: *CountWordEmo*, *EmoLexicon*, and *Unigrams (1-grams)* explained in Section 3.2. For the evaluation of the supervised approach, the datasets are split in 70% for training and 30% for test. And, the optimal set of hyperparameters for each SVM was determined based on an exhaustive search through the parameter space using 10-fold cross-validation. Using this, the parameters selected in each SVM are described below:

- Aman Corpus
  - *CountWordEmo*: an RBF kernel, C value: 1, gamma value: 0.001
  - *EmoLexicon*: an Linear kernel, C value: 1
  - *Unigrams (1-grams)*: an Linear kernel, C value: 1
- Emo-Tweet-5
  - *CountWordEmo*: an Linear kernel, C value: 10
  - *EmoLexicon*: an RBF kernel, C value: 100, gamma value: 0.001
  - *Unigrams (1-grams)*: an RBF kernel, C value: 100, gamma value: 0.001

The results of the *supervised* pre-annotation process for each set of features are shown in Table 9 for Aman corpus, and in Table 10 for EmoTweet-5.

The results of the *supervised* pre-annotation on Aman corpus show that considering the macro-average F1-score, the best result is obtained by the *1-grams* model due to the fact that its F1-score is higher than 75% for all the emotions. With respect to *CountWordEmo* and *EmoLexicon*, the results show these models are not able to detect emotions like FEAR and SURPRISE because these set of features are heavily dependent on the coverage of the lexicon employed.

As for the results of the *supervised* pre-annotation on EmoTweet-5, the conclusion is the same as the one for on Aman corpus since the best performance is obtained by *1-grams* and *CountWordEmo* and *EmoLexicon* continue having

TABLE 9
Precision, Recall and F1-values obtained in the *supervised pre-annotation* using different set of features on Aman corpus.

| | *Supervised Pre-annotation - Aman corpus* | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | CountWordEmo | | | EmoLexicon | | | 1-grams | | |
| | P | R | F1 | P | R | F1 | P | R | F1 |
| Anger | 1.00 | 0.39 | 0.56 | 0.95 | 0.72 | **0.82** | 0.90 | 0.65 | 0.75 |
| Disgust | 1.00 | 0.40 | 0.58 | 1.00 | 0.21 | 0.35 | 0.94 | 0.65 | **0.77** |
| Fear | 0.00 | 0.00 | 0.00 | 0.75 | 0.09 | 0.16 | 0.92 | 0.65 | **0.76** |
| Joy | 1.00 | 0.99 | **0.99** | 0.98 | 0.99 | 0.98 | 0.96 | 0.96 | 0.96 |
| Sadness | 1.00 | 0.56 | 0.72 | 0.88 | 0.29 | 0.43 | 0.97 | 0.73 | **0.84** |
| Surprise | 1.00 | 0.09 | 0.16 | 1.00 | 0.09 | 0.16 | 1.00 | 0.65 | **0.79** |
| Neutral | 0.85 | 1.00 | 0.92 | 0.85 | 1.00 | 0.92 | 0.92 | 1.00 | **0.96** |
| Macro Avg. | 0.84 | 0.49 | 0.56 | 0.92 | 0.48 | 0.55 | 0.95 | 0.75 | **0.83** |
| Micro Avg. | 0.87 | 0.87 | 0.87 | 0.87 | 0.87 | 0.87 | 0.93 | 0.93 | **0.93** |

TABLE 10
Precision, Recall and F1-values obtained in the *supervised pre-annotation* using different set of features on EmoTweet-5 corpus.

| | *Supervised Pre-annotation - EmoTweet-5* | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | CountWordEmo | | | EmoLexicon | | | 1-grams | | |
| | P | R | F1 | P | R | F1 | P | R | F1 |
| Anger | 0.77 | 0.87 | **0.82** | 0.69 | 0.47 | 0.56 | 0.81 | 0.83 | **0.82** |
| Fear | 0.00 | 0.00 | 0.00 | 1.00 | 0.02 | 0.04 | 0.86 | 0.24 | **0.38** |
| Joy | 0.87 | 0.66 | 0.75 | 0.88 | 0.87 | 0.87 | 0.94 | 0.87 | **0.90** |
| Sadness | 1.00 | 0.08 | 0.15 | 0.87 | 0.23 | 0.37 | 0.88 | 0.53 | **0.66** |
| Surprise | 0.00 | 0.00 | 0.00 | 0.75 | 0.06 | 0.12 | 0.73 | 0.23 | **0.35** |
| Neutral | 0.62 | 0.99 | 0.76 | 0.58 | 0.93 | 0.71 | 0.74 | 0.97 | **0.84** |
| Macro Avg. | 0.54 | 0.43 | 0.41 | 0.79 | 0.43 | 0.45 | 0.83 | 0.61 | **0.66** |
| Micro Avg. | 0.72 | 0.72 | 0.72 | 0.69 | 0.69 | 0.69 | 0.82 | 0.82 | **0.82** |

problems to detect FEAR and SURPRISE. In general, the results on EmoTweet-5 are worse than on Aman corpus due to the fact that Twitter is a platform where the text with grammatical errors or not carefully edited is most prevalent. These worse results are most noticeable in those set of features that are exclusively dependent of the lexicon because its coverage in this genre is low.

Comparing the evaluation in both corpora, the results allow concluding that the set of features employed needs to contain information about the text to be processed and not depend exclusively on an emotion lexicon. Naturally, if the *supervised* emotion approach is improved with an advanced set of features or algorithms such as selecting features depending on the genre or domain or applying a classifier combination technique [55], the pre-annotation would improve. However, our aim is to assess the viability of a *supervised* emotion model for pre-annotation, thus a sophisticated feature engineering has not been carried out.

### 5.3 Extrinsic Evaluation

The extrinsic evaluation has as objective the assessment of the work of the annotators in the second phase of EmoLabel. To achieve that, a manual annotation task is carried out for three annotators.

Aman corpus is the dataset employed to assess this phase. Correctly, the test data (30%) previously used for evaluating the pre-annotation processes. This data is split into three datasets of 100 sentences each ($D1$, $D2$, $D3$), whose distribution per emotion is shown in Table 11. The distribution has done in an equitable way with the aim of having the same number of sentences for each emotion.

TABLE 11
Distribution of the number of sentences per emotion annotated in each manual task.

| Anger | Disgust | Fear | Joy | Sadness | Surprise | Neutral | Total |
|---|---|---|---|---|---|---|---|
| 16 | 15 | 11 | 16 | 15 | 11 | 16 | 100 |

As previously introduced in Section 4, the manual annotation task is split into several sub-tasks where all datasets are labeled by all annotators using different setups (Table 3): (i) Pre-ML setup where the *supervised* pre-annotation is employed; (ii) Pre-WE setup which uses the *unsupervised* pre-annotation; and (iii) No-pre setup where the pre-annotation method is not applied. Considering the results of intrinsic evaluation on Aman corpus, the best pre-annotation methods are selected for these tasks. Thus, the approach *1-grams* is applied for Pre-ML setup and the *GloVe* model is used for Pre-WE setup.

Regarding the agreement metrics employed, the *k*-coefficient metrics are well-known in NLP for measuring IAA because these are designed for nominal-scaled variables. Thus, the manual annotation process is assessed calculating Fleiss [48]'s kappa between each annotator and Aman corpus gold standard. Kappa represents the fraction of agreement observed that is not related to chance. This is a statistical measure for assessing the reliability of agreement between a fixed number of annotators when assigning categorical ratings to a number of items or classifying items.

The results of the agreement achieved by each annotator in each setup are shown in Table 12.

TABLE 12
IAA in terms of Fleiss' kappa between each annotator and the Aman corpus gold standard.

| Annotator | Pre-ML | Pre-WE | No-pre | Macro-avg |
|---|---|---|---|---|
| 1 | 0.588 | 0.649 | **0.659** | 0.632 |
| 2 | **0.637** | 0.626 | 0.578 | 0.614 |
| 3 | **0.635** | 0.544 | 0.555 | 0.578 |
| Macro-avg | **0.620** | 0.606 | 0.597 | |

Regarding the agreement evaluation, all tasks reach macro-average scores of 0.60, a "substantial agreement" according to Landis and Koch [56]. Therefore, the results demonstrate that the pre-annotation process does not reduce the IAA or annotation performance. Moreover, it is interesting to mention that two of the three annotators reach their best agreement values in the tasks with Pre-ML, a fact which shows that an accurate pre-annotation process could help human annotators to effectively label emotions.

In terms of time effort, the F8 platform records the time the annotator submitted the judgment and the time at which the annotator started working on the judgment for each page. This allows measuring the time required to complete each task. The macro-average time obtained by each annotator in each setup is shown in Table 13.

TABLE 13
Annotation time of each annotator in all manual annotation tasks.

| Annotator | Pre-ML | Pre-WE | No-pre | Macro-avg |
|---|---|---|---|---|
| 1 | **02:01** | 04:37 | 03:31 | 3:53 |
| 2 | 04:09 | 03:55 | **03:48** | 4:00 |
| 3 | **04:27** | 04:57 | 05:49 | 4:00 |
| Macro-avg | **03:32** | 04:29 | 04:22 | |

As for time effort, the macro-average time shows that Pre-ML reduces annotation time by near 20% (19,1%) with respect to the second-best time (No-pre). And as happen in agreement evaluation, two of the three annotators obtain a time gain of more than 20% (42,6% for Annotator 1 and 23,5% for Annotator 3) when the pre-annotation process is applied (Pre-ML) with respect to their tasks with No-pre. Hence, the evaluation performed demonstrates that the pre-annotation process reduces the annotation time required in emotion labeling.

As far as the comparison between the pre-annotation methods is concerned, whereas there are no significant differences in terms of agreement values, in time evaluation, Pre-ML reduces annotation time by 24,8% with respect to Pre-WE. This indicates that the use of inaccurate pre-annotation methods may have the negative effect of increasing the time needed for the annotation, thus being an obstacle, rather than a support measure. This can be due to the supervised approach generalizing better than the unsupervised one, and thus the emotions proposed as pre-annotated being more useful for annotators. Pre-WE mainly depends on its coverage of the lexicon and its representation in the semantic space. Consequently, its performance is comparable with Pre-ML when the sentences are simple and the vocabulary is not ambiguous, as shown in Table 14 (row 1). However, Pre-WE performance is lower whether the sentence is more complex (Table 14 - row 2).

Concerning the labeling performed by each annotator, it is of remarkable interest the fact that Annotator 3 reaches their best agreement and time scores when the *supervised* pre-annotation process is employed (Pre-ML) since this annotator had some difficulties in understanding the task (Table 12). It may, therefore, be concluded that pre-annotation could be used as a strategy to improve the performance of inaccurate annotators. This is an important factor if we want to carry out emotion annotation in crowdsourcing platforms (AMT or F8) with external contributors, since in this kind of tools does not provide details of the annotators' background.

## 6 Conclusions

As presented in the introductory section of this paper, the rationale behind our research is the need to simplify the emotion annotation task so that to improve its reliability and efficiency. For this purpose, we present EmoLabel: a semi-automatic methodology consisting in two phases: (1) an automatic process to pre-annotate the unlabelled sentences with a reduced number of emotion categories; and (2) a manual refinement process where human annotators determine which is the dominant emotion between the pre-defined set of possibilities. Two pre-annotation strategies are presented: *unsupervised* proposal with the aim of minimizing the human intervention and *supervised* method where simple emotion models are build up, exploiting corpora or models previously developed.

According to the extrinsic evaluation, the experiments performed demonstrate the benefits of pre-annotation processes in emotion labeling since the results on annotation time show a gain of near 20% when the pre-annotation process is applied (Pre-ML) with respect to No-pre. Moreover, the experiments performed show that all tasks reach "substantial agreement" and therefore the pre-annotation process does not reduce the IAA or annotator performance.

With respect to the intrinsic evaluation, the gains of the *supervised* pre-annotation method in terms of and time with respect to the *unsupervised* pre-annotation process, allow concluding that the use of this method is more helpful for annotators than the *unsupervised* approach. Consequently,

TABLE 14
Examples of Pre-ML and Pre-ML performance for the same sentences of Aman corpus

| Sentence | Pre-ML | Pre-WE | Gold Standard |
|---|---|---|---|
| *because you are ANNOYING US.* | anger, neutral, disgust, joy, sadness, fear, surprise | anger, disgust, sadness, surprise, fear, joy, neutral | anger |
| *I'm kind of freaking out about debate but I'm kind of okay.* | neutral, fear, anger, joy, disgust, surprise, sadness | surprise, joy, anger, sadness, fear, disgust, neutral | fear |

the existing *supervised* emotion detection systems developed so far could be employed to annotate new data.

Finally, the improvements reached by Annotator 3 (with the lowest performance) (Table 12) in terms of time and agreement demonstrate the usability of our methodology with inaccurate annotators since his best performances are obtained when a pre-annotation process is employed (Pre-ML). This is one of the most remarkable results of this research since it can be inferred that when the annotation task will be carry out in crowdsourcing platforms, where the knowledge of external contributors is limited, the benefit of using pre-annotation is likely to be even greater.

Our future research will focus on 1) developing new manual tasks with more annotators and a larger amount of the data in a crowdsourcing platform; 2) testing existing supervised emotion detection systems or systems with a sophisticated set of features to improve the *supervised pre-annotation* process; 3) performing a new experiment where one of the three pre-selected categories will be the gold standard in order to analyze the maximal benefits of pre-annotation; 4) evaluating the pre-annotation proposals against methods based on semantic models); and 5) testing emotion annotation frameworks used in other modalities to annotate emotion in text such as CARMA software [14];

## ACKNOWLEDGMENTS

This research has been supported by the Spanish Government (ref. RTI2018-094653-B-C22) and the Valencian Government (grant no. PROMETEU/2018/089). It has also been funded by the FPI grant (BES-2013-065950) and the research stay grant (EEBB-I-17-12578) from the Spanish Ministry of Science and Innovation. This research stay was held in CLiPS Research Center at the University of Antwerp. We thank Irene Molina Pozas and María Leal Lillo for participating in the manual annotation task and offering us their disinterested collaboration to carry out this task.

## REFERENCES

[1] A. Farzindar and D. Inkpen, *Natural Language Processing for Social Media*. Morgan & Claypool Publishers, 2015.

[2] 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.

[3] B. Desmet and V. Hoste, "Emotion detection in suicide notes," *Expert Systems with Applications*, vol. 40, no. 16, pp. 6351–6358, nov 2013.

[4] 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.

[5] 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.

[6] K. Hulliyah, N. S. A. A. Bakar, and A. R. Ismail, "Emotion recognition and brain mapping for sentiment analysis: A review," in *2017 Second International Conference on Informatics and Computing (ICIC)*, Nov 2017, pp. 1–5.

[7] S. Al-Saaqa, H. Abdel-Nabi, and A. Awajan, "A survey of textual emotion detection," in *2018 8th International Conference on Computer Science and Information Technology (CSIT)*, July 2018, pp. 136–142.

[8] L. Deng and D. Yu, *Deep Learning: Methods and Applications*. Now Publishers Inc., may 2014.

[9] R. W. Picard, *Affective computing*. MIT Press Cambridge, MA, USA ©1997, 1997.

[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] S. M. Mohammad, "#Emotional Tweets," in *Proceedings of the First Joint Conference on Lexical and Computational Semantics*, 2012.

[12] M. E. Sánchez-Gutiérrez, E. M. Albornoz, F. Martinez-Licona, H. L. Rufiner, and J. Goddard, "Deep learning for emotional speech recognition," in *Pattern Recognition*, J. F. Martínez-Trinidad, J. A. Carrasco-Ochoa, J. A. Olvera-Lopez, J. Salas-Rodríguez, and C. Y. Suen, Eds. Cham: Springer International Publishing, 2014, pp. 311–320.

[13] A. Meftah, Y. Alotaibi, and S. A. Selouani, "Emotional speech recognition: A multilingual perspective," in *2016 International Conference on Bio-engineering for Smart Technologies (BioSMART)*, Dec 2016, pp. 1–4.

[14] J. Girard, "Carma: Software for continuous affect rating and media annotation," *Journal of Open Research Software*, vol. 2, p. e5, 07 2014.

[15] K. Sharma, C. Castellini, F. Stulp, and E. L. V. den Broek, "Continuous, real-time emotion annotation: A novel joystick-based analysis framework," *IEEE Transactions on Affective Computing*, vol. PP, no. 99, pp. 1–1, 2017.

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

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

[18] F. Vaassen, "Measuring Emotion: Exploring the feasibility of automatically classifying emotional text," Master's thesis, University of Antwerp, Antwerp, Belgium, 2014.

[19] D. Ghazi, "Identifying Expressions of Emotions and Their Stimuli in Text," Ph.D. dissertation, University of Ottawa, 2016.

[20] P. Ekman, "An argument for basic emotions," *Cognition and Emotion*, pp. 169–200, 1992.

[21] 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.

[22] S. Aman and S. Szpakowicz, "Using rogets thesaurus for fine-grained emotion recognition," in *In: Proc. Third International Joint Conf. on Natural Language Processing (IJCNLP)*, 2008, pp. 296–302.

[23] 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.

[24] E. Boldrini, A. Balahur, P. Martínez-Barco, and A. Montoyo, "Using emotiblog to annotate and analyse subjectivity in the new textual genres," *Data Mining and Knowledge Discovery*, vol. 25, no. 3, pp. 603–634, Nov 2012.

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.2019.2927564, IEEE Transactions on Affective Computing

[25] 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.

[26] S. M. Mohammad, F. Bravo-Marquez, M. Salameh, and S. Kiritchenko, "Semeval-2018 task 1: Affect in tweets," in *Proceedings of The 12th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, 2018, pp. 1–17.

[27] 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.

[28] W. Wang, L. Chen, K. Thirunarayan, and A. P. Sheth, "Harnessing Twitter "Big Data" for Automatic Emotion Identification," in *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*. IEEE Computer Society, sep 2012, pp. 587–592.

[29] M. P. Marcus, M. A. Marcinkiewicz, and B. Santorini, "Building a large annotated corpus of english: The penn treebank," *Comput. Linguist.*, vol. 19, no. 2, pp. 313–330, Jun. 1993.

[30] K. Fort and B. Sagot, "Influence of pre-annotation on pos-tagged corpus development," in *Proc. of the Fourth Linguistic Annotation Workshop*, ser. LAW IV '10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 56–63.

[31] I. Rehbein, J. Ruppenhofer, and C. Sporleder, "Assessing the benefits of partial automatic pre-labeling for frame-semantic annotation," in *Proc. of the Third Linguistic Annotation Workshop*. Suntec, Singapore: Association for Computational Linguistics, August 2009, pp. 19–26.

[32] T. Lingren, L. Deleger, K. Molnar, H. Zhai, J. Meinzen-Derr, M. Kaiser, L. Stoutenborough, Q. Li, and I. Solti, "Evaluating the impact of pre-annotation on annotation speed and potential bias: natural language processing gold standard development for clinical named entity recognition in clinical trial announcements," *Journal of the American Medical Informatics Association : JAMIA*, vol. 21, no. 3, pp. 406–413, 2014.

[33] M. Plitt and F. Masselot, "A productivity test of statistical machine translation post-editing in a typical localisation context." *Prague Bull. Math. Linguistics*, vol. 93, pp. 7–16, 2010.

[34] S. Green, J. Heer, and C. D. Manning, "The efficacy of human post-editing for language translation," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '13. New York, NY, USA: ACM, 2013, pp. 439–448.

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

[36] G. Lapesa and S. Evert, "A large scale evaluation of distributional semantic models: Parameters, interactions and model selection," *Transactions of the Association for Computational Linguistics (TACL)*, vol. 2, pp. 531–545, 2014.

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

[38] 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.

[39] E. Cambria, S. Poria, D. Hazarika, and K. Kwok, "Senticnet 5: Discovering conceptual primitives for sentiment analysis by means of context embeddings," in *AAAI Conference on Artificial Intelligence*, 2018.

[40] S. M. Mohammad and P. D. Turney, "Crowdsourcing a Word-Emotion Association Lexicon," *Computation and Language*, vol. 29 (3), pp. 436–465, 2013.

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

[42] L. Canales, C. Strapparava, E. Boldrini, and P. Martnez-Barco, "Intensional learning to efficiently build up automatically annotated emotion corpora," *IEEE Transactions on Affective Computing*, pp. 1–1, 2018.

[43] L. Canales, W. Daelemans, E. Boldrini, and P. Martínez-Barco, "Towards the improvement of automatic emotion pre-annotation with polarity and subjective information," in *Proceedings of the 11th biennial Recent Advances in Natural Language Processing conference (RANLP 2017)*, 2017.

[44] T. De Smedt and W. Daelemans, "Pattern for python," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 2063–2067, Jun. 2012.

[45] D. Ghazi, D. Inkpen, and S. Szpakowicz, "Hierarchical approach to emotion recognition and classification in texts," *Advances in Artificial Intelligence*, vol. 6085/2010, pp. 40–50, 2010.

[46] A. Kennedy and D. Inkpen, "Sentiment classification of movie reviews using contextual valence shifters," *Computational Intelligence*, vol. 22, p. 2006, 2006.

[47] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Nov. 2011.

[48] J. Fleiss *et al.*, "Measuring nominal scale agreement among many raters," *Psychological Bulletin*, vol. 76, no. 5, pp. 378–382, 1971.

[49] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng, "Cheap and Fast—but is It Good?: Evaluating Non-expert Annotations for Natural Language Tasks," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, ser. EMNLP '08. Stroudsburg, PA, USA: Association for Computational Linguistics, 2008, pp. 254–263.

[50] 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.

[51] 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.

[52] S. M. 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.

[53] E. Cambria, J. Fu, F. Bisio, and S. Poria, "Affectivespace 2: Enabling affective intuition for concept-level sentiment analysis," in *Proc. of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA.*, 2015, pp. 508–514.

[54] S. Rothe, S. Ebert, and H. Schütze, "Ultradense word embeddings by orthogonal transformation," *CoRR*, vol. abs/1602.07572, 2016.

[55] F. Enríquez, F. L. Cruz, F. J. Ortega, C. G. Vallejo, and J. A. Troyano, "A comparative study of classifier combination applied to nlp tasks," *Inf. Fusion*, vol. 14, no. 3, pp. 255–267, Jul. 2013.

[56] J. R. Landis and G. G. Koch, "The Measurement of Observer Agreement for Categorical Data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977.

**Lea Canales** obtained her PhD in Computer Science with a specialty in Natural Language Processing from the University of Alicante (2018). Currently, she is collaborating in Language Processing and Information Systems research group (GPLSI) at the University of Alicante. Her research interests are focused on identifying affective states from text. She is especially interested in the development of emotion resources and systems that allow improving the automatic emotion detection from the Web 2.0 text.

**Walter Daelemans** is a professor of Computational Linguistics at the University of Antwerp where he directs the CLiPS computational linguistics research group. His research interests are in machine learning of natural language, for example in the development of Memory-Based Language Processing (CUP, 2005); computational psycholinguistics, especially exemplar-based alternatives to mental rules as representations explaining language acquisition and processing; computational stylometry, with a focus on authorship attribution and author profiling from text; and language technology applications, for example biomedical information extraction and cybersecurity systems for social networks.

**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.

**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 Full 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).