Multimedia Tools and Applications (2023) 82:9871–9890
https://doi.org/10.1007/s11042-022-13750-x

1222: INTELLIGENT MULTIMEDIA DATA ANALYTICS AND COMPUTING

Contextual information usage for the enhancement of basic emotion classification in a weakly labelled social network dataset in Spanish

Juan Pablo Tessore[1,2] • Leonardo Martín Esnaola[1] • Hugo Dionisio Ramón[1] • Laura Lanzarini[3] • Sandra Baldassarri[4,5]

Received: 31 March 2021 / Revised: 6 July 2022 / Accepted: 31 August 2022 /
Published online: 15 September 2022

Abstract
Basic emotion classification is one of the main tasks of Sentiment Analysis usually performed by using several machine learning techniques. One of the main issues in Sentiment Analysis is the availability of tagged resources to properly train supervised classification algorithms. This is of particular concern in languages other than English, such as Spanish, where scarcity of these resources is the norm. In addition, most basic emotion datasets available in Spanish are rather small, containing a few

✉ Juan Pablo Tessore
juanpablo.tessore@itt.unnoba.edu.ar

Leonardo Martín Esnaola
leonardo.esnaola@itt.unnoba.edu.ar

Hugo Dionisio Ramón
hugo.ramon@itt.unnoba.edu.ar

Laura Lanzarini
laural@lidi.info.unlp.edu.ar

Sandra Baldassarri
sandra@unizar.es

[1] Instituto de Investigación y Transferencia en Tecnología (ITT) – (Centro CICPBA), Universidad Nacional del Noroeste de Buenos Aires, Junín, Argentina

[2] Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Ciudad Autónoma de Buenos Aires, Argentina

[3] Instituto de Investigación en Informática LIDI (Centro CICPBA), Facultad de Informática, Universidad Nacional de La Plata, La Plata, Argentina

[4] Departamento de Informática e Ingeniería de Sistemas, Universidad de Zaragoza, Zaragoza, Spain

[5] Instituto de Investigación en Ingeniería (I3A), Universidad de Zaragoza, Zaragoza, Spain

hundred (or thousand) samples. Usually, the samples only contain a short text (frequently a comment) and a tag (the basic emotion), omitting crucial contextual information that may help to improve the classification task results. In this paper, the impact of using contextual information is measured on a recently published Spanish basic emotion dataset and the baseline architecture proposed in the Semantic Evaluation 2019 competition. This particular dataset has two main advantages for this paper. First, it was compiled using Distant Supervision and as a result it contains several hundred thousand samples. Secondly, the authors included valuable contextual information for each comment. The results show that contextual information, such as news headlines or summaries, helps improve the classification accuracy over a dataset of distantly supervised basic emotion labelled comments.

Keywords Distant supervision · Basic emotion classification · Contextual information · Social media

1 Introduction

In recent years, with the massive adoption of social media platforms, people have been expressing their opinions about a wide variety of topics. Politics, economics, sports, business, and other topics are daily discussed in the comments sections of these platforms but reviews about products and services are also included [4, 8, 17, 34]. It is not uncommon for people to consider the feedback from other users before deciding to make a purchase. This phenomenon has created opportunities for governments and businesses which rely on this information respectively to enact better policies and to make better decisions based on the opinions of the market.

However, to exploit these opportunities, a particular set of skills is required, and thus research interest has thrived in areas such as Affective Computing (AC), Sentiment Analysis (SA), Natural Language Processing (NLP) and Machine Learning (ML). According to Cambria [10], AC and SA are key for the advancement of artificial intelligence and these areas can be used for the automated upkeep of product reviews, political issues, brand perception, as well as subsequent components of other systems such as customer relationship management and recommender systems.

Picard [45] defined AC as a field of cognitive computing and artificial intelligence for developing systems that can recognize, interpret, process, and simulate human emotions. Cambria et al. [12] defined SA as a research suitcase problem that requires tackling many NLP tasks, divided into three layers. The first is a syntactic layer that aims at preprocessing texts and includes part-of-speech tagging, lemmatization, and micro text normalization. The second is a semantic layer that aims at deconstructing the normalized text from the previous layer into concepts, resolves entities, and filters neutral content to improve sentiment classification accuracy. The tasks in this layer are, among others, concept extraction, word sense disambiguation, and subjectivity detection [16]. The last is the pragmatics layer, focused on extracting meaning from both sentence structure and semantics obtained from previous layers, and it includes tasks such as polarity and basic emotion detection, aspect recognition, sarcasm detection [36], and personality recognition [35].

Some of the tasks of the last layer, such as polarity and basic emotion detection, are handled differently depending on the approach selected for the research. Thakkar and Patel [56] identify three different approaches. The first relies on pre-compiled affective dictionaries, where each word is associated with a value denoting its correspondence with each class. The input text is scanned for words that are in the dictionary, and the final class of the text is calculated by performing some arithmetical operation with the value associated with the words. Another approach is based on ML, mainly with supervised variants. This approach comprises several stages, namely data collection, preprocessing, data tagging, and classification. Lastly, there is the hybrid approach, which uses the lexicon-based approach to pre-classify the documents, then these documents will represent the training data for the learn-based part [21].

The present work is focused on basic emotion classification with ML-based SA in the Spanish language. One of the main problems with this approach is the scarcity of reliable tagged datasets to train the algorithms used in languages different from English. Even though the availability of datasets in Spanish has been growing [20, 25, 37, 38, 40, 54], it cannot be compared with the number of resources available in English. According to Justo et al. [29], the majority of research in SA is addressed in English. Besides, manual tagging is costly in both time and resources, and usually the resulting datasets are rather small. This has led to the creation of a basic emotion dataset of social media comments in Spanish using Distant Supervision (DS) (i.e., where an already existing noisy label is linked to the content to build a tagged dataset automatically), and validated with the Fleiss Kappa metric over a small sample of the texts [55].

Some other researches, discussed in the following section, have used contextual information (CI) for the enhancement of ML classification; however, most rely on manually tagged datasets or corpus. The goal of this research is to measure and compare the results achieved with a DS tagged dataset using CI. The methodology described in this work allows fast and inexpensive dataset creation which, if the results are similar to those achieved with manual tagging, would help to deal with the scarcity of resources in languages different from English. Furthermore, most studies measure the impact of CI for polarity detection, but only a few perform basic emotion recognition, as will be explained in detail in the next section.

The rest of the paper is organized as follows. In section 2 the literature concerning text classification with CI is reviewed. Section 3 describes the characteristics of the dataset selected for the task and shows some additional metrics. Section 4 explains the experimental setup used in this research, and the experiments. In section 5 the results are presented. Then, section 6 contains the discussion. Finally, section 7 sets out the conclusions and proposals for future work.

2 Background work

Emotion detection in a text is a challenging task as the text format lacks other attributes that may ease the procedure. According to [15], the absence of facial expressions and voice modulations can make the task difficult even for humans, not to mention machines. For example, the phrase “I almost cried” may be interpreted as sadness (or with negative polarity), but if the text is preceded by “I received the gift I so much desired”, then the basic emotion can be happiness (or with positive polarity). Automatic text classification performs well when the context of a short message is extended with knowledge extracted using large collections [42].

Yusof et al. [67] state that sentiment classification is one of the most challenging tasks of NLP because the connotation of sentiment is highly dependent on the context of the text, and that it is imperative to incorporate context in SA because the content by itself may be misleading. This has led many researchers in SA to search for additional information within the text (and some studies using other formats) that may give hints about its connotation and help improve classification results [15].

In [15] a dataset of 38,424 (train + test) text dialogues was compiled and the dialogues were manually tagged into four classes (Happy, Sad, Angry, and Others) by a group of judges. Each content was tagged by 7 people. The study measured the agreement of the dataset by using the Fleiss Kappa metric [23]. The result was 0.58 for the training set and 0.59 for the test set, thus in the moderate agreement region. Something that must be highlighted is that the class "Others" may have helped to improve the score as all the challenging content probably fell into this category. A baseline model using a Long short-term memory (LSTM) recurrent neural network (RNN) achieved a micro F1 score of 0.5861 for three classes (Happy, Sad, Angry). Several research teams [2, 5, 6, 28, 32, 65, 66] participated in SemEval-2019 Task 3 achieving a micro F1 score of 0.7959. However, the highest score for the team that also submitted a paper describing the architecture was 0.7765 [15].

Poria et al. [47] used a model based on RNN (contextual attention-based LSTM) to capture CI among textual, audio, and video utterances. The model using CI showed about 6–8% improvement (81.3% of accuracy) over the state-of-the-art benchmark using the CMU-MOSI Dataset [68]. A refined model of this approach was presented in [27], improving the accuracy to 82.31%.

In the work of Agarwal et al. [1], the authors measured the impact on the classification accuracy over three polar text datasets (software, movies, and restaurant reviews) using (individually and combined): a domain-specific ontology; feature importance; and CI. With regard to the latter, a contextualized sentiment lexicon was built to determine the polarity of ambiguous terms, based on the context in which they appear (context terms). This task was performed using SenticNet [11], SentiWordNet [3], and General Inquirer [53]. The results showed that CI was individually the most impactful addition, as it achieved the most significant accuracy improvement over the three used datasets.

Muhammad et al. [41] built domain-specific lexicons using a distantly supervised approach which, combined with SentiWordNet [3], produced modified sentiment scores (valence + or -) for the terms analyzed (global context). This was combined with a window-based approach in which lexical and non-lexical modifiers are used for term valence within a specific text window (local context). The system built outperformed the baseline in two of the three datasets tested. In a similar approach, Saif et al. [51] used the previous polarity (assigned using SentiWordNet [3], MPQA [64], and Thelwall lexicons [57, 58]) of co-occurrent (context) terms to build what they called a "Sentic circle", which could be later used to determine valence and polarity of a word. These Sentic circles were later used for polarity detection on tweets. The results showed that the approach beat the baseline (SentiStrength) in two out of three datasets used.

Vosoughi et al. [63] also used a distantly supervised approach to collect a dataset of 18 million tweets, which were tagged as positive or negative according to the presence of specific emoticons and later validated using the Fleiss Kappa interrater agreement measure [23]. Then the researchers used CI (such as geolocation, post time and author of the content) to calculate prior probabilities of negative and positive sentiments using a Bayesian model. The accuracy of the model that used all the contextual features was 0.862, an improvement of 0.077 over the baseline (0.785). Also, Vanzo et al. [59] used a set of tweets as context by in the first place obtaining the n preceding tweets in a conversation and in the second place obtaining the n preceding tweets that shared a specific hashtag with the target tweet. The dataset used was the one provided in SemEval-2013 Task 2 [43], and the classifier adopted was a customized support vector machine (SVM). The study achieved an accuracy improvement in almost every experiment conducted; the highest improvement was about 5.2%. The authors concluded that the improvements achieved with the usage of CI are striking as the applicability of their approach does not require additional manually tagged resources.

As seen in this section, most of the papers analyzed are focused on polarity detection. Only the papers presented in SemEval-2019 Task 3 [2, 5, 6, 28, 32, 65, 66] used CI to enhance basic emotion detection. However, the studies described in the papers relied on a manually tagged dataset while in the current approach a basic emotion dataset was collected by using DS and validated by manually classifying a small sample and then measuring its Fleiss Kappa interrater agreement, which resulted in a larger dataset [55]. It is also important to note that the dataset used in SemEval-2019 Task 3 consisted of textual dialogues while the dataset used in the current paper contains comments in response to a certain topic. However, these interactions can be seen as a particular form of dialogue as the users "respond" with what they think or feel about the topic discussed.

It is also relevant to remark that most of the papers presented in this section rely on tools available only for the English language. This presents a difficulty when trying to replicate these studies for other languages, such as Spanish. Nevertheless, the interest in basic emotion classification in Spanish has been growing. For example, in IberLef EmoEvalEs-2021 [46], a competition of emotion classification in Spanish, 70 teams registered, 15 submitted results and 11 presented papers describing their systems [18, 19, 24, 26, 31, 33, 48, 49, 52, 60, 61]. In addition, this competition provided a form of CI for each sample of the used dataset, composed of tweets in Spanish, by adding the domain associated and whether the tweet expressed offensiveness. However, the dataset used in the competition was manually classified by annotators, which explain its small size.

The current approach relies only on automatic or semi-automatic procedures and can be used as a guide for basic emotion classification enhancement through the use of CI for low resource languages.

3 General architecture and dataset characteristics

The general purpose of this research is to measure the impact of CI on the performance achieved in a basic emotion classification task for the Spanish language and also to compare the results achieved by a DS compiled dataset versus a manually classified one. Figure 1 shows the different stages of the research. Some of these, mainly the initial ones, were developed previously [55] but they are briefly described in subsection 3.1. Subsection 3.2 summarizes some metrics of the dataset used. Lastly, subsection 3.3 explains the embedding format and the classification algorithm selected.

3.1 Previously developed stages

Since a proper dataset for this experiment could not be found, it was decided to build one from scratch. The construction process is explained in detail in [55]; however, the steps involved are briefly described in Table 1.

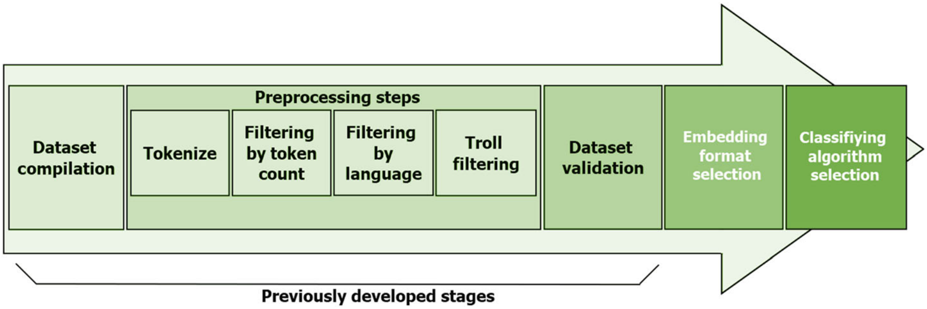


Fig. 1 Dataset construction and evaluation process

3.2 Dataset characteristics

The work described in the previous subsection enabled the possibility to build a dataset of 1,020,557 comments, each one accompanied by the post's title and description, i.e., the CI of the comment. Table 2 shows some important dataset metrics.

It is important to remark that previous to the dataset's publication, no similar alternative could be found, i.e., no basic emotion dataset for the Spanish language of this size also validated by

Table 1 Brief description of the previously developed stages

Step		Brief description
Dataset compilation		Comments and reactions were collected from Facebook. These comments and reactions were taken from the interactions of many different Facebook users that posted across 13 widely read news portals. The extraction process of comments, reactions, and posts was performed using the Facebook API Graph tool. The extraction period was set to 4 years, between January 2016 and December 2019.
Preprocessing	Tokenize	The Tweet Tokenizer class, from the NLTK library, was used for this step. The text was tokenized in order to filter unnecessary tokens like links, signs, non-printable characters, and stop words.
	Filtering by token count	Comments tokenized as described in the previous step and containing less than three valid tokens were excluded from the dataset.
	Filtering by language	Python Bindings to CLD2 library and Google trans were used to filter comments in languages other than Spanish.
	Troll filtering	Users that repeated the same comment across several different posts were identified as trolls and filtered from the dataset.
Dataset validation		This involved taking a small sample of comments from the dataset and having them classified manually by a group of psychologists. The Fleiss Kappa interrater agreement measure was later calculated among the automatically and manually acquired tags. All measures were in the moderate agreement zone of the scale, and thus suitable for training ML based classifiers in the SA field.

Table 2 Relevant metrics and characteristics of the dataset

Feature	Observations
Number of samples (title, new, comment, reaction)	1,020,557
Samples of ANGRY reaction	436,357
Samples of HAHA reaction	338,835
Samples of LOVE reaction	159,830
Samples of SAD reaction	85,535
Total vocabulary size (lowercase and without punctuation)	322,291
News' vocabulary size (lowercase and without punctuation)	81,504
Titles' vocabulary size (lowercase and without punctuation)	39,581
Comments' vocabulary size (lowercase and without punctuation)	281,277
News' average, min and max length (by word)	(22.5; 0; 388)
Titles' average, min and max length (by word)	(12.65; 0; 41)
Comments' average, min and max length (by word)	(19.36; 3; 1218)

specialists. For that reason, the authors believed that the dataset and the construction process that it involved were valuable contributions in themselves, and thus they were published in a different article. The process needed very little manual tagging and could be used as a guide to build such resources for other languages as well.

3.3 Embedding and classification algorithm selection

With regard to the feature representation format, it has been long studied and established that traditional approaches involving sparse vectors like Bag of Words fail to capture information related to the meaning of the word, and have limitations for short text messages such as tweets [42]. This also applies to the bigram/trigram of words/n-characters, etc. Probably for this reason, many studies [2, 5, 6, 28, 32, 65, 66] use neural embeddings as the representation format, from which the most commonly adopted ones are Word2Vec [39], FastText [9], and GloVe [44]. The latter representations have the advantage that they codify the context of a certain word in its embedding.

The original publications for neural embeddings were designed for the English Language. However, over the years such resources have also become available for the Spanish Language [13, 14, 22].

On the other hand, the most common classification algorithms in the literature are SVM and RNN, a specific type of Artificial Neural Network (ANN). Also, it is important to note that the most recent literature is leaning towards variants of RNN, being LSTM the most common [15].

ANNs are based on analogous neural structures found in the brain of living beings. The neuron or node model is based on the following formula:

$$y = g\left(\sum_{i=1}^{n} w_i a_i + b\right)$$

where g acts as an activation function, b is a term that is incorporated called bias, a_i is an input signal and w_i is the weight associated with the previous signal, the above formula corresponds to a neuron with n inputs.

The neurons or units described are grouped in layers and linked to form a neural network. Each layer can have one or more neurons. Traditional ANNs have three layers: an input layer, a hidden layer, and finally an output layer. In the event that the ANN has more than one hidden layer, then it will be a deep neural network.

A special type of deep neural network is the RNN [50]. The RNN allows information from sequential data to be obtained and processed. RNNs are particularly useful for natural language processing (NLP) because they allow capturing the sequential and temporal dependencies of the input data.

RNNs aggregate cycles that connect adjacent nodes and act as a kind of network memory that is used to incorporate data from the past in evaluating the properties of current data. A diagram of these networks can be seen in Fig. 2.

The output of a network node is a function of its input x_t and the historical data received by h_{t-1}. The formula is as follows:

$$h_t = f(h_{t-1}, x_t)$$

A problem that affects RNNs in general is that of gradient vanishing. This makes this type of network have problems remembering patterns that are very extended in time because as the network processes more elements in the chain, it becomes more difficult to recall past information.

To solve this problem, recurrent neural networks for long-short term memory (LSTM) were developed. The novelty of these nodes is that they incorporate three gates: input, forget and output. A diagram of an LSTM cell and its gates can be seen in Fig. 3.

The forget gate is responsible for remembering only some parts of the long-term memory and decides what to remember based on the current input and memory received from the previous step. On the other hand, the input gate remembers only some parts of the current input and previous working memory, and decides what to remember based on the current input and memory received from the previous step. Based on the two previous gates, the long-term memory is updated. Finally, the output gate decides which parts of short-term memory will be remembered and passed to the next iteration.

To conclude this subsection, it is also necessary to remark that one important factor to determine the performance of the system is to establish a baseline to compare it with. Among all the reviewed papers, the most similar one is presented in SemEval-2019 Task 3 [15]. The mentioned work also uses CI for basic emotion classification; however, it does not compare the system behavior with and without CI, and also the dataset used was manually classified and contains a limited number of samples (38,424 train + test). Therefore, in the next section, as Experiment 2, the authors of this study establish a baseline for SemEval-2019 Task 3, i.e., without CI, by recreating their experiment. This recreation is referred to as SemEval-2019 Task 3*.

The SemEval-2019 Task 3′ dataset consists of textual dialogues. The one used in the present paper contains comments in response to posts in social media. However, the authors do not believe this to be a major issue as these responses are a form of communication or dialogue.

In the following section, the compiled and preprocessed DS dataset, the embeddings provided by the Spanish Billion Words Corpus project, and a LSTM based classifier will be

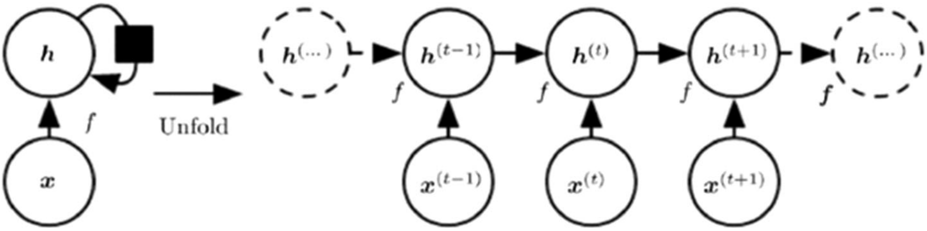


Fig. 2 Recurrent neural network diagram [7]

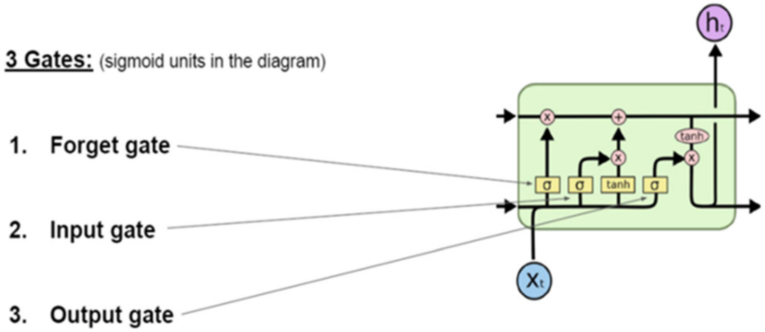


Fig. 3 LSTM cell indicating its input, forget and output gates [62]

combined to measure and compare the effects of the usage of CI in developing a ML based basic emotion classifier.

4 Experimental setup

Two experiments were carried out to measure the impact of CI on a basic emotion classification task over social media content and also to establish how a DS dataset behaves compared to a manually classified counterpart. For these experiments, the baseline model configuration proposed in [15] was adopted. It is necessary to remark that the objective of this study is to measure how much CI improves the classifier behavior rather than building the best classifier. The embedding format used was Glove [44] and the classification algorithm selected was LSTM. This was implemented using Keras [30].

For the neural embeddings for the Spanish language, the Glove embeddings provided by the Spanish Billion Words Corpus were selected. They were computed from a dataset of 1.4 billion Spanish words, containing 855,380 vectors of 300 dimensions each. These vectors were used to build a non-trainable embedding layer. The max number of words for the embeddings was limited to 20,000.

The preceding layer was then connected with a 128-dimension LSTM layer with a 0.2 dropout to prevent overfitting. Lastly, a dense layer was added with the sigmoid function as activation. The loss function was established as categorical cross entropy. A diagram of the architecture used can be seen in Fig. 4, where “None” should be replaced with the number of the samples in the dataset.

The max length for the input was established as 100 words, which is enough for 95% of the samples comprising a news title, a description, and a comment. If all the comments had to be covered, the max length would have been increased significantly and the benefits from this action would not have outweighed the disadvantages in both system complexity and processing time.

For both experiments, the system was trained in 5 folds. In each of the folds executed, the training set was further divided into 80% training and 20% validation. After this cross-validation process the classifier was later trained with all this data (training and validation). The batch size in all cases was established at 200 samples.

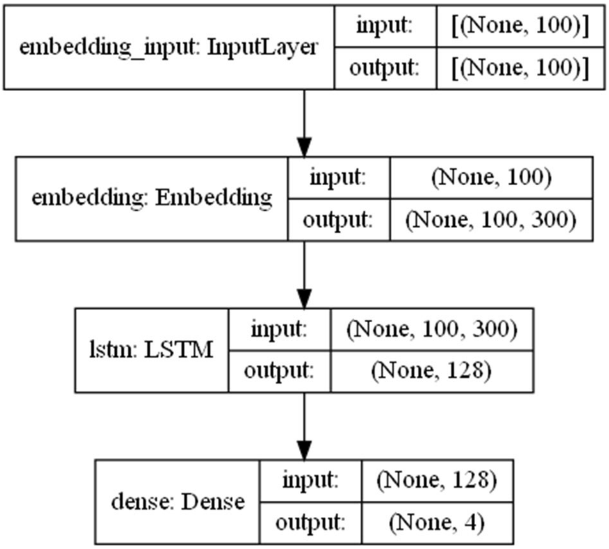


Fig. 4 Architecture of the classifier

4.1 Experiment 1: Effect of considering CI

The goal of this experiment was to measure the impact of CI on the performance of a basic emotion classifier for the Spanish language, using dataset [55]. For this experiment, the classifier described in the previous section was trained two times. First, the samples were composed of comments, each associated with a particular basic emotion, i.e., without CI. Then, a news title and a description, i.e., the CI, were added to the samples.

The classes were balanced by performing under-sampling. The dataset used was randomly split into a training set of 273,712 samples and a test set of 68,428.

4.2 Experiment 2: Comparison against other studies' results

To measure the improvement of the classifier performance through the incorporation of CI, and to establish a comparative base, a baseline was constructed using the datasets with and without CI for both (SemEval-2019 Task 3 and this study).

It should be noted that [15] does not report the classifier performance without CI. However, since the dataset used is available, the parts representing CI are recognizable, and the architecture of the model is described, the authors of the present paper recreated those results, using and without using CI. As previously mentioned, these results are referred to as SemEval-2019 Task 3*, to point out that they do not correspond faithfully to those reported in the original work.

The comparison between both studies is made by measuring, and later comparing, the percentage of improvement obtained in each performance metric evaluated when incorporating CI. It is impossible to produce a direct correlation, since [15] uses a different dataset, English writing, and other embeddings.

Table 3 Validation results without CI

		Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Class ANGRY	Precision	0.453	0.450	0.458	0.441	0.456
	Recall	0.490	0.471	0.477	0.508	0.456
	F1	0.471	0.460	0.468	0.472	0.456
Class SAD	Precision	0.604	0.589	0.595	0.592	0.602
	Recall	0.555	0.576	0.558	0.565	0.563
	F1	0.579	0.583	0.576	0.578	0.582
Class HAHA	Precision	0.518	0.506	0.526	0.527	0.503
	Recall	0.529	0.557	0.522	0.517	0.595
	F1	0.523	0.530	0.524	0.522	0.545
Class LOVE	Precision	0.591	0.619	0.580	0.609	0.622
	Recall	0.577	0.542	0.597	0.557	0.547
	F1	0.584	0.578	0.588	0.582	0.582
All Classes	Accuracy	0.5377	0.5364	0.5385	0.5369	0.5504
	Macro Precision	0.5413	0.5411	0.5399	0.5423	0.5458
	Macro Recall	0.5378	0.5364	0.5384	0.5368	0.5403
	Macro F1	0.5396	0.5388	0.5392	0.5395	0.5431

Besides, to be able to work with datasets comparable in size and class distribution, even though they are different, the dataset [55] was subsampled to match the size of the dataset used in SemEval-2019 Task 3, given that its size is significantly larger.

5 Results

For experiment 1, Table 3 presents the metrics, in each of the folds, on the validation data for the classifier, trained without CI (i.e., comments only). Table 4 shows the same but in this case the classifier was trained with CI.

The metrics for the test data of the classifier, trained with and without CI, can be seen in Table 5 and Fig. 5.

Table 4 Validation results with CI

		Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Class ANGRY	Precision	0.565	0.552	0.540	0.567	0.530
	Recall	0.505	0.529	0.578	0.484	0.562
	F1	0.534	0.540	0.558	0.522	0.545
Class SAD	Precision	0.713	0.733	0.736	0.729	0.733
	Recall	0.706	0.692	0.677	0.689	0.687
	F1	0.709	0.712	0.706	0.708	0.709
Class HAHA	Precision	0.623	0.608	0.623	0.589	0.617
	Recall	0.613	0.648	0.627	0.669	0.633
	F1	0.618	0.627	0.625	0.627	0.625
Class LOVE	Precision	0.663	0.720	0.705	0.687	0.713
	Recall	0.751	0.723	0.708	0.732	0.695
	F1	0.704	0.711	0.707	0.709	0.674
All Classes	Accuracy	0.6435	0.6481	0.6474	0.6433	0.6444
	Micro Precision	0.6409	0.6482	0.6512	0.6430	0.6480
	Micro Recall	0.6437	0.6478	0.6475	0.6436	0.6443
	Micro F1	0.6423	0.6480	0.6493	0.6433	0.6462

Table 5 Test results with and without CI

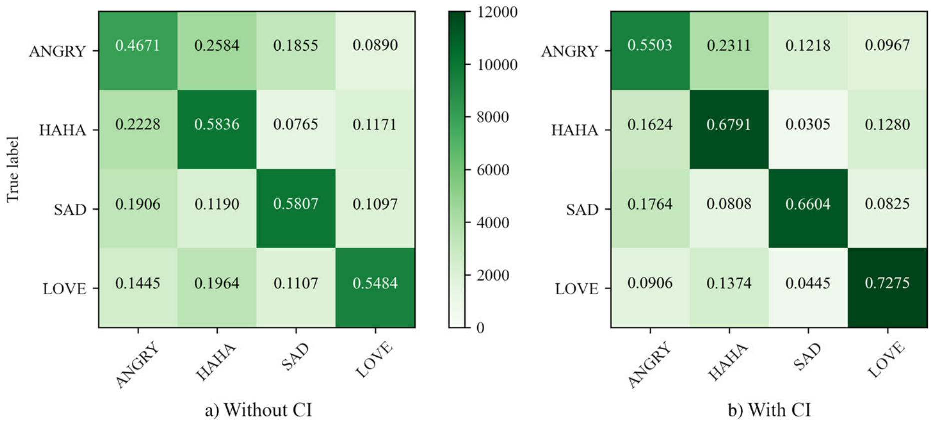
		ANGRY	HAHA	SAD	LOVE	ALL CLASSES
Precision	Without CI	0.456	0.503	0.608	0.637	0.5509
	With CI	0.562	0.601	0.77	0.705	0.6593
Recall	Without CI	0.467	0.584	0.581	0.548	0.5449
	With CI	0.55	0.679	0.66	0.727	0.6543
F1	Without CI	0.461	0.54	0.594	0.589	0.5479
	With CI	0.556	0.637	0.771	0.716	0.6568
Accuracy	Without CI	0.4671	0.5836	0.5807	0.5484	0.5449
	With CI	0.5503	0.6791	0.6304	0.7275	0.6544

As can be seen in the figures and tables in this section, the use of CI improves almost every performance metric of the classifier as the individual and general precision, recall and F1 scores all increased.

For experiment 2, as stated in the previous section, the classifier was first trained and tested with a subsample of the dataset (to match the size of the dataset used in SemEval-2019 Task 3). The test results, with and without CI, are presented in Table 6 for both SemEval-2019 Task 3* and the classifier built in this study. This table shows the percentage of improvement achieved by using CI in order to establish a comparison base between the studies. Figure 6 gives more details with the confusion matrix.

6 Discussion

With experiment 1 this study compares the influence of CI in the process of building a LSTM based emotion classifier. Tables 3 and 4 display consistent results of accuracy, precision, recall and F1 for each fold and class. In addition, the performance measurements shown for the classifier trained with CI outperforms the one trained without it. This is also the case for test results which are displayed in Table 5, showing a performance gain around 10% for every metric.



a) Without CI

b) With CI

Fig. 5 Confusion matrix for test results

Table 6 Test results for the classifier trained with the subsampled dataset

	SemEval-2019 Task 3*			This study		
	Without Context	With Context	Improvement (%)	Without Context	With Context	Improvement (%)
Accuracy	0.8403	0.8448	0.5355	0.4742	0.5444	14.8039
Precision	0.4755	0.4719	−0.7571	0.4783	0.5570	16.4541
Recall	0.6689	0.7215	7.8637	0.4740	0.5437	14.7046
F1	0.5559	0.5706	2.6444	0.4762	0.5503	15.5607

The magnitude of performance improvement with CI in accuracy, precision, recall and F1 is also visible in Figs. 7 and 8. The first presents the metrics for all classes and the second desegregated by metric and class.

Table 7 shows some test samples categorized correctly with the classifier trained with CI and wrongly with the one trained without it. Some of the samples are sarcastic, for example comments A, F and G, while others do not include enough information in the comment for the classifier to discern them correctly, for example B, C, D and E. A more extensive list of this kind of sample is included as Electronic Supplementary Material.

Experiment 2 subsampled the dataset used in this study to match the size of the one used in SemEval-2019 Task 3 and compared the performance gain obtained by training a classifier with CI for both. Figure 9 shows that, while the SemEval-2019 Task 3* classifier obtained little to no gain with CI in almost every metric for the corresponding dataset, its use had quite an impact for the dataset used in this article, as a significant improvement was achieved for every metric.

7 Conclusions and future work

The experiments performed showed that the use of CI produced an improvement in the metrics both individually for each class and globally. As ML based classifiers can benefit from CI,

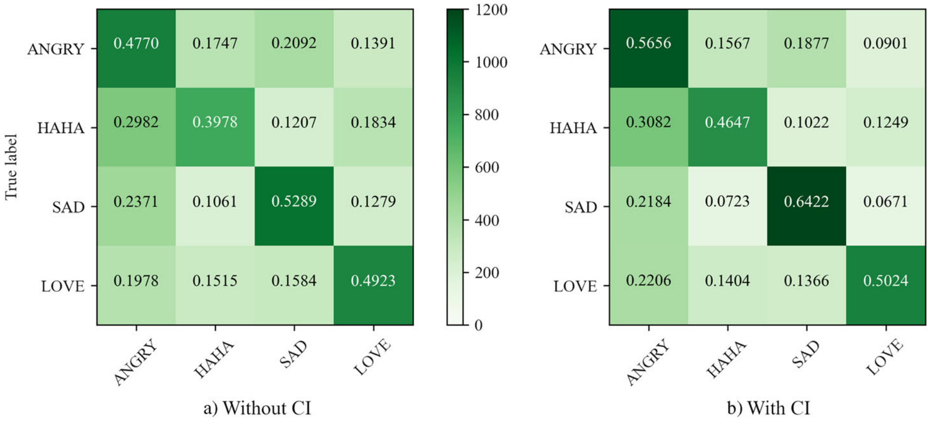


Fig. 6 Confusion matrix for the test (subsampled dataset)

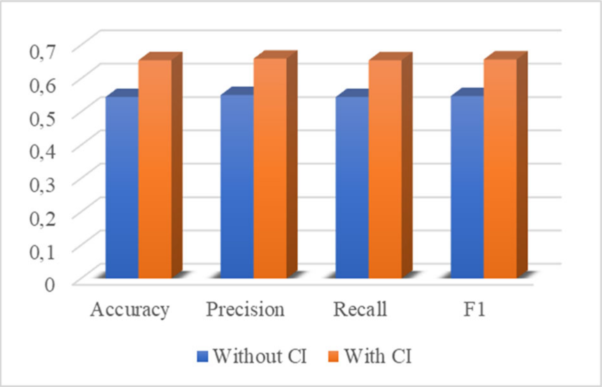


Fig. 7 Metrics for the classifier with and without CI

researchers who are planning to build basic emotion datasets should consider also capturing this data which is generally available but usually ignored.

Experiment 2 showed that the performance gain of CI in a dataset of social media titles, subtitles and comments proved to be higher than that obtained for a dataset of textual dialogues, at least for the datasets analyzed in this study. Moreover, as Table 6 shows, the F1 scores for both classifiers trained with CI are similar, and it is important to point out that the dataset used in SemEval2019-Task 3 was manually annotated while that used in this study was built semi-automatically using DS.

It is also important to point out that the classification task described in this article was more difficult than the task in the SemEval-2019 Task 3 competition, as it required classifying 4 classes instead of 3. In addition, the SemEval-2019 Task 3 dataset included the category “others”, which usually contains the more challenging content to be classified in a category.

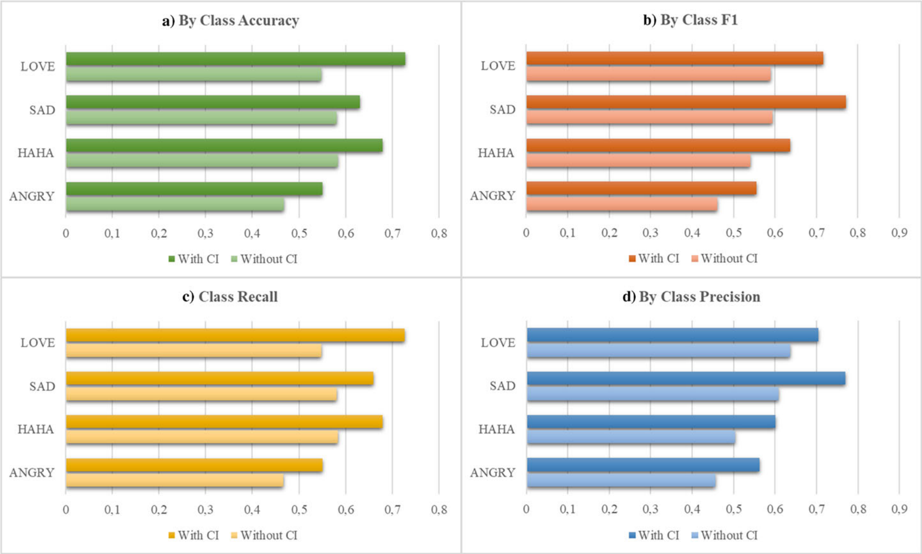


Fig. 8 Metrics for the classifier with and without CI by classes

Table 7 Correctly classified samples using CI

Sample				Label w/o CI	Actual label
A	**CI**	**Title**	*“Pogba me admira mucho”* In English: “Pogba admires me very much”	LOVE	HAHA
		Subtitle	*“POGBA ME ADMIRA MUCHO” Carlitos Tevez contó que, cuando compartían vestuario en Juventus, el propio Paul le contaba que era gran fanático suyo.* In English: “POGBA ADMIRES ME VERY MUCH” Carlitos Tevez said that, when they shared a dressing room at Juventus, Paul himself told him that he was a big fan of his.		
	Comment		*Que humilde es este pibe....Pogba digo...* In English: What a modest guy.... Pogba I mean...		
B	**CI**	**Title**	*Misiones: ordenaron la detención de un perro que ahora se encuentra “prófugo de la Justicia” - TN.com.ar* In English: Misiones: They ordered the arrest of a dog that is now a “fugitive from Justice” - TN.com.ar	ANGRY	HAHA
		Subtitle	*¡Un prófugo inesperado! La polícia de Wanda, en Misiones, busca intensamente a un perro por cometer un delito: se lo acusa de haber comprometido “la seguridad física de las personas” que transitan por el centro de la localidad* In English: An unexpected fugitive! The Wanda police, in Misiones, are carrying out an intensive search for a dog for committing a crime: it is accused of endangering “the physical safety of people” who pass through the center of the town		
	Comment		*‘Por eso la justicia esta como esta. Que pavada Dios’* In English: ‘That’s why justice is the way it is. My God, what a load of nonsense’		
C	**CI**	**Title**	*Murió el humorista Martín Rocco - TN.com.ar* In English: The humorist Martín Rocco has died – TN.com.ar	HAHA	SAD
		Subtitle	*Murió el humorista Martín Rocco.* In English: The humorist Martín Rocco has died		
	Comment		*Hacía stand up! Un geniooo. Me reía mucho con su monólogo del gato. Buen viaje* In English: He did stand up comedy! A real genius. I laughed a lot with his cat monologue. Have a good trip		
D	**CI**	**Title**	*Juan Martín se vio superado por Djokovic y la derrota fue fuerte.* In English: Juan Martín was beaten by Djokovic and it was a heavy defeat.	LOVE	SAD
		Subtitle	*EL LLANTO DE DELPO Juan Martín se vio superado por Djokovic y la derrota fue fuerte.* In English: THE TEARS OF DELPO Juan Martín lost to Djokovic and it was a heavy defeat.		
	Comment		*Dejaste todo nene!!! Sos un grande* In English: You left everything kid!!! You are really great		
E	**CI**	**Title**	*“Está realmente mal”* In English: “It’s really bad”	LOVE	SAD

Table 7 (continued)

Sample				Label w/o CI	Actual label
		Subtitle	*Es uno de los buzos más experimentados del mundo; trabaja en el rescate de los niños atrapados en una cueva de Tailandia y así contó cómo es la misión más extrema que tuvo en su vida.* In English: He is one of the most experienced divers in the world; he took part in the rescue of children trapped in a cave in Thailand and told how it was the most extreme mission he had in his life.		
	Comment		*Dios bendeci a chicos profe y rescatistas.amén* In English: God bless the kids, the teacher and the rescuers. Amen		
F	**CI**	**Title**	*El ministro de Transporte, Guillermo Dietrich, anuncia aumentos de colectivos y trenes* In English: Transport Minister Guillermo Dietrich announces increases in bus and train fares	LOVE	ANGRY
		Subtitle	*El ministro de Transporte, Guillermo Dietrich, anuncia aumentos de colectivos y trenes* In English: Transport Minister Guillermo Dietrich announces increases in bus and train fares		
	Comment		*Que lindo cuanto los queremos!!* In English: How cute! Don't we just love them!!!		
G	**CI**	**Title**	*YPF aumentó las naftas 4,5 por ciento en todo el país - TN.com.ar* In English: YPF increased gasoline 4.5% throughout the country – TN.com.ar	HAHA	ANGRY
		Subtitle	*Ahora, YPF: se sumó a los aumentos y el litro de nafta súper, por ejemplo, cotiza a $26,35 en Capital. En el interior, la premium está muy cerca de los $30* In English: Now, YPF: it has joined the increases and a liter of super gasoline, for example, is trading at $26.35 in Capital. In the interior, premium gasoline is very close to $30		
	Comment		*Querian un cambio ahi tienen el cambio jajajaj* In English: They wanted a change there and now they've got one hahaha		

In addition, the larger number of samples available in the dataset used allowed the classifier trained with CI to outperform the F1 score obtained in SemEval-2019 Task 3*. As this dataset was compiled semi-automatically with DS, the recompilation process can serve as a guide for researchers working with low resource languages not only for building greater datasets, but also for constructing more robust basic emotion classifiers.

The next steps in this research will be: tuning the classifier using in part the systems presented in the SemEval-2019 Task 3 competition as suggestions [2, 5, 6, 28, 32, 65, 66] and verifying whether it behaves similarly, testing other types of CI to enhance the classification process as shown in the articles discussed in the background work, and lastly verifying whether the behavior shown is also present in other types of datasets. The use of semantic information should also be explored, as it may help improve classification accuracy [31].

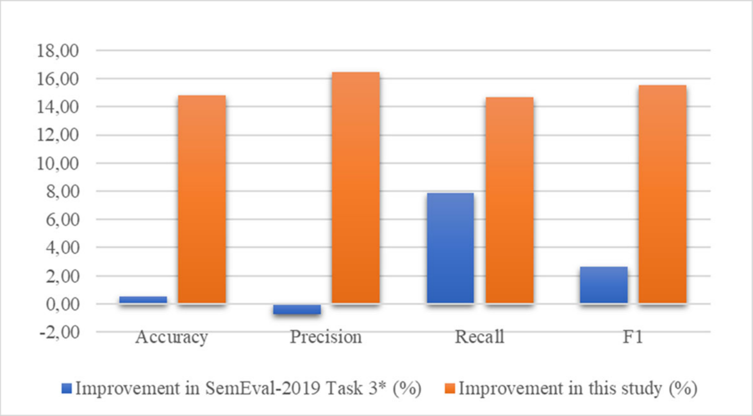


Fig. 9 Improvement comparison between SemEval-2019 Task 3* and this study

This study, however, has some limitations. First, relating to the use of CI in DS datasets, as the tags are already present in the data the compilers cannot choose how to categorize the content. In addition, further research should be performed to prove the consistency of this kind of CI enhancement with other types of datasets.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11042-022-13750-x.

Acknowledgments This publication was supported by public funds, and the authors would like to thank the following institutions and/or agencies:

Instituto de Investigación y Transferencia en Tecnología (Centro CICPBA), Universidad Nacional del Noroeste de Buenos Aires, Argentina.

Secretaría de Investigación, Desarrollo y Transferencia de la Universidad Nacional del Noroeste de Buenos Aires, Argentina.

Gobierno de Aragón, a través del proyecto AffectiveLab-T60-20R, España.

Instituto de Investigación en Informática LIDI (Centro CICPBA), Facultad de Informática, Universidad Nacional de La Plata, Argentina.

Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina.

Comisión de Investigaciones Científicas de la Provincia de Buenos Aires (CIC-PBA), Argentina.

Ministerio de Ciencia, Innovación y Universidades (MCIU), España (contrato RTI2018-096986-B-C31).

Code availability Not relevant.

Funding This study was funded by:

Consejo Nacional de Investigaciones Científicas y Técnicas, Argentina, (RESOL-2020-131-APN-DIR#CONICET).

Convocatoria para la Acreditación de Proyectos y Solicitud de Subsidios de Investigación Bianuales 2019, Secretaría de Investigación Desarrollo y Transferencia, Universidad Nacional del Noroeste de la Provincia de Buenos Aires, Argentina, (EXP 536–2019).

Gobierno de Aragón, España, (AffectiveLab-T60-20R).

Ministerio de Ciencia, Innovación y Universidades, España, (contrato RTI2018–096986-B-C31).

Data availability Not relevant.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethics approval Not relevant.

Consent to participate Not relevant.

Consent for publication Not relevant.

References

1. Agarwal B, Mittal N, Bansal P, Garg S (2015. Epub ahead of print 2015) Sentiment analysis using common-sense and context information. Comput Intell Neurosci 2015:715730. https://doi.org/10.1155/2015/715730
2. Agrawal P, Suri A (2019) NELEC at SemEval-2019 Task 3: Think Twice Before Going Deep. In: May J, Shutova E, Herbelot A, Zhu X, SMM MA (eds) Proceedings of The 13th International Workshop on Semantic Evaluation (SemEval-2019). Association for Computational Linguistics, Minneapolis, pp 266–271
3. Baccianella S, Esuli A, Sebastiani F (2006) SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. In: Proceedings of the 7th Conference on Language Resources and Evaluation LREC10, pp. 417–422
4. Bae Y, Lee H (2012) Sentiment analysis of twitter audiences: measuring the positive or negative influence of popular twitterers. J Am Soc Inf Sci Technol 63:2521–2535
5. Bae S, Choi J, Lee S (2019) SNU IDS at SemEval-2019 Task 3: addressing training-test class distribution mismatch in conversational classification. In: Proceedings of the 13th International Workshop on Semantic Evaluation. Association for Computational Linguistics, Stroudsburg, pp 312–317
6. Basile A, Franco-Salvador M, Pawar N et al (2019) SymantoResearch at SemEval-2019 Task 3: combined neural models for emotion classification in human-chatbot conversations. In: Proceedings of the 13th International Workshop on Semantic Evaluation. Association for Computational Linguistics, Stroudsburg, pp 330–334
7. Goodfellow I, Bengio,Y, Courville A (2016). Deep learning. MIT Press, Cambridge
8. Bi J-W, Liu Y, Fan Z-P, Cambria E (2019) Modelling customer satisfaction from online reviews using ensemble neural network and effect-based Kano model. Int J Prod Res 57:7068–7088
9. Bojanowski P, Grave E, Joulin A, Mikolov T (2017) Enriching word vectors with subword information. Trans Assoc Comput Linguist 5:135–146
10. Cambria E (2016) Affective computing and sentiment analysis. IEEE Intell Syst 31:102–107
11. Cambria E, Havasi C, Hussain A (2012) SenticNet 2: A semantic and affective resource for opinion mining and sentiment analysis. In: Proceedings of the 25th International Florida Artificial Intelligence Research Society Conference, FLAIRS-25. pp. 202–207
12. Cambria E, Poria S, Gelbukh A, Thelwall M (2017) Sentiment analysis is a big suitcase. IEEE Intell Syst 32:74–80
13. Cañete J (2019) FastText at Bot Center repo. Retrieved March 31, 2021, from https://github.com/BotCenter/spanishWordEmbeddings
14. Cardellino C (2016) Spanish Billion Words Corpus and Embeddings. Retrieved March 31, 2021, from https://crscardellino.github.io/SBWCE/
15. Chatterjee A, Narahari KN, Joshi M et al SemEval-2019 Task 3: EmoContext Contextual Emotion Detection in Text. In: May J, Shutova E, Herbelot A et al (eds) Proceedings of the 13th International Workshop on Semantic Evaluation. Association for Computational Linguistics, Stroudsburg, pp 39–48
16. Chaturvedi I, Cambria E, Vilares D (2016) Lyapunov filtering of objectivity for Spanish sentiment model. In: 2016 International Joint Conference on Neural Networks (IJCNN). IEEE, pp 4474–4481
17. Chen L, Qi L (2011) Social opinion mining for supporting buyers' complex decision making: exploratory user study and algorithm comparison. Soc Netw Anal Min 1:301–320
18. Chiruzzo L, Rosá A (2021) Retuyt-inco at emoevales 2021: multiclass emotion classification in spanish. CEUR Workshop Proc 2943:72–77
19. De Arriba A, Oriol M, Franch X (2021) Applying sentiment analysis on spanish tweets using beto. CEUR Workshop Proc 2943:86–93

20. Díaz-Galiano MC, García-Vega M, Edgar C et al (2019) Overview of TASS 2019 : One more further for the global Spanish sentiment analysis Corpus. Proc Iber Lang Eval forum (IberLEF 2019) co-located with 35th Conf Spanish Soc Nat Lang process (SEPLN 2019); 2421: 550–560
21. El Alaoui I, Gahi Y, Messoussi R et al (2018) A novel adaptable approach for sentiment analysis on big social data. J Big Data 5. Epub ahead of print 2018:12. https://doi.org/10.1186/s40537-018-0120-0
22. FastText by fastText Team (2017) Retrieved March 31, 2021, from https://github.com/facebookresearch/fastText
23. Fleiss JL (1971) Measuring nominal scale agreement among many raters. Psychol Bull 76:378–382
24. Fu Y, Yang Z, Lin N et al (2021) Sentiment analysis for spanish tweets based on continual pre-training and data augmentation. CEUR Workshop Proc 2943:27–34
25. Gambino OJ, Calvo H (2019) Predicting emotional reactions to news articles in social networks. Comput Speech Lang 58:280–303
26. García-Díaz JA, Colomo-Palacios R, Valencia-García R (2021) Umuteam at emoevales 2021: emotion analysis for spanish based on explainable linguistic features and transformers. CEUR Workshop Proc 2943: 59–71
27. Ghosal D, Akhtar MS, Chauhan D et al (n.d.) Contextual Inter-modal Attention for Multi-modal Sentiment Analysis. In: Proceedings of the 2018 Conference on empirical methods in natural language processing. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 3454–3466
28. Huang C, Trabelsi A, Zaïane O (2019) ANA at SemEval-2019 Task 3: Contextual Emotion detection in Conversations through hierarchical LSTMs and BERT. In: Proceedings of the 13th International Workshop on Semantic Evaluation. Association for Computational Linguistics, Stroudsburg, pp 49–53
29. Justo R, Alcaide JM, Torres MI, Walker M (2018) Detection of sarcasm and nastiness: new resources for Spanish language. Cognit Comput 10:1135–1151
30. Keras (2021) Retrieved March 31, 2021, from https://keras.io/
31. Li K (2021) Haha at emoevales 2021: sentiment analysis in spanish tweets with cross-lingual model. CEUR Workshop Proc 2943:49–58
32. Liang X, Ma Y, Xu M (2019) THU-HCSI at SemEval-2019 Task 3: hierarchical ensemble classification of contextual emotion in conversation. In: Proceedings of the 13th International Workshop on Semantic Evaluation. Association for Computational Linguistics, Stroudsburg, pp 345–349
33. Luo H (2021) Emotion detection for spanish with data augmentation and transformer-based models. CEUR Workshop Proc 2943:35–42
34. Mahata D, Friedrichs J, Hitkul et al (2018) #phramacovigilance - Exploring deep learning techniques for identifying mentions of medication intake from twitter. Retrieved March 31, 2021, from http://arxiv.org/abs/1805.06375
35. Majumder N, Poria S, Gelbukh A, Cambria E (2017) Deep learning-based document modeling for personality detection from text. IEEE Intell Syst 32:74–79
36. Majumder N, Poria S, Peng H, Chhaya N, Cambria E, Gelbukh A (2019) Sentiment and sarcasm classification with multitask learning. IEEE Intell Syst 34:38–43
37. Martín C, Aguilar RM, Torres JM et al (2020) Supervisión remota en el entrenamiento de un clasificador de sentimientos en comentarios turísticos. In: XXXIX Jornadas de Automática. pp. 644–650
38. Mercado V, Villagra A, Errecalde M (2020) Political alignment identification : a study with documents of Argentinian journalists. J Comput Sci Technol 20:43–52
39. Mikolov T, Sutskever I, Chen K et al (2013) Distributed representations of words and phrases and their compositionality. NIPS'13 Proc 26th Int Conf Neural Inf Process Syst; 2. Retrieved March 31, 2021, from http://arxiv.org/abs/1310.4546
40. Moctezuma D, Graff M, Miranda-Jiménez S et al (2017) A genetic programming approach to sentiment analysis for twitter: TASS'17. CEUR Workshop Proc 1896:23–28
41. Muhammad A, Wiratunga N, Lothian R (2016) Contextual sentiment analysis for social media genres. Knowledge-Based Syst 108:92–101
42. Mukherjee I, Sahana S, Mahanti PK (2017) An improved information retrieval approach to short text classification. Int J Inf Eng Electron Bus 9:31–37
43. Nakov P, Kozareva Z, Ritter A et al (2013) SemEval-2013 task 2: Sentiment analysis in Twitter. In: SEMEVAL 2013 - 2nd Joint Conference on Lexical and Computational Semantics. pp. 312–320
44. Pennington J, Socher R, Manning C (2014) Glove: global vectors for word representation. In: Proceedings of the 2014 Conference on empirical methods in natural language processing (EMNLP). Association for Computational Linguistics, Stroudsburg, pp 1532–1543
45. Picard RW (1997) Affective computing. MIT Press, Cambridge
46. Plaza-Del-Arco FM, Jiménez-Zafra SM, Montejo-Ráez A et al (2021) Overview of the EmoEvalEs task on emotion detection for Spanish at IberLEF 2021. Proces Leng Nat 67:155–161

47. Poria S, Cambria E, Hazarika D et al (2017-Novem) Multi-level multiple attentions for contextual multimodal sentiment analysis. Proc - IEEE Int Conf Data Mining, ICDM 2017; 1033–1038
48. Qu Y, Jia S, Zhang Y (2021) Emotion analysis for spanish tweets: the model based on xlm-roberta and bi-gru. CEUR Workshop Proc 2943:101–109
49. Qu S, Yang Y, Que Q (2021) Emotion classification for spanish with xlm-roberta and textcnn. CEUR Workshop Proc 2943:94–100
50. Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back-propagating errors. Nature 323:533–536
51. Saif H, He Y, Fernandez M, Alani H (2016) Contextual semantics for sentiment analysis of twitter. Inf Process Manag 52:5–19
52. Sánchez JAF, Herranz SM, Unanue RM (2021) Urjc-team at emoevales 2021: Bert for emotion classification in spanish tweets. CEUR Workshop Proc 2943:43–48
53. Stone PJ, Hunt EB (1963) A computer approach to content analysis: studies using the general inquirer system. In: AFIPS conference proceedings - 1963 spring joint computer conference, AFIPS 1963. pp. 241–256.
54. Taller de Análisis de sentimientos en Español (TASS) Retrieved March 31, 2021, from http://tass.sepln.org
55. Tessore JP, Esnaola LM, Lanzarini L, et al. Distant Supervised Construction and Evaluation of a Novel Dataset of Emotion-Tagged Social Media Comments in Spanish. Cognit Comput. Epub ahead of print 18 January 2021. https://doi.org/10.1007/s12559-020-09800-x, 14, 407, 424, 2022.
56. Thakkar H, Patel D (2015) Approaches for sentiment analysis on twitter: a state-of-art study. Retrieved March 31, 2021, from http://arxiv.org/abs/1512.01043
57. Thelwall M, Buckley K, Paltoglou G, Cai D, Kappas A (2010) Sentiment strength detection in short informal text. J Am Soc Inf Sci Technol 61:2544–2558
58. Thelwall M, Buckley K, Paltoglou G (2012) Sentiment strength detection for the social web. J Am Soc Inf Sci Technol 63:163–173
59. Vanzo A, Croce D, Basili R (2014) A context-based model for sentiment analysis in twitter. COLING 2014 - 25th Int Conf Comput linguist proc COLING 2014 tech pap; 2345–2354
60. Vera D, Araque O, Iglesias CA (2021) Gsi-upm at iberlef2021: emotion analysis of spanish tweets by fine-tuning the xlm-roberta language model. CEUR Workshop Proc 2943:16–26
61. Vitiugin F, Barnabó G (2021) Emotion detection for spanish by combining laser embeddings, topic information, and offense features. CEUR Workshop Proc 2943:78–85
62. Voleti V (2018) Intuition behind LSTM. Retrieved March 31, 2021, from https://voletiv.github.io/docs/presentations/20180202_IIITH_Intuition_behind_LSTMs.pdf
63. Vosoughi S, Zhou H, Roy D (2015) Enhanced twitter sentiment classification using contextual information. In: Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. Association for Computational Linguistics, Stroudsburg, pp 16–24
64. Wilson T, Wiebe J, Hoffmann P (2010) Recognizing contextual polarity in phrase-level sentiment analysis. Int J Comput Appl 7:12–21
65. Winata GI, Madotto A, Lin Z et al (2019) CAiRE_HKUST at SemEval-2019 Task 3: hierarchical attention for dialogue emotion classification. In: Proceedings of the 13th International Workshop on Semantic Evaluation. Association for Computational Linguistics, Stroudsburg, pp 142–147
66. Xiao J (2019) Figure eight at SemEval-2019 Task 3: ensemble of transfer learning methods for contextual emotion detection. In: Proceedings of the 13th International Workshop on Semantic Evaluation. Association for Computational Linguistics, Stroudsburg, pp 220–224
67. Yusof NN, Mohamed A, Abdul-Rahman S (2018) A review of contextual information for context-based approach in sentiment analysis. Int J Mach Learn Comput 8:399–403
68. Zadeh A, Zellers R, Pincus E, Morency LP (2016) Multimodal sentiment intensity analysis in videos: facial gestures and verbal messages. IEEE Intell Syst 31:82–88

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.