Overview of TASS 2019: One More Further for the Global Spanish Sentiment Analysis Corpus*

Manuel Carlos Díaz-Galiano[1], Manuel García-Vega[1], Edgar Casasola[2], Luis Chiruzzo[3], Miguel Á. García-Cumbreras[1], Eugenio Martínez Cámara[4], Daniela Moctezuma[5], Arturo Montejo Ráez[1], Marco Antonio Sobrevilla Cabezudo[6], Eric Tellez[7], Mario Graff[7], and Sabino Miranda[7]

[1] Universidad de Jaén, Jaén, España {mcdiaz,mgarcia,magc,amontejo}@ujaen.es
[2] Universidad de Costa Rica, San José, Costa Rica, edgar.casasola@ucr.ac.cr
[3] Universidad de la República Montevideo, Uruguay luischir@fing.edu.uy
[4] Andalusian Research Institute in Data Science and Computational Intelligence (DaSCI), Universidad de Granada, España emcamara@decsai.ugr.es
[5] CONACyT-CentroGEO dmoctezuma@centrogeo.edu.mx
[6] Universidade de São Paulo msobrevillac@usp.br
[7] CONACyT-INFOTEC {eric.tellez, mario.graff, sabino.miranda}@infotec.mx

Abstract. In September 2019, the eighth edition of TASS workshop (Task of Sentiment Analysis at SEPLN) was held in Bilbao, Spain as part of the first edition of IberLEF (Iberian Languages Evaluation Forum), which joined the efforts of the IberEval and TASS workshops. In this edition, the natural evolution from previous editions was proposed: sentiment analysis at tweet level. It includes two subtasks, monolingual and cross-lingual sentiment analysis, with different subsets of the InterTASS corpus (ES-Spain, PE-Peru, CR-Costa Rica, UR-Uruguay and MX-Mexico). This paper summarizes the approaches and the results of the submitted systems of the different groups for each task.

Keywords: Sentiment Analysis · Opinion Mining · Social Media.

1 Introduction

After seven editions of the workshop on Semantic Analysis at SEPLN (TASS) as an independent workshop co-located with the International Conference of the Spanish Society on Natural Language Processing (SEPLN), TASS has been incorporated into the Iberian Languages Evaluation Forum (IberLEF)[8]. IberLEF

* This work has been partially supported by a grant from the Spanish Government under the LIVING-LANG project (RTI2018-094653-B-C21) and the REDES project (TIN2015-65136-C2-1-R). Eugenio Martínez Cámara was supported by the Spanish Government Programme Juan de la Cierva Formación (FJCI-2016-28353).

[8] https://iberlef.sepln.org/

is the result of the association of some workshops in the Natural Language Processing (NLP) domain for Spanish and other languages spoken in the Iberian peninsula, and the aim of joining forces of different NLP research communities in order to provide a common forum for assessing NLP systems and interchanging research ideas, issues, challenges and experiences.

Since the edition of TASS 2017 [3], the main aim is the elaboration of a corpus of sentiment tweets written in different Spanish variants, in order to provide a representative corpus of Spanish posts written in microblogs all over the world and not only of the usage of the Spanish language in Spain. Accordingly, the International TASS corpus was released for the first time in the edition of 2017, and it was only composed of tweets written in the Spanish used in Spain. The second version of InterTASS was released in the edition of TASS 2018 [4], and set of tweets written in the Spanish used in Perú and Costa Rica were added. The novelty of the edition of TASS 2019 lies in the incorporation of two new Spanish variants, namely the Spanish written in social media in Mexico and Uruguay.

The goal behind the aim of compiling a global Spanish corpus of tweets is to study the differences among different versions of Spanish, and fostering the cross-lingual research on the Spanish language. Consequently, the TASS 2019 proposed two subtasks, specifically a mono-lingual polarity classification task (Subtask 1) and a cross-lingual polarity classification task (Subtask 2) (see Section 2.1).

Seven research teams submitted several classification results to the Subtask 1, and four teams submitted to the Subtask 2. The systems submitted go in the line of the state of the art in similar workshops, and the participants developed classification systems based on Recurrent Neural Networks, Transformer Networks and fine-tunning models built upon BERT [2]. The details of the systems submitted are described in Sections 2.2 and 2.3.

2 Spanish Semantic Analysis Tasks

The workshop "Sentiment Analysis at SEPLN (TASS)" has been held since 2012, under the umbrella of the International Conference of the Spanish Society for Natural Language Processing (SEPLN).

Spanish is the second language used in Twitter, what calls for the development of new language comprehension systems and the opportunity of creation of resources for NLP and, more specifically, for sentiment analysis.

Many resources have been developed under TASS tasks. In this edition, we have completed the InterTASS corpus [4] with Uruguayan and Mexican Spanish variants. The workshop has been built over 2 general task: monolingual and multilingual approaches, over combinations of the five different datasets of Spanish language variants.

In this section we describe the entire InterTASS corpus and the two proposed tasks.

2.1 Corpus datasets

International TASS Corpus (InterTASS) is a corpus released in 2017 [3] that was updated in 2018 [4]. In this edition, it has been extended with new texts written in two new Spanish variants: Uruguayan and Mexican.

Therefore, this last version contains tweets written in five different variants of Spanish from Spain, Peru, Costa Rica, Uruguay and Mexico, and it exhibits a large amount of lexical and even structural differences in each variant. In this edition, participants have had to face this five different variants of Spanish to train and tests their systems.

Spanish dataset The Spanish dataset was released in 2017 as the first version of InterTASS. Its contains 3,401 tweets in Spanish by users from Spain, and it is a subset of a biggest corpus collected from July 2016 to January 2017. Each tweet was labeled with its level of polarity, which can be positive (P), neutral (NEU), negative (N) and no sentiment tag (NONE). Each tweet was annotated at least by three annotators. The dataset was originally split into three datasets that have been reorganized this year, in order to homogenize all the datasets. The new partitions contain a training set with 1,126 tweets, a development set with 569 tweets and a test set with 1,706 tweets. Table 1 shown the general statistics of Spanish dataset.

Table 1. Number of tweets per partition and class of the Spanish (from Spain) dataset

	Training	Dev.	Test
P	354	156	594
NEU	140	83	195
N	475	266	663
NONE	157	64	254
Total	1,126	569	1,706

Costa Rican dataset The Costa Rican dataset was created in 2018. It contains 2,363 tweets. The annotation methodology replicated the one used to label the Spanish dataset. Each tweet was labeled as positive (P), neutral (NEU), negative (N) and no sentiment tag (NONE). Every tweet was labeled by three annotators. Agreement was reached for 2,048 tweets. For the extra tweets, two more annotators were required to obtain agreement. The dataset also has been reorganized in order to homogenize the entire corpus. Table 2 shows the new composition.

Peruvian dataset This dataset is comprised by 3,005 tweets in the Peruvian Spanish variant. The annotation of the dataset was performed as follows. First, three annotators labeled all the tweets independently. Then, tweets with total

Table 2. Number of tweets per partition and class of the Costa Rican dataset

	Training	Dev.	Test
P	221	120	366
NEU	91	55	151
N	310	143	459
NONE	155	72	220
Total	777	390	1,196

or partial agreement (with agreement between two annotators at least) were included into the dataset. Tweets where annotators totally disagreed were labeled by two additional annotators. After this, the first annotator decided the label of the tweets where the disagreement continued. Finally, all tweets were included into the dataset. This partition has also been re-balanced. Table 3 shows the distribution of tweets according to classes in the Peruvian Spanish variant.

Table 3. Number of tweets per partition and class of the Peruvian dataset

	Training	Dev.	Test
P	216	105	435
NEU	170	163	368
N	228	107	485
NONE	352	200	176
Total	966	575	1,464

Uruguayan dataset The Uruguayan dataset is comprised of 2,857 tweets in the Uruguayan Spanish variant. The annotation process consisted of two phases. First, of all three annotators independently labeled all the tweets. After this first step, two more annotators relabeled the tweets that got three different votes in the first round. The few tweets that were still ambiguous after this process were discussed between the annotators in order to get a consensus. Table 4 shows the distribution of tweets in the Uruguayan Spanish variant.

Table 4. Number of tweets per partition and class of the Uruguayan dataset

	Training	Dev.	Test
P	290	153	469
NEU	192	90	290
N	367	192	587
NONE	94	51	82
Total	943	486	1,428

Mexican dataset The Mexican dataset contains 3,000 tweets in the Mexican Spanish variant. It was generated through a labeling process done with four annotators. This labeling process consisted of the following steps: 1) for each tweet, each annotator assigned a polarity of the set {P, N, NEU, NONE}; 2) from the labels assigned by all annotators, if there is a predominant label, this is assigned as the class of the tweet, and 3) in the case of no predominant label, another annotator intervened to obtain a predominant label for final assignment. The resulting distribution of tweets can be seen in Table 5.

Table 5. Number of tweets per partition and class of the Mexican dataset

	Training	Dev.	Test
P	313	159	525
NEU	79	51	119
N	505	252	745
NONE	93	48	111
Total	990	510	1,500

2.2 Task 1: Monolingual

The main goal of this task is the evaluation of polarity classification systems at tweet level for tweets written in Spanish in a monolingual environment. That is, the aim is to evaluate systems designed and trained for each individual variant.

The submitted systems will have to face up with the following challenges:

- Lack of context: the source elements are tweets.
- Informal language: misspelling, emojis and onomatopoeia are common.
- Multilinguality (local): the datasets have been developed with tweets written in the Spanish language spoken in Spain, Peru, Costa Rica, Uruguay and Mexico.
- Generalization: the systems will be assessed with several datasets of tweets written in the Spanish language spoken in different countries.

In this task, the participating teams could only perform monovariety experiments using InterTASS dataset (ES-Spain, PE-Peru, CR-Costa Rica, UR-Uruguay and MX-Mexico), so five rankings have been prepared, one for each Spanish variant.

Systems presented Seven teams presented their systems and results for this first task, whose main features are detailed below.

Atalaya Team [8] System inspired in [9]. Different representations of the data have been used, such as bag-of-words, bag-of-characters and tweet embeddings and they have trained robust subword-aware word embeddings and computed

tweet representations using a weighted-averaging strategy. The novelty of the system is the use of two data augmentation techniques to deal with data scarcity: two-way translation augmentation, and a novel technique that generates new instances by combining halves of tweets.

LaSTUS/TALN Team [1] The system proposes a deep learning approach based on bidirectional LSTM (biLSTM) models to face both sub-tasks. The tweets are tokenized keeping emojis and full hashtags and they are transformed in a embedding process.

GTH-UPM Team [6] The developed system consisted of three classifiers: a system based on feature vectors extracted from the tweets, a neural-based classifier using FastText and a deep neural network classifier using contextual vector embeddings created using BERT (Bidirectional Encoder Representations from Transformers). The averaged probability of the three classifiers was calculated to get the final score.

ELiRF-UPV [7] proposed a system focused mainly on employing the encoders of the Transformer model, based on self-attention mechanisms. The Transformer model dispenses with convolution and recurrences to learn long-range relationships. They use only the encoder part in order to extract vector representations that are useful to perform sentiment analysis. They denote this encoding part of the Transformer model as Transformer Encoder. The results obtained were very promising, being the first or second ranked system on almost all the Spanish variants.

The Titans [5] use a bidirectional LSTM based approach to capture information from both the past and future context followed by an attention layer consisting of initializers and regularizers.

RETUYT-InCo [10] presents three approaches for classifying the sentiment of tweets for different Spanish variants. In the first one, they consider multiple variants to perform a classification of the sentence word vectors mean, performing the classification through layered fully connected neural networks and support vector machines. The second approach relies on transfer learning from a pretrained Spanish BERT. The third approach is based on the use of FastText embeddings as input to an LSTM neural network. The MLP based approach achieved good results in monolingual experiments while the BERT based system performed better in the crosslingual task.

ITAINNOVA [11] explores two different deep learning approaches. The first one with an embedding-based strategy combined with bidirectional recurrent neural networks (an architecture that learns the representation of input documents as a concatenation of self-learned char-embeddings with sequence word-embeddings), and the second one using the new method of pre-trained BERT. Although the

performance of the second approach has not been presented as official results, it is reasonably remarkable and higher than the winner approach.

Tables 6, 7, 8, 9 and 10 show the results obtained on the Spain, Peru, Costa Rica, Uruguay and Mexico test datasets respectively. ELiRF-UPV team obtained the overall best results.

Table 6. Task-1: Monolingual Sentiment Analysis - Spain

Team	Macro F1	Macro Precision	Macro Recall
ELiRF-UPV	0.507	0.505	0.508
Atalaya	0.484	0.533	0.444
LaSTUS/TALN	0.464	0.47	0.457

Table 7. Task-1: Monolingual Sentiment Analysis - Peru

Team	Macro F1	Macro Precision	Macro Recall
Atalaya	0.454	0.462	0.446
ELiRF-UPV	0.447	0.456	0.439
RETUYT-InCo	0.438	0.437	0.439

Table 8. Task-1: Monolingual Sentiment Analysis - Costa Rica

Team	Macro F1	Macro Precision	Macro Recall
RETUYT-InCo	0.512	0.588	0.454
ELiRF-UPV	0.496	0.498	0.493
Atalaya	0.469	0.472	0.467

2.3 Task 2: Crosslingual

The purpose of this task is similar to that of Task 1, but systems must be trained with one or more Spanish variants and tested with a different Spanish variant. The Spanish variant of training set had to be different from the evaluation one, in order to test the dependency of systems on a language.

Six teams have participated in this task: *Atalaya Team*, *LaSTUS/TALN Team*, *GTH-UPM Team*, *The Titans Team*, *ITAINNOVA Team* and *RETUYT-InCo*. The systems are the same as those described in section 2.2.

Table 9. Task-1: Monolingual Sentiment Analysis - Uruguay

Team	Macro F1	Macro Precision	Macro Recall
ELiRF-UPV	0.515	0.497	0.536
Atalaya	0.499	0.498	0.499
GTH-ETSIT-UPM	0.492	0.521	0.466

Table 10. Task-1: Monolingual Sentiment Analysis - Mexico

Team	Macro F1	Macro Precision	Macro Recall
ELiRF-UPV	0.501	0.490	0.512
GTH-ETSIT-UPM	0.487	0.497	0.477
RETUYT-InCo	0.486	0.487	0.485

Tables 11, 12, 13, 14 and 15 show the results obtained on the test sets for Spanish variants of Spain, Peru, Costa Rica, Uruguay and Mexico respectively. In three of the five experiment results the *Atalaya team* obtained the best results, being the second in the evaluation of Spanish for Costa Rica.

The values obtained in the evaluation of this task are very similar to those of Task 1, although slightly lower, which is reasonable as no training data from the target Spanish variant was allowed.

Table 11. Task-2: Crosslingual - Spain

Team	Macro F1	Macro Precision	Macro Recall
RETUYT-InCo	0.460	0.456	0.465
LaSTUS/TALN	0.459	0.456	0.462
Atalaya	0.454	0.433	0.477

3 Conclusions

The 2019 edition of TASS has attracted the participation of 13 systems, seven for the first task (monolingual sentiment analysis), and six for the second task (crosslingual). Seven papers with the description of the evaluated systems were presented. This year, new datasets for the InterTASS corpus have been added, enlarging this reference corpus for the Spanish sentiment analysis task.

The submitted systems are in the line the state-of-the-art approaches in other similar workshops, and most of them are grounded in Deep Learning and the use of hand-crafted linguistic features.

As future work, we plan to consolidate the InterTASS corpus to the Spanish-speaking community, with new challenges for the next year. Moreover, we will

Table 12. Task-2: Crosslingual - Peru

Team	Macro F1	Macro Precision	Macro Recall
Atalaya	0.474	0.468	0.48
GTH-ETSIT-UPM	0.456	0.456	0.457
LaSTUS/TALN	0.448	0.442	0.454

Table 13. Task-2: Crosslingual - Costa Rica

Team	Macro F1	Macro Precision	Macro Recall
GTH-ETSIT-UPM	0.476	0.484	0.469
Atalaya	0.474	0.479	0.47
LaSTUS/TALN	0.465	0.472	0.458

Table 14. Task-2: Crosslingual - Uruguay

Team	Macro F1	Macro Precision	Macro Recall
Atalaya	0.514	0.517	0.510
GTH-ETSIT-UPM	0.481	0.458	0.507
LaSTUS/TALN	0.469	0.450	0.491

Table 15. Task-2: Crosslingual - Mexico

Team	Macro F1	Macro Precision	Macro Recall
Atalaya	0.473	0.474	0.471
GTH-ETSIT-UPM	0.471	0.465	0.476
RETUYT-InCo	0.465	0.455	0.474

keep working in the development of new corpora and linguistic resources for the research community.

References

1. Altin, L.S.M., Bravo, A., Saggion, H.: Lastus/taln at tass 2019: Sentiment analysis for spanish language variants with neural networks. In: Díaz-Galiano, M.C., García-Vega, M., Casasola, E., Chiruzzo, L., García-Cumbreras, M.A., Martínez-Cámara, E., Moctezuma, D., Montejo Ráez, A., Sobrevilla Cabezudo, M.A., Tellez, E., Graff, M., Miranda, S. (eds.) Proceedings of TASS 2019: Workshop on Semantic Analysis at SEPLN (TASS 2019). CEUR Workshop Proceedings, vol. ??? CEUR-WS, Bilbao, Spain (September 2019)
2. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019), https://www.aclweb.org/anthology/N19-1423
3. Díaz-Galiano, M.C., Martínez-Cámara, E., García Cumbreras, M.A., García Vega, M., Villena Román, J.: The democratization of deep learning in TASS 2017. Procesamiento del Lenguaje Natural **60**, 37–44 (2018), http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/5556
4. Daz-Galiano, M.C., Garca-Cumbreras, M.., Garca-Vega, M., Gutirrez, Y., Martnez Cmara, E., Piad-Morffis, A., Villena-Romn, J.: Tass 2018: The strength of deep learning in language understanding tasks. Procesamiento del Lenguaje Natural **62**, 77–84 (2019), http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/5955
5. Garain, A., Mahata, S.K.: Sentiment analysis at sepln (tass)-2019: Sentiment analysis at tweet level using deep learning. In: Díaz-Galiano, M.C., García-Vega, M., Casasola, E., Chiruzzo, L., García-Cumbreras, M.A., Martínez-Cámara, E., Moctezuma, D., Montejo Ráez, A., Sobrevilla Cabezudo, M.A., Tellez, E., Graff, M., Miranda, S. (eds.) Proceedings of TASS 2019: Workshop on Semantic Analysis at SEPLN (TASS 2019). CEUR Workshop Proceedings, vol. ??? CEUR-WS, Bilbao, Spain (September 2019)
6. Godino, I.G., DHaro, L.F.: Gth-upm at tass 2019: Sentiment analysis of tweets for spanish variants. In: Díaz-Galiano, M.C., García-Vega, M., Casasola, E., Chiruzzo, L., García-Cumbreras, M.A., Martínez-Cámara, E., Moctezuma, D., Montejo Ráez, A., Sobrevilla Cabezudo, M.A., Tellez, E., Graff, M., Miranda, S. (eds.) Proceedings of TASS 2019: Workshop on Semantic Analysis at SEPLN (TASS 2019). CEUR Workshop Proceedings, vol. ??? CEUR-WS, Bilbao, Spain (September 2019)
7. Ángel González, J., Hurtado, L.F., Pla, F.: Elirf-upv at tass 2019: Transformer encoders for twitter sentiment analysis in spanish. In: Díaz-Galiano, M.C., García-Vega, M., Casasola, E., Chiruzzo, L., García-Cumbreras, M.A., Martínez-Cámara, E., Moctezuma, D., Montejo Ráez, A., Sobrevilla Cabezudo, M.A., Tellez, E., Graff, M., Miranda, S. (eds.) Proceedings of TASS 2019: Workshop on Semantic Analysis at SEPLN (TASS 2019). CEUR Workshop Proceedings, vol. ??? CEUR-WS, Bilbao, Spain (September 2019)
8. Luque, F.M.: Atalaya at tass 2019: Data augmentation and robust embeddings for sentiment analysis. In: Díaz-Galiano, M.C., García-Vega, M., Casasola, E.,

Chiruzzo, L., García-Cumbreras, M.A., Martínez-Cámara, E., Moctezuma, D., Montejo Ráez, A., Sobrevilla Cabezudo, M.A., Tellez, E., Graff, M., Miranda, S. (eds.) Proceedings of TASS 2019: Workshop on Semantic Analysis at SEPLN (TASS 2019). CEUR Workshop Proceedings, vol. ??? CEUR-WS, Bilbao, Spain (September 2019)

9. Luque, F.M., Pérez, J.M.: Atalaya at tass 2018: Sentiment analysis with tweet embeddings and data augmentation. In: Martínez-Cámara, E., Almeida Cruz, Y., Díaz-Galiano, M.C., Estévez Velarde, S., García-Cumbreras, M.A., García-Vega, M., Gutiérrez Vázquez, Y., Montejo Ráez, A., Montoyo Guijarro, A., Muñoz Guillena, R., Piad Morffis, A., Villena-Román, J. (eds.) Proceedings of TASS 2018: Workshop on Semantic Analysis at SEPLN (TASS 2018). CEUR Workshop Proceedings, vol. 2172. CEUR-WS, Sevilla, Spain (September 2018)
10. Pastorini, M., Pereira, M., Zeballos, N., Chiruzzo, L., Rosá, A., Etcheverry, M.: Retuyt-inco at tass 2019: Sentiment analysis in spanish tweets. In: Díaz-Galiano, M.C., García-Vega, M., Casasola, E., Chiruzzo, L., García-Cumbreras, M.A., Martínez-Cámara, E., Moctezuma, D., Montejo Ráez, A., Sobrevilla Cabezudo, M.A., Tellez, E., Graff, M., Miranda, S. (eds.) Proceedings of TASS 2019: Workshop on Semantic Analysis at SEPLN (TASS 2019). CEUR Workshop Proceedings, vol. ??? CEUR-WS, Bilbao, Spain (September 2019)
11. nés Salas, R.M.M., del Hoyo-Alonso, R., Aznar-Gimeno, R.: From recurrency to attention in opinion analysis. comparing rnn vs transformer models. In: Díaz-Galiano, M.C., García-Vega, M., Casasola, E., Chiruzzo, L., García-Cumbreras, M.A., Martínez-Cámara, E., Moctezuma, D., Montejo Ráez, A., Sobrevilla Cabezudo, M.A., Tellez, E., Graff, M., Miranda, S. (eds.) Proceedings of TASS 2019: Workshop on Semantic Analysis at SEPLN (TASS 2019). CEUR Workshop Proceedings, vol. ??? CEUR-WS, Bilbao, Spain (September 2019)