information

MDPI

Article

Emotion Classification in Spanish: Exploring the Hard Classes

Aiala Rosá *,† and Luis Chiruzzo †

Facultad de Ingeniería, Instituto de Computación, Universidad de la República, Julio Herrera y Reissig 565, Montevideo 11300, Uruguay; luischir@fing.edu.uy
* Correspondence: aialar@fing.edu.uy; Tel.: +598-27142714
† These authors contributed equally to this work.

Abstract: The study of affective language has had numerous developments in the Natural Language Processing area in recent years, but the focus has been predominantly on Sentiment Analysis, an expression usually used to refer to the classification of texts according to their polarity or valence (positive vs. negative). The study of emotions, such as joy, sadness, anger, surprise, among others, has been much less developed and has fewer resources, both for English and for other languages, such as Spanish. In this paper, we present the most relevant existing resources for the study of emotions, mainly for Spanish; we describe some heuristics for the union of two existing corpora of Spanish tweets; and based on some experiments for classification of tweets according to seven categories (*anger, disgust, fear, joy, sadness, surprise*, and others) we analyze the most problematic classes.

Keywords: emotion classification; affective language; text classification

check for updates

Citation: Rosá, A.; Chiruzzo, L. Emotion Classification in Spanish: Exploring the Hard Classes. *Information* **2021**, *12*, 438. https://doi.org/10.3390/info12110438

Academic Editors: Salud María Jiménez-Zafra and Miguel Ángel García Cumbreras

Received: 8 September 2021
Accepted: 18 October 2021
Published: 21 October 2021

1. Introduction

The study of sentiments and emotions expressed in texts has been part of the Natural Language Processing field since the 1990s,increasing in interest in the last two decades due to the large amount of texts available on the Internet [1], especially messages on social networks where large numbers of people give their opinion on all relevant topics and express their personal emotions. Within this area, work has been done mainly on the classification of texts by their polarity or valence (positive or negative, differentiating in general neutral texts), this problem is usually called Sentiment Analysis. The detection of more complex emotions, on the other hand, is a problem of greater complexity, where a finer classification of texts is made, taking into account different emotions, such as joy, sadness, anger, surprise, among others. Emotion classification requires more expensive resources than the usual ones for polarity classification, as it requires more examples to cover the full set of classes, more annotators, and more attention to differentiate a varied number of classes and achieve a reliable inter-annotator agreement.

This effort is worthwhile as it could help in the detection of different problems that people experience and frequently express in social networks. Situations of harassment, signs of mental health problems, hate speech, and many other situations, can be detected by analyzing the affective content of texts. Furthermore, the study of emotions can be a tool for decision making in the political or business environment, where knowing the opinions and feelings of citizens or users can be very useful.

In order to face the problem of detecting emotions in texts, it is necessary to define the set of classes to be considered, and for this purpose work has been done on the basis of psychological studies. Several studies have worked on the scheme of three dimensions (valence, arousal, and dominance) presented by Wundt [2], carrying out different experiments that support this approach [3–5]. Other authors have defined sets of basic emotions, such as the set of six classes proposed by Ekman [6]: *anger, disgust, fear, happiness, sadness, and surprise*; or the Plutchik's set of eight emotions [7]: *anger, fear, sadness, disgust, surprise, anticipation, trust, and joy*. These theoretical proposals have led to the creation of different emotion lexicons (see Section 2).

Although having specific lexicons can help with emotion detection, the main resources for text classification currently are annotated datasets for training machine learning models. Large datasets can be used with deep neural network approaches, usually without the need to define attributes such as emotion word counts using a lexicon. Creating these resources is costly and imprecise, since emotions must be interpreted, which places us in a clearly subjective field. Nonetheless, corpus annotation tasks have been carried out on the basis of different sets of emotions, using different types of texts and different annotation schemes, e.g., single-label or multi-label classification.

In this paper, we present a revision of the resources available for emotion analysis, focusing on resources for the Spanish language, and previous work on automatic emotion classification. We also describe a new dataset built by merging two existing emotion datasets for Spanish and then we present some experiments performed on the new dataset, taking as a starting point the systems we sent to the EmoEvalEs task at IberLEF 2021 [8,9]. Finally, we analyze the most problematic classes.

The paper is organized as follows. In Section 2, we present the related work. In Section 3 we describe the materials (the new corpus) and the experiments on automatic detection of emotions. In Section 4, we show and analyze the results of the experiments. Finally, in Section 5 we present the conclusions of the work.

2. Related Work

In this section, we present different resources, lexicons, and annotated corpora, for emotion analysis, and related work on automatic emotion classification.

2.1. Lexicons for Emotion Analysis

Based on different emotion schemes, several lexical resources have been constructed, mainly for the English language. On the one hand, different resources have been developed based on the three basic emotion dimensions: valence, arousal and dominance. The ANEW (Affective Norms for English Words) resource [10] contains 1034 English words that were annotated with values for the three dimensions by a set of annotators. This resource was adapted to Spanish, based on a new annotation process involving 720 annotators [11]. The ANEW lexicon was expanded to have a larger resource [12] and then, from this new resource, a second expansion was made [13]. These authors also worked on generalizing the method to expand or create lexicons for other languages, performing some experiments for Spanish, Russian, and Farsi.

On the other hand, other lexical resources have been created classifying words according to the emotion sets proposed by Ekman or Plutchik. The NRC lexicon (Emolex) is based on Plutchik's eight categories and was created by crowdsourcing [14]. It contains more than 10,000 words, each of which can have several categories (or none). This lexicon was adapted into many different languages, including Spanish (https://saifmohammad.com/WebPages/AccessResource.htm, accessed on 17 October 2021).

Some efforts have been made for the construction of emotion lexicons for Spanish. The Spanish Emotion Lexicon (SEL) [15] contains 2036 words annotated with the probability of being used to express one of Ekman's six categories. In [16], the NRC lexicon was used to generate a Spanish lexicon for four emotions: anger, fear, joy, and sadness. The lexicon has about 5000 words and was created by machine translation, applying automatic post-processing and manual revision. The generated resource was validated against the corpus of one of the subtasks of SemEval-2018 Task1 [17] and compared with the existing corpus for Spanish, SEL.

2.2. Annotated Corpora for Emotion Classification

For the SemEval-2007 Task 14: Affective Text [18], an English corpus of 1250 news titles was annotated using Ekman's six categories plus valence labels (positive/negative). In [19], the annotation of an English corpus of tweets with intensities for four different emotions (anger, fear, joy, and sadness) is described, generating a dataset of 7097 tweets fairly balanced in terms of the four categories. The dataset was used in the WASSA-2017 Shared Task on Emotion Intensity [20].

In SemEval 2018—Task 1: Affect in Tweets [17] different subtasks were included for three languages: English, Spanish, and Arabic, taking as a starting point the previously mentioned English corpus used in WASSA-2017, which was extended for SemEval [21]. New datasets were created for Arabic and Spanish. On the one hand, some subtasks focused on the detection of some basic emotions and their degree of intensity. On the other hand, a subtask oriented to multi-label classification of emotions was presented, based on a set of eleven emotions (*anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise, and trust*) and the neutral class. This set of categories represents emotions that are frequent in tweets. This is the first experience that includes a corpus annotated with emotions for Spanish (7094 tweets in total, with multi-labels for eleven emotions or neutral). In SemEval-2019 Task 3: EmoContext, a task for detecting emotions in dialogues was proposed for English, based on four classes: *happy, sad, angry, and others*.

For Spanish, a new dataset with 8409 annotated tweets, the EmoEvent corpus [22], was recently created. The tweets are classified according to the six Ekman's categories, and the extra category *others*. The corpus was annotated by crowd-sourcing, each tweet received a label from three different annotators, and the final tweet label is chosen by majority vote. Tweets annotated with three different classes are assigned to the *others* class. Inter-annotator agreement was measured for the seven classes, showing that *disgust*, *fear*, and *surprise* have the lowest values. The corpus is also annotated with some extra information: topic and offensive content. As the tweets were selected by filtering those referring to certain predefined relevant events (Game of Thrones, Notre Dame Cathedral Fire, Venezuela's institutional crisis, etc.), a tag indicating which event the tweet belongs to is included. In addition, each tweet has a binary tag for offensive content annotation. In 2020, the Iberian Languages Evaluation Forum (IberLEF 2020) introduced for the first time a task on emotion evaluation in Spanish tweets, as part of the TASS 2020 task [23], which traditionally addressed tweets polarity classification. In IberLEF 2021 [24], a task only for emotion classification, EmoEvalEs [9], was proposed. Both editions of the task used the EmoEvent corpus.

2.3. Automatic Emotion Classification

Automatic emotion classification has not been as extensively investigated as polarity classification, which is the most commonly addressed problem when talking about sentiment analysis. Most of the work on emotion detection has been developed in the framework of evaluation campaigns, such as the aforementioned tasks on Affect Analysis from SemEval, the WASSA shared tasks, and, specifically for Spanish, TASS 2020 and EmoEvalEs 2021, both as part of the IberLEF forum. Below we discuss some methods that reached the best results in these tasks, in addition to some other recent work done outside these events. A survey of the area over the years can be found in [25].

In SemEval-2018 Task 1 [17], datasets were available for three languages: English, Arabic and Spanish. Twelve teams participated in the E-c subtask (emotion multi-label classification). The best results for Macro F1 for each language were: English, 0.530; Spanish, 0.440; Arabic, 0.475. In TASS2020-Task2 [23] on single-label emotion classification, two teams participated, and the best result was 0.447 of Macro F1 (on a different test corpus). In both competitions (SemEval and EmoEval) the team that obtained the best results for Spanish was ELiRF-UPV. For SemEval2018-Task1 they represented each tweet as the concatenation of word embeddings, including information from polarity lexicons, for feeding a CNN + LSTM neural network [26]. For their participation in TASS2020-Task2 [27] the team proposed using TWilBERT, a model generated by fine-tuning an adaptation of BERT, pre-trained on a corpus of 91 million Spanish tweets collected by the authors [28].

Some experiments on one of the Spanish corpora used in SemEval2018-Task1 (the corpus with four classes: *anger, fear, sadness, and joy*) have been performed by [29], who trained different machine learning classifiers with features from the SEL lexicon and the Spanish adaptation of the NRC lexicon. They achieved 0.74 of Accuracy with an SVM classifier.

In [22], an experiment was carried out, using an SVM classifier with tf-idf features, for validating the corpus created for emotion classification. This experiment reached an Accuracy of 0.64 with cross-validation.

The EmoEvalEs task at IberLEF 2021 [9] used the EmoEvent corpus as the official dataset. Most of the submitted systems used variants of BERT in their approaches, mainly BETO [30], a version of the BERT model trained entirely with Spanish texts. Several teams worked on corpus augmentation, applying different techniques, such as back translation and masked language models. Some teams used the extra information provided by the event and offensiveness labels: the first and fifth teams used both tags and the third team used only the offensiveness tag. The first place system [31] uses multilingual RoBERTa, they experimented adding the event and offensiveness information as features, but this decreased the performance. Their best model reached 0.73 of Accuracy and 0.72 of Weighted F1.

The WASSA Shared Task 2021 [32] included a sub-track on emotion classification using a corpus of English news stories, annotated with Ekman's six categories plus non-emotion. The corpus is not too large, containing 2655 examples. The best result in emotion classification was obtained by the WASSA@IITK team: 0.55 Macro-F1 and 0.62 Accuracy.

Emotion detection for other languages— specifically Chinese and Hindi—using similar datasets as described above, reaches results similar to the mentioned shared tasks. For Chinese [33], experiments have been performed on two corpora, reaching 48.31 Macro F1 and 60.76 Accuracy in the Ren-CECps corpus (eight classes plus neutral) and 49.42 Macro F1 and 63.02 Accuracy in the NLPCC2018 corpus (five classes, Chinese/English code-switching). For Hindi [34], some experiments were conducted by applying transfer learning from English resources, which achieved a Macro F1 of 53.2 (eight classes).

Some works propose the automatic creation of corpora, using hashtags to filter tweets conveying different emotions, without neutral category (*others*). Since these resources are not manually curated, they are larger than the standard datasets we mentioned above. For English, in [35] this method combined with distant supervision is used to build a corpus for the full Pluthick schema (three levels of emotions with eight classes each). Experiments with the eight basic emotions achieved an Accuracy of 95.68, using Gated Recurrent Neural Networks. Another work based on a corpus built using hashtags is presented in [36]. They generate different versions of the corpus, based on different sets of emotions, and with single-label and multi-label annotation. The best results on the single-label corpus annotated with Ekman's six classes are 61.8 for Macro F1 and 73.0 for Accuracy, training RNNs on sequences of characters. A similar approach for corpus creation was performed for Hindi/English mixed-code tweets [37], using Ekman's six categories. A BERT-based system achieved an Accuracy of 71.43

It is important to point out that it is really complex to compare different works on emotion detection, given the existing variety in different aspects. On the one hand, there are differences in the set of emotions; as seen above, sets with four, six, eight or eleven classes have been used. On the other hand, the inclusion or not of a neutral class seems to significantly affect the results, the works that do not include it achieve better accuracy. Other differences are the type of annotation—single-label or multi-label—and the metrics reported, the most frequent being Accuracy and Macro F1, but also Weighted F1. Different strategies for corpus construction were also observed, in particular, with manual curation or fully automatically annotated, filtering tweets based on hashtags.

3. Materials and Methods

Next, we describe a new emotion corpus, built by merging the EmoEvent corpus and the corpus used in the subtask on emotion classification at SemEval-2018 Task 1. Then, we present some experiments on emotion classification using the expanded corpus for training.

For development and evaluation we used the original corpora from EmoEvalEs, in order to make some comparisons with previous results.

3.1. Corpora

We worked on the expansion of the EmoEvent corpus, which was used in the TASS-2020 and EmoEvalEs-2021 tasks (https://competitions.codalab.org/competitions/28682, accessed on 17 October 2021), by adapting the corpus used in the subtask on emotion classification (E-c) at SemEval-2018 Task1 (https://competitions.codalab.org/competitions/17751, accessed on 17 October 2021). Table 1 shows the statistics of the EmoEvent dataset, taking the partition in train/development/test used in the EmoEvalEs task. The corpus contains 8223 tweets: 5723 for training, 844 for development, and 1656 for testing.

Table 1. Number of tweets per category in the EmoEvent dataset.

Emotion	Train	Dev	Test
Anger	589	85	168
Disgust	111	16	33
Fear	65	9	21
Joy	1227	181	354
Sadness	693	104	199
Surprise	238	35	67
Others	2800	414	814
Total	5723	844	1656

In order to compare the results with those published in the proceedings of EmoEvalEs, we kept the original EmoEvalEs development and test corpora and worked on extending only the training corpus. Since the corpora of both tasks are annotated based on different schemes, different options were tested for the creation of the extended corpus.

On the one hand, the sets of categories are not the same: EmoEvent uses Ekman's six classes, while SemEval uses a set of eleven categories, which includes Ekman's. We tried two different approaches: the first approach was to keep only the six EmoEvent categories and discard the remaining ones; the second approach was to assign the label *others* to the five extra categories. As already stated, this class was used in the EmoEvent to annotate the tweets which received different emotions according to different annotators, or neutral tweets, so this decision appears consistent.

On the other hand, the annotation of the two corpora differs in the number of classes per tweet. In the EmoEvent corpus each tweet is assigned a single category, while in the SemEval corpus multi-label tagging was performed. Again, two different approaches were tested: filtering only tweets in the SemEval corpus that have a single label or duplicating tweets with multiple labels, generating an instance for each value.

After testing these different options, we decided to keep the tweets that do not have emotions included in the six-value schema, assigning them the category *others*, and to create multiple instances of the tweets that have more than one class from the six-value schema (one instance for each class assigned to the tweet).

We also tried to extend the corpus by translating the English and Arabic datasets from SemEval, and applying back-translation for data augmentation (using the huggingface MarianMT models (https://huggingface.co/transformers/model_doc/marian.html, accessed on 17 October 2021)), but the experiments performed with these versions of the training corpus always performed worse than using just the Spanish set from SemEval.

Table 2 shows the final EmoEvent + SemEval training corpus statistics.

Table 2. Number of tweets per category in the expanded training corpus (EmoEvent + SemEval).

Emotion	Tweets
Anger	2872
Disgust	1153
Fear	810
Joy	3388
Sadness	2325
Surprise	566
Others	3816
Total	14,930

A particular problem with these corpora, mainly the EmoEval corpus, is the significant imbalance between classes. Two classes stand out as particularly sparse: *surprise* and *fear*. Although these two classes are better represented after merging both corpora, the imbalance is still significant.

3.2. Preprocessing

First of all we removed URLs that could be present in the tweets. This was the only preprocessing done to the text when using the tweet for generating a BERT-based encoding, as we used a case-sensitive model. On the other hand, when using the tweet for extracting word embeddings or detecting occurrences of lexicon words, we removed all special characters (particularly the hashtag symbols) and kept only letters and numbers, then we converted all the text to lowercase and split tokens based on blank spaces.

EmoEvalEs [9] reported replacing hashtag mentions with a single token "HASHTAG" in the tweets, and also replacing user mentions with a "USER" token. We noticed that the user mentions were replaced, but not the hashtags, and we decided to keep it that way as the hashtag text might convey important information about tweet topics, which could influence the emotions.

3.3. Experiments

Based on the experiments we carried out for our participation in EmoEvalEs, we tested different variants of the merged corpus with the best model presented in the task. This model consists of a LSTM network that takes as input information generated by BETO for each tweet. The information we take from BETO is the encoding of the CLF token, and the centroid of the encoding of each token. This combination gave us better results than using only the CLF embedding.

In addition, we added features to indicate whether the tweet contains any words from a list of relevant words obtained from the training corpus. These are the most relevant words for the discrimination of the different classes in the training corpus, generated by the k_best function, which uses the ANOVA F-value method. We tested different values of k, choosing the ones that gave the best results with a simple baseline using a SVM model with bag of words type features and tested them with the LSTM. While large values for k give good results for the baseline, with 2000 being the best value we found, for the LSTM model the best value turned out to be 30.

Besides these relevant word features, we tried including features for each one of the words in the NRC lexicon adapted to Spanish (https://saifmohammad.com/WebPages/AccessResource.htm, accessed on 17 October 2021). However, the experiments using this extended set of features from the lexicon did not result in any improvement.

In our experiments we did not use the extra information about the event to which the tweet belongs or whether it is offensive or not. This is information that is not available in new instances to be classified, so it does not make sense to rely on it. There is also no such information in the SemEval corpus with which we extended the EmoEvent corpus.

The final LSTM model uses a single bi-directional LSTM layer of size 96 and a dense layer of size 64 (with tanh activation), the output of the LSTM is concatenated with the BERT features and features for the 30 best words.

For all the experiments carried out with SVM, mentioned in later sections, we used the default parameters configuration of the *svc* function of the sklearn library (https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html, accessed on 17 October 2021).

4. Results and Discussion

Table 3 shows the results on the development and test corpora from EmoEvalEs of the neural model trained with the EmoEvent corpus, on the one hand, and with the EmoEvent + SemEval corpus, on the other hand. The metrics we used for evaluation are Accuracy (Acc) and Weighted F1 (W-F1).

Table 3. Results on the development and test sets of the neural model trained with EmoEvent and with EmoEvent + SemEval.

Training Corpus	Acc on Dev	W-F1 on Dev	Acc on Test	W-F1 on Test
EmoEvent	0.7026 ± 0.0219	0.6815 ± 0.0224	0.6781 ± 0.0224	0.6573 ± 0.0228
EmoEvent + SemEval	0.7121 ± 0.0217	0.6884 ± 0.0222	0.6860 ± 0.0223	0.6620 ± 0.0227

Table 4 shows the results of our current best model, which uses data from EmoEvent and the Spanish set of SemEval data, compared to the best and worst systems in EmoEvalEs. We also include our own submission to EmoEvalEs for comparison.

Table 4. Comparison of the best, worst, and our system (originally described in [8]) in the EmoEvalEs competition, together with the result of the new model trained with data from EmoEvent and SemEval. Metrics: Accuracy (Acc), Weighted Precision (W-P), Weighted Recall (W-R), and Weighted F1 (W-F1).

System	Acc	W-P	W-R	W-F1
GSI-UPM	0.7276	0.7094	0.7276	0.7170
EmoEvent + SemEval	0.6860	0.6683	0.6860	0.6620
RETUYT-InCo (EmoEvent)	0.6781	0.6583	0.6781	0.6573
qu	0.4498	0.6188	0.4498	0.4469

The results seem to indicate a slight improvement training with the extended train corpus, over training only with the EmoEvent corpus, but, as the confidence intervals are not completely separate, further experiments would have to be performed to confirm this. This result is of particular interest since the new corpus contains tweets on different topics, not only tweets on some specific events, as is the case with EmoEvent. The combination of the two corpora could have had a negative effect on the results, compared to the training with the original corpus, since the test corpus contains exclusively tweets related to the events selected for building the EmoEvent corpus.

Looking at the confusion matrix of the model trained with the EmoEvent + SemEval corpus, shown in Figure 1, we see that almost all classes tend to be confused with the class *others*. Due to the way this class was generated [22], it is expected that many of these tweets express some emotion, since tweets that received different emotions by different annotators were assigned to the *others* category. It is not a category representing tweets without emotion, but tweets with some emotion or a mixture of several ones, and probably also neutral tweets. Something similar happened with the *neutral* class of the dataset for sentiment analysis of the TASS task, where tweets with a neutral polarity and also tweets with mixed polarity, i.e., with both positive and negative nuances, could be found. In [38], we discuss this problem and show a detailed analysis of tweets belonging to that category.

	anger	disgust	fear	joy	others	sadness	surprise
anger	101	1	1	4	47	13	1
disgust	13	0	0	2	16	2	0
fear	1	0	5	0	14	1	0
joy	5	0	0	192	149	6	2
others	25	0	1	76	700	11	1
sadness	9	0	2	2	53	133	0
surprise	4	0	0	14	43	1	5

Figure 1. Confusion matrix over the test set for the best model. Rows show expected values and columns show predicted values.

This can be seen graphically in Figure 2 as well. In this diagram, the crossing arcs represent cases in which the classifier is wrong, and the bumps inside categories represent cases when the classifier is right. We can see in this diagram that the *others* category is the most numerous, and approximately a quarter of their expected values are classified as other categories. The class *joy*, on the other hand, has close to half of its examples classified as *others*. Note that categories *surprise*, *fear*, and *anger* have almost no visible bump, as almost no example of these categories is correctly classified. Furthermore, we can see that the *disgust* class is wrongly mistaken with the *anger* and *others* category in similar proportions.

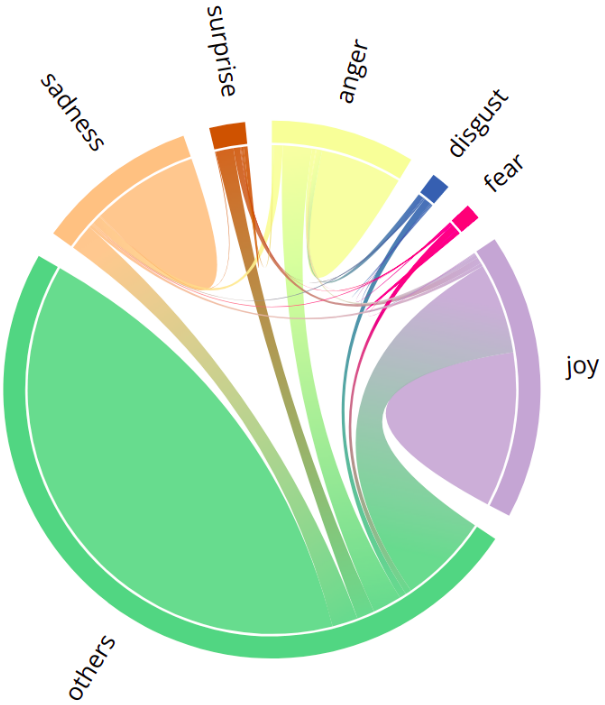

Figure 2. Chord diagram representing the confusion matrix for the best model. An incoming arc from class X to Y represents that class Y was expected, but X was predicted.

An experiment performed with a new version of the corpus excluding the *others* category allows us to see how the remaining categories are better classified. In Table 5, we show the results for each category using both versions of the corpus: with and without the *others* class.

Table 5. Results on the test set training with the EmoEvent + SemEval corpus, with and without the *others* class.

Corpus Version	Class	F1 on Test
With others	anger	0.6196
	disgust	0.0000
	fear	0.3333
	joy	0.5963
	sadness	0.7268
	surprise	0.1316
	others	0.7625
Accuracy:	0.6860	
Weighted-F1:	0.6620	
Without others	anger	0.7049
	disgust	0.0000
	fear	0.5500
	joy	0.8475
	sadness	0.7749
	surprise	0.2917
Accuracy:	0.7447	
Weighted-F1:	0.7170	

As can be seen, the overall measures improve significantly. On the test corpus, Accuracy rises 5.87 points and Weighted F1 rises 5.50 points, the increase being even greater in the *fear* (+22) and *joy* (+25) categories. However, the most problematic class (*disgust*) does not improve.

Some examples of *others* tweets from the training corpus show the diversity we can find in this class:

- A sad tweet: *Guardaré en mis ojos tu última mirada... #notredame #paris #francia #photography #streetphotography*;
- A clearly positive tweet, that could even have been annotated as *joy*: *Que clase táctica están dando estos dos Equipos... bendita #ChampionsLeague*;
- An informative tweet, with no emotion: *El escrutinio en el Senado va mucho más lento. Solo el 14.85% del voto escrutado #28A #EleccionesGenerales28A*.

Besides the *others* class, we carried some more experiments on the three more difficult classes to detect (*disgust*, *fear*, and *surprise*), to try to understand what made these categories so difficult. One interesting experiment is trying to analyze which words are the most relevant ones for each class, the ones that would let us tell apart a tweet in one of these categories with the highest confidence. In order to do this, we trained several variants of SVM classifiers using different lists of BOW features found with the ANOVA F-value method. For these classifiers, we were only trying to classify a class against all the rest, for example: *disgust* vs. *no-disgust*, or *fear* vs. *no-fear*.

4.1. The Disgust Class

Regarding the *disgust* class, our best neural models are not able to correctly spot even a single tweet from this class. Some examples from this class are:

- Tweet transmitting a very low level of dislike: *Me cuesta mucho entender la fiesta de Ciudadanos...#EleccionesGenerales28A*;
- Tweet transmitting a very high level of dislike *Caterva de hijueputas venezolanos que le hacen juego al pilche golpe. Háganse los valientes en #Venezuela y no jodan en Ecuador. Dan asco....*;
- Informative tweet: *Los gobiernos de #Argentina #Brasil #Canada #Chile #Colombia #CostaRica #Guatemala #Honduras #Panamá #Paraguay #Peru y #Venezuela, miembros del #GrupoDe-*

Lima, "conminan a USER a cesar la usurpación, para que pueda empezar la transición democrática" en #Venezuela;

- Tweet that could have been annotated as *anger*: *Si no fuésemos estúpidos/as,los gestores de nuestro sistema alimentario estarían en prisión, por actos criminales contra la naturaleza y la salud pública.#ExtinctionRebellion #GretaThunberg*.

As we can see, there is a great disparity of opinions inside this category, many of them convey different levels of dislike, but some could perhaps be best represented as other categories. This is consistent with the fact that it was one of the categories with the lowest inter-annotator agreement [22], which might indicate the difficulty to properly characterize this category.

As the confusion matrix showed in Figure 1, this class is frequently confused with the *anger* class. This is to be expected because they are two classes with a strong negative content; they could even be seen as two degrees of the same emotion.

Our SVM analysis showed that, even using the 200 best words as features, it is only possible to achieve 0.1 F1 on the development set, and it still gets an absolute 0.0 F1 on the test set. This might indicate that the vocabulary overlap between those sets is rather small, which renders the problem even harder.

4.2. The Fear Class

When analyzing the *fear* class, we realized that a limited number of words occur in many tweets: *miedo, ansiedad, temor*, among others. We performed some experiments with an SVM classifier using features from BOW, selecting the best words provided by the k_best function, replacing all the classes but *fear* by the tag *no-fear* on the corpus. We found that with just 20 words we could reach good results: 0.60 of F1 on the development corpus and 0.45 on the test corpus. The 20 words used as features are:

- Ansiedad;
- Ansiosa;
- Ansioso;
- Asustada;
- Asustado;
- Asustar;
- Co;
- https;
- Miedo;
- Nerviosa;
- Nervioso;
- Peligroso;
- Pesadilla;
- Preocupación;
- Preocupada;
- Pánico;
- Susto;
- Temblor;
- Temor;
- Terror.

4.3. The Surprise Class

For the *surprise* class, the experiments with the SVM plus BOW features that gave the best results used the 200 words from the k_best function, reaching 0.23 for F1 both on the development and the test set.

Analyzing some tweets from this class, we observed that more than the vocabulary, what many tweets have in common is the use of intensifying punctuation marks such as repeated exclamation or question marks (as can be seen in the first examples below).

Another characteristic of this class is the use of some explicit references to a surprise or being surprised, or some particular idioms like *"sin palabras"* (breathtaking).

- *Liverpool está paseando al barcelona, hace tiempo no lo veía tan presionado al barca ..!! #ChampionsLeague*;
- *Tremendo liderazgo de #GretaThunberg !! Tiene 16 años y nos está haciendo a TODOS mirar el mundo con otros ojos! Entremos en pánico, salvemos el planeta!! #CambioClimatico*;
- *El Messi de hoy deslumbra! Que nivel #ChampionsLeague*;
- *Primera vez que veo a Messi exagerar una falta. #ChampionsLeague*;
- *El único episodio que me ha dejado sin palabras. #JuegoDeTronos #GameofThrones*;
- *Menudo sorpresón final. Eso sí que no me lo esperaba. #JuegoDeTronos*;
- *Lo que vino a sacarles #NotreDame, es TODA su amargura. A la madere, me tienen sorprendida.*

5. Conclusions

Emotion detection and classification is a very hard task, but it has started to gain traction quickly in the latest years. Although resources for this task, particularly for Spanish, are still scarce, we consider this situation might change in the future if the task keeps being developed for Spanish and for other languages. For the time being, the performance for emotion classification in Spanish still has a lot of room for improvement. We performed a series of experiments using different variants of the available data for classifying emotions in Spanish: tweet datasets EmoEvent and SemEval, and an adaptation of an English lexicon specialized in emotion detection to Spanish.

The best model we found combines a LSTM neural network enriched with encoding features calculated with BERT in Spanish and features representing the most salient words from the corpus. This model is trained using the training set from EmoEvent, plus the Spanish set from the SemEval emotion classification challenge. It achieves an accuracy of 0.6860 and a weighted F1 score of 0.6620 over the EmoEvent test set. We compared our performance against our previous results and other results from the EmoEvalEs competition.

Despite our results being promising, we found we are still a few points below the state of the art for this task in Spanish. Furthermore, even if our new model seems to show improvements for some of the least numerous classes (*fear* and *surprise*), it still cannot detect any tweets from the *disgust* class, and we must consider that the confidence intervals for the new and previous model are not completely separate as well. We analyzed the three categories with fewer examples (*disgust*, *fear*, and *surprise*, to try to understand why they behave that way. We conclude that the great disparity of negative emotions conveyed in the *disgust* tweets and the large variability of vocabulary might be one of the causes behind the difficulty of this class. We also analyzed the class *others*, which contains tweets with very diverse content, and observed that excluding these tweets from the corpus significantly increases the results.

Author Contributions: Conceptualization, A.R.; Formal analysis, A.R. and L.C.; Investigation, A.R. and L.C.; Software, L.C.; Writing—original draft, A.R. and L.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not Applicable, the study does not report any data.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

BERT	Bidirectional Encoder Representations from Transformers
BOW	Bag of Words
ANEW	Affective Norms for English Words
LSTM	Long Short-Term Memory
SEL	Spanish Emotion Lexicon
SVM	Support Vector Machine

References

1. Liu, B. *Sentiment Analysis. Mining Opinions, Sentiments, and Emotions*, 2nd ed.; Cambridge University Press: New York, NY, USA, 2020.
2. Wundt, W. *Grundriss der Psychologie (Outlines of Psychology)*; Engelmann: Leibzig, Germany, 1874.
3. Osgood, C.; Suci, G.; Tenenbaum, P. *The Measurement of Meaning*; University of Illinois Press: Urbana, IL, USA, 1957.
4. Mehrabian, A.; Russell, J. *An Approach to Environmental Psychology*; The MIT Press: Cambridge, MA, USA, 1974.
5. Bradley, M.M.; Lang, P.J. Measuring emotion: The self-assessment manikin and the semantic differential. *J. Behav. Ther. Exp. Psychiatry* **1994**, *25*, 49–59. [CrossRef]
6. Ekman, P. Facial Expressions of Emotion: New Findings, New Questions. *Psychol. Sci.* **1992**, *3*, 34–38. [CrossRef]
7. Plutchik, R. A psychoevolutionary theory of emotions. *Soc. Sci. Inf.* **1982**, *21*, 529–553. [CrossRef]
8. Chiruzzo, L.; Rosá, A. RETUYT-InCo at EmoEvalEs 2021: Multiclass Emotion Classification in Spanish. In Proceedings of the Iberian Languages Evaluation Forum Co-Located with the XXXVII International Conference of the Spanish Society for Natural Language Processing (SEPLN 2021), Málaga, Spain, 21–24 September 2021.
9. Plaza-del-Arco, F.M.; Jiménez-Zafra, S.M.; Montejo-Ráez, A.; Molina-González, M.D.; Ureña-López, L.A.; Martín-Valdivia, M.T. Overview of the EmoEvalEs task on emotion detection for Spanish at IberLEF 2021. *Proces. Leng. Nat.* **2021**, *67*, 155–161.
10. Bradley, M.; Lang, P. *Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings*; Technical Report; Center for Research in Psychophysiology, University of Florida: Gainesville, FL, USA, 1999.
11. Redondo, J.; Fraga, I.; Padrón, I.; Comesaña, M. The Spanish adaptation of ANEW (Affective Norms for English Words). *Behav. Res. Methods* **2007**, *39*, 600–605. [CrossRef] [PubMed]
12. Warriner, A.B.; Kuperman, V.; Brysbaert, M. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behav. Res. Methods* **2013**, *45*, 1191–1207. [CrossRef] [PubMed]
13. Shaikh, S.; Cho, K.; Strzalkowski, T.; Feldman, L.; Lien, J.; Liu, T.; Broadwell, G.A. ANEW+: Automatic Expansion and Validation of Affective Norms of Words Lexicons in Multiple Languages. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), Portorož, Slovenia, 23–28 May 2016; pp. 1127–1132.
14. Mohammad, S.; Turney, P.D. Crowdsourcing a Word-Emotion Association Lexicon. *Comput. Intell.* **2013**, *29*, 436–465. [CrossRef]
15. Rangel, I.; Sidorov, G.; Guerra, S. Creación y evaluación de un diccionario marcado con emociones y ponderado para el español. *Onomázein* **2014**, *29*, 31–46. [CrossRef]
16. Plaza-del-Arco, F.M.; Molina-González, M.D.; Jiménez-Zafra, S.M.; Martín-Valdivia, M.T. Lexicon Adaptation for Spanish Emotion Mining. *Proces. Leng. Nat.* **2018**, *61*, 117–124.
17. Mohammad, S.M.; Bravo-Marquez, F.; Salameh, M.; Kiritchenko, S. SemEval-2018 Task 1: Affect in Tweets. In Proceedings of the International Workshop on Semantic Evaluation (SemEval-2018), New Orleans, LA, USA, 5–6 June 2018.
18. Strapparava, C.; Mihalcea, R. SemEval-2007 Task 14: Affective Text. In Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007), Prague, Czech Republic, 23–24 June 2007; pp. 70–74.
19. Mohammad, S.; Bravo-Marquez, F. Emotion Intensities in Tweets. In Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017), Vancouver, QC, Canada, 3–4 August 2017; pp. 65–77. [CrossRef]
20. Mohammad, S.M.; Bravo-Marquez, F. WASSA-2017 Shared Task on Emotion Intensity. In Proceedings of the Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA), Copenhagen, Denmark, 8 September 2017.
21. Mohammad, S.; Kiritchenko, S. Understanding Emotions: A Dataset of Tweets to Study Interactions between Affect Categories. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC), Miyazaki, Japan, 7–12 May 2018.
22. Plaza-del-Arco, F.; Strapparava, C.; Ureña-López, L.A.; Martín-Valdivia, M.T. EmoEvent: A Multilingual Emotion Corpus based on different Events. In Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, France, 11–16 May 2020; pp. 1492–1498.
23. García-Vega, M.; Díaz-Galiano, M.C.; García-Cumbreras, M.A.; del Arco, F.M.P.; Montejo-Ráez, A.; Jiménez-Zafra, S.M.; Martínez Cámara, E.; Aguilar, C.A.; Cabezudo, M.A.S.; Chiruzzo, L.; et al. Overview of TASS 2020: Introducing Emotion Detection. In Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020) Co-Located with 36th Conference of the Spanish Society for Natural Language Processing (SEPLN 2020), Málaga, Spain, 23 September 2020; pp. 163–170.

24. Montes, M.; Rosso, P.; Gonzalo, J.; Aragón, E.; Agerri, R.; Álvarez-Carmona, M.Á.; Álvarez Mellado, E.; Carrillo-de Albornoz, J.; Chiruzzo, L.; Freitas, L.; et al. Iberian Languages Evaluation Forum 2021. In Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021), Málaga, Spain, 21–24 September 2021. Available online: http://ceur-ws.org/Vol-2943/ (accessed on 17 October 2021).
25. Acheampong, F.A.; Wenyu, C.; Nunoo-Mensah, H. Text-based emotion detection: Advances, challenges, and opportunities. *Eng. Rep.* **2020**, *2*, e12189. [CrossRef]
26. González, J.Á.; Hurtado, L.F.; Pla, F. ELiRF-UPV at SemEval-2018 Tasks 1 and 3: Affect and Irony Detection in Tweets. In Proceedings of the 12th International Workshop on Semantic Evaluation, New Orleans, LA, USA, 5–6 June 2018; pp. 565–569. [CrossRef]
27. Ángel González, J.; Moncho, J.A.; Hurtado, L.F.; Pla, F. ELiRF-UPV at TASS 2020: TWilBERT for Sentiment Analysis and Emotion Detection in Spanish Tweets. In Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020) Co-Located with 36th Conference of the Spanish Society for Natural Language Processing (SEPLN 2020), Málaga, Spain, 23 September 2020; pp. 179–186.
28. Gonzalez, J.A.; Hurtado, L.F.; Pla, F. TWilBert: Pre-trained deep bidirectional transformers for Spanish Twitter. *Neurocomputing* **2021**, *426*, 58–69. [CrossRef]
29. Plaza-del-Arco, F.; Martín-Valdivia, M.T.; Ureña-López, L.A.; Mitkov, R. Improved emotion recognition in Spanish social media through incorporation of lexical knowledge. *Future Gener. Comput. Syst.* **2020**, *110*, 1000–1008. [CrossRef]
30. Cañete, J.; Chaperon, G.; Fuentes, R.; Ho, J.H.; Kang, H.; Pérez, J. Spanish Pre-Trained BERT Model and Evaluation Data. PML4DC at ICLR 2020. 2020. Available online: https://users.dcc.uchile.cl/~jperez/papers/pml4dc2020.pdf (accessed on 17 October 2021).
31. Vera, D.; Araque, O.; Iglesias, C.A. GSI-UPM at IberLEF2021: Emotion Analysis of Spanish Tweets by Fine-tuning the XLM-RoBERTa Language Model. In Proceedings of the Iberian Languages Evaluation Forum co-located with the XXXVII International Conference of the Spanish Society for Natural Language Processing (SEPLN 2021), Málaga, Spain, 21–24 September 2021.
32. Tafreshi, S.; De Clercq, O.; Barriere, V.; Buechel, S.; Sedoc, J.; Balahur, A. WASSA 2021 Shared Task: Predicting Empathy and Emotion in Reaction to News Stories. In Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, Online, 19 April 2021; pp. 92–104.
33. Deng, J.; Ren, F. Multi-label Emotion Detection via Emotion-Specified Feature Extraction and Emotion Correlation Learning. *IEEE Trans. Affect. Comput.* **2020**. [CrossRef]
34. Ahmad, Z.; Jindal, R.; Ekbal, A.; Bhattachharyya, P. Borrow from rich cousin: Transfer learning for emotion detection using cross lingual embedding. *Expert Syst. Appl.* **2020**, *139*, 112851. [CrossRef]
35. Abdul-Mageed, M.; Ungar, L. EmoNet: Fine-Grained Emotion Detection with Gated Recurrent Neural Networks. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vancouver, QC, Canada, 7–12 August 2017; pp. 718–728. [CrossRef]
36. Colnerič, N.; Demšar, J. Emotion Recognition on Twitter: Comparative Study and Training a Unison Model. *IEEE Trans. Affect. Comput.* **2020**, *11*, 433–446. [CrossRef]
37. Wadhawan, A.; Aggarwal, A. Towards Emotion Recognition in Hindi-English Code-Mixed Data: A Transformer Based Approach. In Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, Online, 19 April 2021; pp. 195–202.
38. Chiruzzo, L.; Etcheverry, M.; Rosá, A. Sentiment analysis in Spanish tweets: Some experiments with focus on neutral tweets. *Rev. Proces. Leng. Nat.* **2020**, *64*, 109–116.