Overview of TASS 2020: Introducing Emotion Detection

Manuel García-Vega[a], Manuel Carlos Díaz-Galiano[a], Miguel Á. García-Cumbreras[a], Flor Miriam Plaza del Arco[a], Arturo Montejo-Ráez[a], Salud María Jiménez-Zafra[a], Eugenio Martínez Cámara[b], César Antonio Aguilar[c], Marco Antonio Sobrevilla Cabezudo[d], Luis Chiruzzo[e] and Daniela Moctezuma[f]

[a]*CEATIC. Universidad de Jaén, Jaén, España*
[b]*Andalusian Research Institute in Data Science and Computational Intelligence (DaSCI), Universidad de Granada, España*
[c]*Universidad Católica de Chile, Chile*
[d]*Universidade de São Paulo, Brasil*
[e]*Universidad de la República Montevideo, Uruguay*
[f]*CONACyT-CentroGEO, México*

Abstract

The Task on Semantic Analysis at SEPLN (TASS task within IberLEF 2020 workshop) took place on September 22, reaching its ninth edition. Due to the COVID-19 pandemic, the number of participants is lower compared to past campaigns. Also, the organizers decided to held it remotely. In this edition, the classical polarity classification subtask was, again, organized. As a novelty, a second subtask was proposed to foster research in emotion detection of Spanish texts on a new dataset. This paper summarizes the different approaches of the teams who participated, the key insights of their systems and the results obtained for all the proposed solutions.

Keywords

Sentiment Analysis, Opinion Mining, Social Media.

1. Introduction

Sentiment Analysis (SA) is still a challenging task, because the difficulties of processing some underlying linguistic phenomena as negation, irony, sarcasm and more broadly speaking the subjectivity of the opinions. However, there are additional uses of subjective language to

Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020)
EMAIL: mgarcia@ujaen.es (M. García-Vega); mcdiaz@ujaen.es (M.C. Díaz-Galiano); magc@ujaen.es (M.Á. García-Cumbreras); fmplaza@ujaen.es (F.M.P.d. Arco); amontejo@ujaen.es (A. Montejo-Ráez); sjzafra@ujaen.es (S.M. Jiménez-Zafra); emcamara@decsai.ugr.es (E.M. Cámara); caguilara@uc.cl (C.A. Aguilar); msobrevillac@usp.brs (M.A.S. Cabezudo); luischir@fing.edu.uy (L. Chiruzzo); dmoctezuma@centrogeo.edu.mx (D. Moctezuma)
URL: https://www.ujaen.es/ (M. García-Vega); https://www.ujaen.es/ (M.C. Díaz-Galiano); https://www.ujaen.es/ (M.Á. García-Cumbreras); https://www.ujaen.es/ (F.M.P.d. Arco); https://www.ujaen.es/ (A. Montejo-Ráez); https://www.ujaen.es/ (S.M. Jiménez-Zafra)
ORCID: 0000-0003-2850-4940 (M. García-Vega); 0000-0001-9298-1376 (M.C. Díaz-Galiano); 0000-0003-1867-9587 (M.Á. García-Cumbreras); 0000-0002-3020-5512 (F.M.P.d. Arco); 0000-0002-8643-2714 (A. Montejo-Ráez); 0000-0003-3274-8825 (S.M. Jiménez-Zafra)

express private states beyond opinions [1]. Humans also express their private states through the description of their emotional state. Hence, emotion analysis (EA) represents an additional challenge to the processing of subjective language.

The task on semantic analysis at SEPLN (TASS), belonging to the IberLEF workshop, addressed as a shared task the main challenges in SA in Spanish. The edition of TASS 2020 has as main novelty to face up the task of EA in Spanish. Specifically, TASS 2020 proposes two subtasks:

1. General polarity at three levels. It corresponds to the default subtask of TASS, which consists of classifying the opinion meaning of tweets written in Spanish. Since the edition of 2017, one of the aims of TASS is the elaboration of a representative Spanish dataset for SA in Twitter. Accordingly, the InterTASS dataset was presented and used for the first time in the edition of 2017 [2]. The InterTASS dataset is composed of tweets written in the Spanish spoken in Spain and in some countries of America. In this edition, the InterTASS dataset has been enlarged with tweets written in the Spanish spoken in Chile.
2. Emotion detection. It proposes to classifies the emotion expressed in tweets written in Spanish at seven levels of emotion intensity, namely the six Ekman's basic emotions [3] and "*neutral or no emotion*".

Three research teams submitted several classification results to the Subtask 1, and two teams submitted to the Subtask 2.

The details of the systems submitted are described in Sections 2.2 and 2.3.

2. Spanish Semantic Analysis Tasks

The International Conference of the Spanish Society for Natural Language Processing (SEPLN) supports it and the workshop is held within the SEPLN 2020 event.

Spanish is the fourth most spoken language and the second one on Twitter. These facts inspire TASS to create new systems of language comprehension and new resources for NLP and, more specifically, for sentiment analysis.

Many resources have been developed within the framework of the TASS tasks. In this edition, InterTASS [4] and EmoEvent [5] have been used. The workshop has been built on 2 general tasks.

Task 1, General polarity at three levels, with two sub-tasks:

- **Monolingual**, in which only the dataset of the variety of Spanish to be undertaken has been used.
- **Multivariant**, in which the test dataset has been built with a set of tweets from all the varieties in the corpus.

Task 2, Emotion detection. It has been designed to encourage research in the understanding the emotions expressed by users on social media. This is a hard task due to the absence of voice modulations and facial expressions.

The task consists of classifying the emotion expressed in a tweet as "*neutral or no emotion*" or as one of the six Ekman's basic emotions: *anger*, *disgust*, *fear*, *joy*, *sadness* and *surprise*.

In this section we describe in detail the corpus used and the two proposed tasks.

2.1. Systems presented

In this edition, only three teams presented their systems and results, whose main features are detailed below.

Palomino Team [6] system applies transfer learning techniques using the BERT language model and a procedure to augment the training data set and prevent overfitting.

UMUTeam [7] system is based on the use of linguistic features in combination with word-embeddings. It tests convolutional neural networks and supporting vector machines with sentence embedding. Moreover, it shows that the use of this combination performs better than separately. The system uses UMUTextStats, a linguistic tool developed in-house, to extract the above-mentioned linguistic features.

ELiRF-UPV [8] system adapt BERT model on Spanish tweets since the datasets contain texts from Twitter. The authors compare the adapted system with the Deep Averaging Networks system they stablished as baseline.

2.2. Task 1: General polarity at three levels

International TASS Corpus (InterTASS) was released in 2017 [9], updated in 2018 [10] and completed in 2019 [4]. This is the corpus used in this task.

In this edition, we have introduced a novelty: the delivered corpora contains only 3 different tags or levels of opinion intensity (P, N, NEU), where NEU includes NONE.

Evaluation measures: accuracy and the macro-averaged versions of Precision, Recall, and F1 have been used as evaluation measures. Systems were ranked by the Macro-F1 and Accuracy measures.

2.2.1. Datasets

For sub-task 1.1, monolingual classification, the training and evaluation used each data set of the InterTASS varieties (ES-Spain, CR-Costa Rica, PE-Peru, UY-Uruguay and MX-Mexico), but, unlike in past editions, the use of any other corpora or linguistic resources has been allowed.

For sub-task 1.2, multivariant classification, a new test dataset was delivered, with tweets extracted from the different test subsets of Spain, Costa Rica, Peru, Uruguay, and Mexico. These tweets were specially selected from the most difficult tweets of the last edition (tweets that were missed by most teams). Again, unlike in past editions, the use of any other corpora or linguistic resources has been allowed.

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

- Lack of context: the source elements are tweets.
- Informal language: misspelling, emojis and onomatopoeia are common.
- Multivariety: the datasets have been developed with tweets written in the Spanish language spoken in Spain, Peru, Costa Rica, Uruguay and Mexico.

Table 1
Final ranking of task 1.1: General polarity at three levels. Monolingual: Spanish (from Spain) variant sub-task

Team	Set	F1	Precision	Recall
ELiRF-UPV	ES	0.671	0.673	0.670
Palomino-Ochoa	ES	0.665	0.665	0.664
UMUTeam	ES	0.503	0.561	0.456

Table 2
Final ranking of task 1.1: General polarity at three levels. Monolingual: Costa Rican variant sub-task

Team	Set	F1	Precision	Recall
ELiRF-UPV	CR	0.6464	0.647	0.646
Palomino-Ochoa	CR	0.6463	0.644	0.649
UMUTeam	CR	0.3498	0.350	0.350

Table 3
Final ranking of task 1.1: General polarity at three levels. Monolingual: Peruvian variant sub-task

Team	Set	F1	Precision	Recall
ELiRF-UPV	PE	0.635577	0.672	0.603
Palomino-Ochoa	PE	0.633577	0.642	0.626
UMUTeam	PE	0.389992	0.397	0.383

Table 4
Final ranking of task 1.1: General polarity at three levels. Monolingual: Uruguayan variant sub-task

Team	Set	F1	Precision	Recall
Palomino-Ochoa	UY	0.669	0.671	0.667
ELiRF-UPV	UY	0.654	0.667	0.642
UMUTeam	UY	0.520	0.518	0.522

- Generalization: the systems will be assessed with several datasets of tweets written in the Spanish language spoken in different countries

2.2.2. Subtask 1: Monolingual

The goal with this task is to evaluate how systems classify polarity at tweet level, for tweets written in Spanish, in an environment of a previously known variety of Spanish. That is, each variety of Spanish has a test corpus that must be evaluated using any source.

Tables 1, 2, 3, 4 and 5 show the final results for Spanish, Costa Rican, Peruvian, Uruguayan and Mexican variant sub-tasks.

2.2.3. Subtask 2: Multivariant

The goal with this task is to evaluate how systems classify polarity at tweet level, for tweets written in Spanish, in a multivariant environment.

Table 5
Final ranking of task 1.1: General polarity at three levels. Monolingual: Mexican variant sub-task

Team	Set	F1	Precision	Recall
ELiRF-UPV	mx	0.634	0.637	0.633
Palomino-Ochoa	mx	0.633	0.636	0.630
UMUTeam	mx	0.397	0.397	0.396

Table 6
Final ranking of task 1.2: General polarity at three levels. Multivariant sub-task

Team	F1	Precision	Recall
Palomino-Ochoa	0.498	0.487	0.510
ELiRF-UPV	0.358	0.359	0.357

The training corpus that has been offered is the entire InterTASS. Furthermore, a new test dataset has been developed, with tweets extracted from most difficult tweets, based on last year results, from the different variant subsets. The use of any linguistic body or resource has been allowed.

The process of generating the test dataset was as follows:

- A ranking has been created for the test dataset of each variety, according to the number of teams that got the right answer in task 2 of TASS'2019.
- We have selected the 300 most difficult tweets of each variety and included them in the test dataset of sub-task 1.2.

Table 6 shows the final results for multivariant sub-task

2.3. Task 2: Emotion detection

This task has been introduced for the first time this year (TASS 2020) with the purpose of encouraging research in the emotion detection task for Spanish. While polarity classification is a well-established task with many standard datasets and well-defined methodologies, emotion detection has received less attention due to its complexity. In fact, it can be considered a further step in the task of polarity classification since it consists of detecting fine-grained emotions in text, not just positive or negative polarity.

The goal of Task 2: Emotion detection is to classify the main emotion expressed in a tweet as "*neutral or no emotion*" or as one of the six Ekman's basic emotions [3] that best represent the mental state of the tweeter. The emotion categories are listed below, along with some synonyms:

- *Joy*, including serenity and ecstasy.
- *Sadness*, including pensiveness and grief.
- *Anger*, including annoyance and rage.

Table 7
Distribution of emotions by subset (Train, Dev, Test) in EmoEvent dataset for Task 2.

Emotion	Train	Dev	Test
Joy	1,270	185	360
Sadness	706	103	200
Anger	600	87	170
Surprise	241	35	68
Disgust	113	16	32
Fear	67	10	19
Others	2,889	421	817
Total	5,886	857	1,666

- *Surprise*, including distraction and amazement.
- *Disgust*, including disinterest, dislike, and loathing.
- *Fear*, including apprehension, anxiety, concern and terror.
- *Others*, no emotion or neutral.

2.3.1. Emotion corpus

For this task, we use EmoEvent [5], a multilingual emotion corpus of tweets based on events that took place in April 2019. They are related to different domains such as entertainment, catastrophe, political, global commemoration, and global strike. Each instance in the dataset were labeled with the main emotion expressed in the tweet by three annotators according to the following categories: *anger*, *disgust*, *fear*, *joy*, *sadness*, *surprise*, "*neutral or no emotion*". The authors decided the final emotion label of the tweet as the majority emotion labeled by the annotators, but in case the three annotators labeled the tweet with different emotions, the final label is "*neutral or no emotion*". In particular, for this task we use the Spanish version of EmoEvent that contains a total of 8,409 tweets written in Spanish.

With the purpose of providing the dataset to the participants, we decided to replace the hashtags by the keyword *HASHTAG* in order to prevent the automatic classifier from relying on hashtags to categorize the emotion associated with a tweet. Moreover, we replaced the user mentions by *@USER* to anonymize mentions to users. Finally, training, development and test sets have been released for the participants. Table 7 shows the number and percentage of tweets corresponding to each partition by emotion.

2.3.2. System participant results

The systems submitted have had to cope with the following challenges:

- Lack of context: the source elements are tweets and the length of each tweet is limited (up to 240 characters).
- Informal language: misspellings, emojis, onomatopoeias are common in tweets.

Table 8
Final raking of task 2: Emotion detection

Team	F1	Precision	Recall
ELiRF-UPV	0.447	0.443	0.450
UMUTeam	0.379	0.420	0.345

- Multiclass classification: Each tweet is labeled with one of the following seven different categories: *anger*, *fear*, *sadness*, *joy*, *disgust*, *surprise*, "*neutral or no emotion*".

For Task 2: Emotion Detection, two teams have submitted their systems. Table 8 shows the final results obtained by the participant systems on the test set of Task 2. The ELiRF-UPV team obtained the best macro averaged F1-score of 0.447. They take advantage of BERT by adapting the model on Spanish tweets, then they established a baseline based on Deep Averaging Networks to compare their results. The UMUTeam has achieved the performance of 0.379, presenting a system based on the combination of linguistic features and word embeddings. As a classifier, they use the sequential minimal optimisation algorithm which is based on support vector machines.

3. Conclusions

The 2020 edition of TASS continues to work with the InterTASS dataset, introducing a new task that works with a multivariant test set, common to all participating teams, and introduces as a novelty a new task: emotion detection.

Although 14 participating teams were registered, in the end only three presented results. In the different tasks with the InterTASS corpus we highlight the robustness of the systems, without significant variations between monolingual tasks and the new multivariate task.

Regarding the task of emotion detection, taking into account that the dataset provided is not balanced among the eight categories of emotion raised, the systems presented results that pose many improvements in the future to address this complex task, pending an in-depth analysis of each system, which will be reflected in the paper of each participant.

The future of TASS is to establish as a main line of work this detection of emotions as a natural evolution of the analysis of feelings.

Acknowledgments

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

References

[1] R. Quirk, S. Greenbaum, G. Leech, J. Svartvik, A Comprehensive Grammar of the English Language, Longman, London, 1985.

[2] M. C. Díaz-Galiano, E. Martínez-Cámara, M. A. García Cumbreras, M. García Vega, J. Villena Román, The democratization of deep learning in TASS 2017, Procesamiento del Lenguaje Natural 60 (2018) 37–44. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/5556.

[3] P. Ekman, Are there basic emotions? 99(3) (1992) 550––553.

[4] M. C. Díaz-Galiano, M. García-Vega, E. Casasola, L. Chiruzzo, M. A. García-Cumbreras, E. Martínez-Cámara, D. Moctezuma, A. Montejo Ráez, M. A. Sobrevilla Cabezudo, E. Tellez, M. Graff, S. Miranda, Overview of tass 2019: One more further for the global spanish sentiment analysis corpus, in: M. C. Díaz-Galiano, M. García-Vega, E. Casasola, L. Chiruzzo, M. A. García-Cumbreras, E. Martínez-Cámara, D. Moctezuma, A. Montejo Ráez, M. A. Sobrevilla Cabezudo, E. Tellez, M. Graff, S. Miranda (Eds.), Proceedings of TASS 2019: Workshop on Semantic Analysis at SEPLN (TASS 2019), volume 2421 of *CEUR Workshop Proceedings*, CEUR-WS, Bilbao, Spain, 2019.

[5] F. M. Plaza del Arco, C. Strapparava, L. A. Urena Lopez, M. Martin, EmoEvent: A multilingual emotion corpus based on different events, in: Proceedings of The 12th Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 1492–1498. URL: https://www.aclweb.org/anthology/2020.lrec-1.186.

[6] D. Palomino, J. Ochoa-Luna, Palomino-ochoa at TASS 2020: Single BERT-based approach to overcome spanish sentiment analysis, in: Proceedings of TASS 2020: Workshop on Semantic Analysis at SEPLN (TASS 2020), volume ??? of *CEUR Workshop Proceedings*, CEUR-WS, Málaga, Spain, 2020.

[7] J. A. García-Díaz, A. Álmela, R. Valencia-García, Umuteam at tass 2020: Combining linguistic features and machine-learning models for sentiment classification, in: Proceedings of TASS 2020: Workshop on Semantic Analysis at SEPLN (TASS 2020), volume ??? of *CEUR Workshop Proceedings*, CEUR-WS, Málaga, Spain, 2020.

[8] J. Ángel González, L.-F. Hurtado, F. Pla, J. A. Moncho, ELiRF-UPV at TASS 2020: TWilBERT for sentiment analysis and emotion detection in spanish tweets, in: Proceedings of TASS 2020: Workshop on Semantic Analysis at SEPLN (TASS 2020), volume ??? of *CEUR Workshop Proceedings*, CEUR-WS, Málaga, Spain, 2020.

[9] E. Martínez-Cámara, M. C. Díaz-Galiano, M. A. García-Cumbreras, M. García-Vega, J. Villena-Román, Overview of TASS 2017, in: E. Martínez-Cámara, M. C. Díaz-Galiano, M. A. García-Cumbreras, M. García-Vega, J. Villena-Román (Eds.), Proceedings of TASS 2017: Workshop on Semantic Analysis at SEPLN (TASS 2017), volume 1896 of *CEUR Workshop Proceedings*, CEUR-WS, Murcia, Spain, 2017.

[10] E. Martínez-Cámara, Y. Almeida-Cruz, M. C. Díaz-Galiano, S. Estévez-Velarde, M. A. García-Cumbreras, M. García-Vega, Y. Gutiérrez, A. Montejo-Ráez, A. Montoyo, R. Muñoz, A. Piad-Morffis, J. Villena-Román, Overview of TASS 2018: Opinions, health and emotions, in: E. Martínez-Cámara, Y. Almeida Cruz, M. C. Díaz-Galiano, S. Estévez Velarde, M. A. García-Cumbreras, M. García-Vega, Y. Gutiérrez Vázquez, A. Montejo Ráez, A. Montoyo Guijarro, R. Muñoz Guillena, A. Piad Morffis, J. Villena-Román (Eds.), Proceedings of TASS 2018: Workshop on Semantic Analysis at SEPLN (TASS 2018), volume 2172 of *CEUR Workshop Proceedings*, CEUR-WS, Sevilla, Spain, 2018.