TASS 2018: Workshop on Semantic Analysis at SEPLN, septiembre 2018, págs. 13-27

Overview of TASS 2018: Opinions, Health and Emotions

Resumen de TASS 2018: Opiniones, Salud y Emociones

Eugenio Martínez-Cámara[1], **Yudivián Almeida-Cruz**[2], **Manuel Carlos Díaz-Galiano**[3] **Suilan Estévez-Velarde**[2], **Miguel Á. García-Cumbreras**[3], **Manuel García-Vega**[3], **Yoan Gutiérrez**[4], **Arturo Montejo-Ráez**[3], **Andrés Montoyo**[4], **Rafael Muñoz**[4], **Alejandro Piad-Morffis**[2], **Julio Villena-Román**[5]

[1]Andalusian Research Institute in Data Science and Computational Intelligence (DaSCI) Universidad de Granada, España

[2]Universidad de La Habana, Cuba

[3]Centro de Estudios Avanzados en Tecnologías de la Información y de la Comunicación (CEATIC) Universidad de Jaén

[4]Universidad de Alicante, España

[5]MeaningCloud

Abstract: This is an overview of the Workshop on Semantic Analysis at the SEPLN congress held in Sevilla, Spain, in September 2018. This forum proposes to participants four different semantic tasks on texts written in Spanish. Task 1 focuses on polarity classification; Task 2 encourages the development of aspect-based polarity classification systems; Task 3 provides a scenario for discovering knowledge from eHealth documents; finally, Task 4 is about automatic classification of news articles according to safety. The former two tasks are novel in this TASS's edition. We detail the approaches and the results of the submitted systems of the different groups in each task.

Keywords: Sentiment Analysis, Opinion Mining, Affect Computing, eHealth, Social Media

Resumen: Este artículo ofrece un resumen sobre el Taller de Análisis Semántico en la SEPLN (TASS) celebrado en Sevilla, España, en septiembre de 2018. Este foro propone a los participantes cuatro tareas diferentes de análisis semántico sobre textos en español. La Tarea 1 se centra en la clasificación de la polaridad; la Tarea 2 anima al desarollo de sistemas de polaridad orientados a aspectos; la Tarea 3 consiste en descubrir conocimiento en documentos sobre salud; finalmente, la Tarea 4 propone la clasificación automática de noticias periodísticas según un nivel de seguridad. Las dos últimas tareas son nuevas en esta edición. Se ofrece una síntesis de los sistemas y los resultados aportados por los distintos equipos participantes, así como una discusión sobre los mismos.

Palabras clave: Análisis de Sentimientos, Minería de Opiniones, Informática Afectiva, e-Salud, Medios Sociales

1 Introduction

The Workshop on Semantic Analysis at the SEPLN[1] (in Spanish *Taller de Análisis Semántico en la SEPLN*, TASS) is the evolution of the Workshop on Sentiment Analysis at the SEPLN which is being held since 2012. The aim of the workshop is the furtherance of the research in semantic tasks on texts written in Spanish, roughly speaking in Spanish data. The edition 2018 has proposed two new challenges (Tasks 3 and 4), and provided several linguistic resources.

The processing of health data is attracting the attention of the Natural Language Processing (NLP) research community (Denecke,

[1] http://www.sepln.org/workshops/tass

ISSN 1613-0073

2015). In this line, Task 3 proposes modelling the human language in a scenario in which Spanish electronic health documents could be machine readable from a semantic point of view. This Task 3 consists of detecting and classifying concepts for semantic relating them. Task 4 is related to the brand safety concept, which is crucial for the reputation of a brand or the company of the brand. Task 4 proposes the classification of the level of safety of a news for the publication of a ads spot of a brand according to the headline of that news.

Tasks 3 and 4 provided specific datasets for accomplishing the proposed challenge, and are described in Sections 2.3.1 and 2.4.1 respectively. Task 1 provided an extension of the InterTASS corpus, that was presented in the edition of 2017 (Martínez-Cámara et al., 2017). The main novelty of the new version of InterTASS is the incorporation of tweets written in the Spanish language spoken in Spain and in the several other countries of America. Since the difficulty of Task 2 is high, the organisation proposed the same setting of the task as in previous editions.

The paper is organised as follows: Section 2 describes all the tasks proposed in the edition of year 2018. The specific details of each Subtask are in Section 2.1, 2.2, 2.3 and 2.4 respectively. Section 3 exposes the conclusions of the paper.

2 Spanish Semantic Analysis Tasks

As mentioned before, TASS is a relevant workshop for semantic analysis tasks, particularly for Spanish. In 2018, new resources and challenges were introduced to evolve Sentiment Analysis systems to a semantic level. In the last editions, several research groups from different countries, like Uruguay or Costa Rica, presented their systems, and it was mandatory to make an effort to build adequate resources for their languages.

In addition, society and companies are interested in new specific challenges, and for this reason new tasks arise, while maintaining the main task (global polarity).

In this Section, we describe the four tasks of the edition of 2018, namely Section 2.1 expose the details of Task 1; Section 2.2 describes the corpus and the systems submitted to Task 2; Section 2.3 is focused on the Task 3; and Section 2.4 describes all details of Task 4.

2.1 Task 1

This task focused on the evaluation of polarity classification systems at tweet level of tweets written in Spanish.

The submitted systems had to face, as usual, the lack of context due to length of tweets written in an informal language with misspelling or emojis, even onomatopeias. But this edition brought new challenges to this task:

- Multilinguality: *training*, *tests* and *development* corpus contain tweets written in Spanish from Spain, Peru and Costa Rica.
- Generalization: Several corpora have been used. One of them is the *development* set, so it follows a similar distribution. The second corpus is the *test* set of the General Corpus of TASS, which was compiled some years ago, so it may be lexically and semantically different from the *training* and *development* data. Furthermore, the system will be evaluated with *test* sets of tweets written in the Spanish language spoken in different American countries.

The General Corpus of TASS has been provided in the same way as previous editions. Further details in (Martínez-Cámara et al., 2017).

However, International TASS Corpus (InterTASS) is a corpus released in 2017 that has been updated for this edition with new texts. It is composed of tweets written in different varieties of Spanish (for Spain, Peru and Costa Rica), so it exhibits a large amount of lexical and even structural differences in each variant. The main purpose of compiling and using an inter-varietal corpus of Spanish for the evaluation tasks is to challenge participating systems to cope with the many faces of this language worldwide.

Datasets were annotated with 4 different polarity labels POSITIVE, NEGATIVE, NEUTRAL and NONE), and systems had to identify the orientation of the opinion expressed in each tweet in any of those 4 polarity levels.

The Spanish variety part was released in 2017 and its description can be found in (Martínez-Cámara et al., 2017). Table 1 shows the tweets distribution for *training*, *development* (dev.) and *test* corpora.

	Training	Dev.	Test
P	317	156	642
NEU	133	69	216
N	416	219	767
NONE	138	62	274
Total	1,008	506	1,899

Table 1: Tweets distribution in InterTASS-ES

The Peru and Costa Rica varieties have been released for this edition. The tweets distributions are shown in Tables 2 and 3 respectively for both variants.

	Training	Dev.	Test
P	231	95	430
NEU	166	61	367
N	242	106	472
NONE	361	238	159
Total	1,000	500	1,428

Table 2: Tweets distribution in InterTASS-PE

	Training	Dev.	Test
P	230	93	354
NEU	94	39	164
N	311	110	491
NONE	165	58	224
Total	800	300	1,233

Table 3: Tweets distribution in InterTASS-CR

Four sub-tasks were proposed, working with the datasets of the different countries:

Subtask-1: Monolingual ES. *training* and *test* were the InterTASS ES datasets.

Subtask-2: Monolingual PE. *training* and *test* were the InterTASS PE datasets.

Subtask-3: Monolingual CR. *training* and *test* were the InterTASS CR datasets.

Subtask-4: Cross-lingual. The *training* could be done with any dataset, but using a different one for the evaluation, in order to test the dependency of systems on a language.

Results were submitted in a plain text file with the following format:

```
tweet_id \t polarity
```

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

2.1.1 Analysis of the Results

For task 1 five system were presented. Most of them make use of deep learning algorithms, combining different ways of obtaining the word embeddings.

INGEOTEC. Moctezuma1 et al. (2018) present a polarity classification system based on the combination of different labelling systems. The main component is the EvoMSA system, based on genetic algorithms, which combines the outputs of the other systems. EvoMSA is based on the B4MSA system for the adjustment of the different parameters (how the text is normalised, how the tokens are calculated or how the tokens are weighted) and on the EvoDAG program that carries out the classification. As for the input systems, various systems are used based on lexicons of affectivity or aggressiveness. It also uses the algorithm of word embeddings called FastText, using the Wikipedia in Spanish to train it. Vectors are generated for each document and SVM is used for training. Their approach performs better when it is trained with tweets from Spain and test with other Spanish varieties.

RETUYT-InCo. Chiruzzo and Rosá (2018) submitted three approaches: SVM using word embedding centroids and manually crafted features, CNN using word embeddings as input, and Long Short Term Memory (LSTM) using word embeddings, trained with focus on improving the recognition of neutral tweets. In all cases, embedding improves results and LSTM has the best behaviour for neutral tweets. The use of a mixed-balanced training method for the LSTM resulted in a significant improvement in the detection of neutral tweets.

ITAINNOVA. Montanés, Aznar, and del Hoyo (2018) analyse the use of convolutional network models (CNN), LSTM, Bidirectional LSTM (BiLSTM) and a hybrid approach between CNN and LSTM. The combination CNN-LSTM has been chosen as it integrates the benefits of both models. They choose the CNN-LSTM combination because it integrates the benefits provided from both models.

Run	M. F1	Acc.
elirf-es-run-1	0.503	0.612
retuyt-lstm-es-1	0.499	0.549
retuyt-lstm-es-2	0.498	0.514
retuyt-combined-es	0.491	0.602
elirf-es-run-2	0.489	0.593
atalaya-ubav3-100-3-syn	0.476	0.544
retuyt-svm-es-2	0.473	0.584
atalaya-lr-50-2-bis	0.468	0.599
atalaya-lr-50-2	0.461	0.598
atalaya-ubav3-50-3	0.460	0.583
retuyt-cnn-es-1	0.458	0.592
atalaya-lr-50-2-roc	0.455	0.595
ingeotec-run1	0.445	0.530
retuyt-cnn-es-2	0.445	0.574
atalaya-svm-50-2	0.431	0.583
itainnova-cl-base	0.383	0.433
itainnova-cl-proc1	0.320	0.395
retuyt-cnn-es-1	0.097	0.096

Table 4: Task 1: InterTASS Monolingual ES

ELiRF-UPV. González, Hurtado, and Pla (2018b) explore different approaches based on Deep Learning. Specifically, they study the behaviour of the CNN, Attention Bidirectional Long Short Term Memory (Att-BLSTM) and Deep Averaging Networks (DAN). In order to study the behaviour of the different models, they carry out an adjustment process. They get the best results in InterTASS-ES. However, linguistic variability affects the choice of architecture and its hyperparameters, so the application of the same system to InterTASS-CR and InterTASS-PE tasks, without making any adjustment, has not allowed to obtain results as competitive as in InterTASS-ES.

ATALAYA. Luque and Pérez (2018) presented a system that uses a weighted scheme to average the subword-aware embeddings obtained from preprocessed tweets that have been enriched with data obtained from machine translation. This novel solution involves translating tweets into another language and back into the source language, to lexically and grammatically increase them.

Tables 4, 5 and 6 show the results obtained in the monolingual subtasks (Spain, Costa Rica and Peru variants).

For the cross-lingual runs, the participants selected an InterTASS dataset to train their systems and a different one to test, in order to test the dependency of systems on a language. Tables 7, 9 and 8 show the results obtained in these cross-lingual subtasks.

Run	M. F1	Acc.
retuyt-lstm-cr-2	0.504	0.537
retuyt-svm-cr-2	0.499	0.577
retuyt-svm-cr-1	0.493	0.567
elirf-cr-run-2	0.482	0.561
retuyt-cnn-cr-1	0.477	0.569
atalaya-cr-lr-50-2	0.475	0.582
ingeotec-run1	0.474	0.522
retuyt-lstm-cr-1	0.473	0.530
retuyt-cnn-cr-2	0.469	0.563
elirf-intertass-cr-run-1	0.463	0.544
atalaya-mlp-300-sentiment	0.439	0.520
atalaya-mlp-ubav3-50-3	0.436	0.560
ingeotec-run1	0.384	0.398
elirf-cr-run-1	0.317	0.288

Table 5: Task 1: InterTASS Monolingual CR

Run	M. F1	Acc.
retuyt-cnn-pe-1	0.472	0.494
atalaya-pe-lr-50-2	0.462	0.451
retuyt-lstm-pe-2	0.443	0.488
retuyt-svm-pe-2	0.441	0.471
ingeotec-run1	0.439	0.447
elirf-intertass-pe-run-2	0.438	0.461
atalaya-mlp-sentiment-ubav3-50-3	0.437	0.520
retuyt-svm-pe-1	0.437	0.474
elirf-intertass-pe-run-1	0.435	0.440
atalaya-mlp-300-sentiment	0.429	0.395
atalaya-mlp-50-sentiment	0.427	0.501
retuyt-svm-pe-2	0.425	0.477
retuyt-cnn-pe-2	0.425	0.477
retuyt-lstm-pe-1	0.419	0.420
elirf-intertass-pe-run-1	0.225	0.210

Table 6: Task 1: InterTASS Monolingual PE

The overall results, in terms of F1, obtained with the monolingual and multilingual systems for the Spanish and Costa Rica collections are quite comparable, but the one with the Peru collection fall by around 10%.

2.2 Task 2

Task 2, Aspect-based Sentiment Analysis, proposes the development of aspect-based polarity classification systems. Similar to previous editions (Martínez-Cámara et al., 2017), two datasets were used to evaluate the different approaches: Social-TV and STOMPOL. Both datasets were annotated with

Run	M. F1	Acc.
retuyt-svm-cross-es-2	0.471	0.555
retuyt-lstm-cross-es-2	0.470	0.466
retuyt-svm-cross-es-1	0.464	0.572
retuyt-cnn-cross-es-1	0.450	0.524
retuyt-cnn-cross-es-2	0.448	0.563
ingeotec-run1	0.445	0.530
atalaya-mlp-300-sentiment	0.441	0.485
retuyt-lstm-cross-es-1	0.438	0.498

Table 7: Task 1: InterTASS Cross-lingual with ES as test

Run	M. F1	Acc.
ingeotec-run1	0.447	0.506
retuyt-svm-cross-pe-2	0.445	0.514
retuyt-svm-cross-pe-1	0.444	0.505
retuyt-lstm-cross-pe-2	0.444	0.465
atalaya-mlp-300-sentiment	0.438	0.523
retuyt-lstm-cross-pe-1	0.425	0.472
retuyt-cnn-cross-pe-1	0.409	0.481
retuyt-cnn-cross-pe-2	0.391	0.438
itainnova-cl-base-cross-PE	0.367	0.382

Table 8: Task 1: InterTASS Cross-lingual with PE as test

Run	M. F1	Acc.
retuyt-svm-cross-cr-1	0.476	0.569
retuyt-svm-cross-cr-2	0.474	0.542
retuyt-lstm-cross-cr-1	0.473	0.530
retuyt-cnn-cross-cr-2	0.462	0.551
ingeotec-run2	0.454	0.538
retuyt-lstm-cross-cr-2	0.444	0.468
retuyt-cnn-cross-cr-1	0.421	0.423
itainnova-cl-base-cross-CR	0.409	0.440
ingeotec-run1	0.384	0.398

Table 9: Task 1: InterTASS Cross-lingual with CR as test

aspect-related metadata: the main category of the aspect, and the polarity of the opinion about the aspect. Systems had to classify the opinion about the given aspect in 3 different polarity labels (POSITIVE, NEGATIVE, NEUTRAL).

Participants were expected to submit up to 3 experiments for each provided collection, each in a plain text file with tweet identification, aspect and polarity.

For evaluation, exact match with a single label combining "aspect-polarity" was used. Similarly to Task 1, the macro-averaged version of Precision, Recall, F1, and Accuracy were considered, and Macro-F1 was used for a final ranking of proposed systems.

2.2.1 Collections

The Social-TV corpus was collected during the 2014 Final of "Copa del Rey" championship in Spain. After filtering out useless information, a subset of 2,773 tweets was obtained. The details of the corpus are described in (Villena-Román et al., 2015; García-Cumbreras et al., 2016; Martínez-Cámara et al., 2017).

STOMPOL (corpus of Spanish Tweets for Opinion Mining at aspect level about POLitics) is a corpus for the task of Aspect Based Sentiment Analysis. The corpus is composed of 1,284 tweets manually annotated by two annotators, and a third one in case of disagreement. The details of the corpus are described in (Villena-Román et al., 2015; García-Cumbreras et al., 2016; Martínez-Cámara et al., 2017).

2.2.2 Results

Only the research group ELiRF (González, Hurtado, and Pla, 2018c) participated in this edition. They explored different approaches based on Deep Learning. Specifically, they studied the behaviour of the CNN, Attention Bidirectional Long Short Term Memory (Att-BLSTM) and Deep Averaging Networks (DAN), similar to the proposal of the team for Task 1. In order to study the performance of the different models, they carried out an adjustment process. Tables 10 and 11 show the results obtained in their experiments.

Run	M. F1	Acc.
ELiRF-UPV-run1	0.485	0.627
ELiRF-UPV-run3	0.483	0.628
ELiRF-UPV-run2	0.476	0.625

Table 10: Task 2 Social-TV corpus results

Run	M. F1	Acc.
ELiRF-UPV-run2	0.526	0.633
ELiRF-UPV-run1	0.490	0.613
ELiRF-UPV-run3	0.447	0.576

Table 11: Task 2 STOMPOL corpus results

2.3 Task 3

NLP methods are increasingly being used to mine knowledge from unstructured content of

health (Liu et al., 2013; Doing-Harris and Zeng-Treitler, 2011; Gonzalez-Hernandez et al., 2017) and other domains (Estevez-Velarde et al., 2018). Over the years, many eHealth challenges have taken place, such as SemEval[2], CLEF[3] campaigns and others (Augenstein et al., 2017). These tasks have mainly dealt with identification, classification, extraction and linking of knowledge. The Task 3: eHealth Knowledge Discovery (eHealth-KD) proposes modelling the human language in a scenario in which Spanish electronic health documents could be machine readable from a semantic point of view. This task is designed to encourage the development of software technologies to automatically extract a large variety of knowledge from eHealth documents written in the Spanish language.

In order to capture the semantics of a broad range of health related text, eHealth-KD proposes the identification of two types of elements: **Concepts** and **Actions**. Concepts are key phrases that represent actors or entities which are relevant in a domain, while Actions represent how these Concepts interact with each other. Actions and Concepts can be linked by two types of relations: **subject** and **target**, which describe the main roles that a **Concept** can perform. Also, four specific semantic relations between Concepts are defined: **is-a**, **part-of**, **property-of** and **same-as**. Figure 1 provides an example.

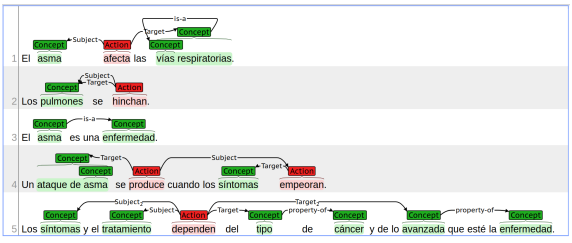

Figure 1: Example annotation of a small set of documents.

To simplify and normalise the extraction process, the overall task is divided into three subtasks:

- Subtask A is concerned with the extraction of the relevant key phrases.
- Subtask B is concerned with the classification of the key phrases identified in Subtask A as either **Concept** or **Action**.
- Subtask C is concerned with the discovery of the semantic relations between pairs of entities.

To compute the evaluation metrics for each subtask, we define the following sets for comparing the annotations between both the expected output (gold standard) and the actual output in each subtask:

Correct matches (C): in all subtasks, when one gold and one given annotation exactly match.

Partial matches (P): in subtask A, when two key phrases have a non-empty intersection.

Missing matches (M): in subtasks A and C, when an annotation in the gold output is not provided by the system.

Spurious matches (S): in subtasks A and C, when an annotation given by the system does not appear in the gold output.

Incorrect matches (I): in subtask B, when one assigned label is incorrect.

To measure the individual subtasks results as well as overall results, the eHealth-KD challenge proposes three evaluation scenarios.

Scenario 1. The first scenario requires all subtasks (i.e. A, B and C) to be performed sequentially. The input in this scenario consists of plain text (100 sentences), and participants must submit the three output files corresponding to subtasks A, B and C. In this scenario the overall quality of the participant systems is evaluated. So, a combined micro F_1 metric was defined, taking into account results of the three tasks:

$$F_{1ABC} = \frac{2 \cdot P_{ABC} \cdot R_{ABC}}{P_{ABC} + R_{ABC}} \quad (1)$$

$$P_{ABC} = \frac{T_{ABC} + \frac{1}{2}P_A}{T_{ABC} + P_A + M_A + I_B + M_C} \quad (2)$$

$$R_{ABC} = \frac{T_{ABC} + \frac{1}{2}P_A}{T_{ABC} + P_A + S_A + I_B + S_C} \quad (3)$$

$$T_{ABC} = C_A + C_B + C_C \quad (4)$$

[2]International Workshop on Semantic Evaluation
[3]Conference and Labs of the Evaluation Forum

Scenario 2. In the second scenario only subtasks B and C are performed. Hence, participants receive plain text inputs and the corresponding outputs for subtask A (a different subset of 100 sentences). This scenario allows participants to focus on the key phrases classification, without being affected by errors related to the extraction of key phrases. Like Scenario 1, a combined micro F_1 is defined which takes into account the results for subtasks B and C:

$$F_{1BC} = \frac{2 \cdot P_{BC} \cdot R_{BC}}{P_{BC} + R_{BC}} \quad (5)$$

$$P_{BC} = \frac{T_{BC}}{T_{BC} + I_B + M_C} \quad (6)$$

$$R_{BC} = \frac{T_{BC}}{T_{BC} + I_B + S_C} \quad (7)$$

$$T_{BC} = C_B + C_C \quad (8)$$

Scenario 3. Finally, the third scenario evaluates only subtask C. Participants are provided with plain text inputs and the corresponding output of subtasks A and B (a final subset of another 100 sentences). In this scenario, the following metric is defined for evaluation:

$$F_{1C} = 2 \cdot \frac{P_C \cdot Rec_C}{P_C + R_C} \quad (9)$$

$$P_C = \frac{C_C}{C_C + S_C} \quad (10)$$

$$R_C = \frac{C_C}{C_C + M_C} \quad (11)$$

For competition purposes, the best system is defined as the submission that maximises the macro-average F_1 across all three scenarios:

$$F_1 = \frac{F_{1ABC} + F_{1BC} + F_{1C}}{3} \quad (12)$$

2.3.1 Corpora

For evaluation purposes, a corpus of health-related sentences in Spanish was manually built and tagged. The corpus consists of a selection of articles collected from the MedlinePlus[4] website. These files contain several entries related to health and medicine topics, and environmental topics strongly related to health care. Spanish language items were converted to a plain text document, processed, and manually tagged using the Brat annotation tool[5] by 15 human annotators divided into seven groups. The final 1,173 tagged sentences were organised in three collections: *training*, *development* and *test*. Table 12 summarises the main statistics of the corpus.

	Train	**Dev.**	**Test**
Files	6	1	3
Sentences	559	285	300
Annotations	5976	3573	3310
Entities	3280	1958	1805
- Concepts	2431	1524	1305
- Actions	849	434	500
Roles	1684	843	988
- subject	693	339	401
- target	991	504	587
Relations	1012	772	517
- is-a	434	370	235
- part-of	149	145	96
- property-of	399	244	178
- same-as	30	13	8

Table 12: Statistics of the eHealth-KD v1.0 corpus.

2.3.2 Analysis of the Results

eHealth-KD challenge attracted the attention of a total 31 registered teams of which six of then successfully concluded their participation. Their results are summarised in Table 13. The following tag labels are designed to provide an overview of the main characteristics of each participant system:

S: Uses shallow supervised models such as CRF, logistic regression, SVM, decision trees, etc.

D: Uses deep learning models, such as LSTM or convolutional networks.

E: Uses word embeddings or other embedding models trained with external corpora.

K: Uses external knowledge bases, either explicitly or implicitly (i.e, through third-party tools).

R: Uses hand crafted rules based on domain expertise.

N: Uses natural language processing techniques or features, i.e., POS-tagging, dependency parsing, etc.

[4] https://medlineplus.gov/xml.html

[5] http://brat.nlplab.org/

Baseline description: A baseline, trained on the training corpus, was defined. This strategy consists of a dummy approach based solely on the text of key phrases. This technique collects all training data and stores three maps: (1) key phrases associated with their most common class (either Concept or Action); (2) pairs of concepts associated with their most common relation; and (3) tuples of <Action,Concept> associated with their most common role. At prediction time, these maps are used to select a key phrase, decide its class, and predict relations and roles.

Once the shared subtask ended, the official results were published. However, some participants noticed that their systems provided duplicated outputs on some occasions. These duplicated outputs, even if correct, were being counted as spurious after the first match. To account for this duplication, the evaluation script was modified to remove duplicated outputs from the participants submissions prior to calculating the evaluation metrics. Table 14 shows this second version of the metrics, where variations in scores are highlighted in bold text. This proved not to be a significant problem, since only two participants were affected, and even though their metrics improved marginally, the overall results or the main conclusions of the shared subtask did not change.

The results of this task, eHealth-KD, show that a variety of approaches, on the whole, deal effectively with the health knowledge discovery problem. However, issues still need to be resolved to obtain highly competitive systems. The best performing submissions include classic supervised learning, deep learning and knowledge-based techniques. In subtask A, the best approach (`UC3M`) is based on a CRF model with pre-trained embeddings as features. This approach obtains similar scores in subtask B. In general, subtask B appears to be easier than the rest, which is understandable given that there are only two classes and there is a large correlation between word lemmas and their classes (as shown by the relatively high performance of the baseline).

Subtask C, in concordance with Scenario 3, does not exceed 45% in F-score. This reinforces the belief that this task is difficult to deal with, even after having applied novel approaches (i.e. `TALP` and `LaBDA`) based on convolutional neural networks.

The best-performing systems in each scenario highly coincide with all three task results. For Scenario 1, the top performing strategy belongs to `UC3M`, which achieves the best scores in subtasks A and B, and the overall best result in the shared subtask (averaged across all three scenarios), pretty close to SINAI. Likewise, the best strategy in Scenario 3 corresponds to `TALP`, which achieves the best score for subtask C. However, for the overall results, other participants such as `SINAI` and `UPF-UPC` achieve higher average scores, even though their performance in subtask C and Scenario 3 is practically negligible. In contrast, these teams obtain relatively high scores in subtasks A and B.

The diverse nature and complexity of the three subtasks make it difficult to design a single fair evaluation metric. For this reason, we consider that each system submission gets more accurate results related to the specific sub-problems that it tackles. Although generalisation across the three subtasks is a desirable characteristic, advances in any particular subtask are also very valuable.

In general, the most competitive approaches in individual subtasks are dominated by state-of-the-art machine learning. In the particular case of subtask C, where modern deep learning approaches seem to outperform classic techniques. In addition, incorporating domain-specific knowledge provides a significant boost to the performance. Most participants use NLP features, either explicitly, or implicitly captured in word embeddings and other representations. An interesting phenomenon is that the best systems in subtask A do not correlate with the best systems in subtask C. This suggests that the optimal approach for either subtask is different, giving rise to an interesting research line that would explore integrated approaches to simultaneously solving these three subtasks. The overall results show that general purpose knowledge discovery in domain-specific documents is potentially a prolific research area, particularly for the Spanish language. We expect similar future initiatives to provide fruitful evaluation scenarios where researchers can deploy techniques from several domains, and compete in friendly contests to improve the state-of-the-

[7] This extracts lexical and syntactic features for each token. Afterwards, it applies a set of handcrafted heuristics for each subtask.

	UC3M†	SINAI†	UPF-UPC†	TALP†	LaBDA†	UH	Baseline
Tags	**SDEN**	**KRN**	**SKN**	**DEN**	**D**	**RN**	
Subtask A	**0.872**	0.798	0.805	-	0.323	0.172	0.597
Subtask B	**0.959**	0.921	0.954	0.931	0.594	0.639	0.774
Subtask C	-	-	0.036	**0.448**	0.420	0.018	0.107
Average	**0.610**	0.573	0.598	0.460	0.446	0.276	0.493
Scenario 1	**0.744**	0.710	0.681	-	0.297	0.181	0.566
Scenario 2	0.648	0.674	0.622	**0.722**	0.275	0.255	0.577
Scenario 3	-	-	0.036	**0.448**	0.420	0.018	0.107
Average	**0.464**	0.461	0.446	0.390	0.331	0.151	0.417

Table 13: Summary of systems and results for the TASS 2018 Task 3 event. The best scores are in bold text. More details in UC3M (Zavala, Martínez, and Segura-Bedmar, 2018), SINAI (López-Ubeda et al., 2018), UPF-UPC (Palatresi and Hontoria, 2018), TALP (Medina and Turmo, 2018), LaBDA (Suarez-Paniagua, Segura-Bedmar, and Martínez, 2018) and UH[7]. The symbol † means that the group submitted a system description paper.

	UPF-UPC	LaBDA
Tags	**SKN**	**DN**
Subtask A	0.805	0.323
Subtask B	0.954	0.594
Subtask C	0.036	**0.444**
Average	0.598	**0.454**
Scenario 1	0.681	**0.310**
Scenario 2	**0.626**	**0.294**
Scenario 3	0.036	**0.444**
Average	**0.448**	**0.349**

Table 14: Summary of results of submissions that changed once duplicated entries were removed. Variations in score are highlighted in bold text.

art.

2.4 Task 4

When news are about natural disasters, readers usually feel negative emotions (sadness, for instance), whereas when those news are about the last championship won by your favourite team, readers feel positive emotions like happiness. Moreover, it is commonly assumed in marketing that emotions aroused in the reader by news articles have an impact in the perception of the advertisements displayed along with those articles. Thus, from that marketing perspective, if a company wants to promote their brand, the ads should better be associated to (i.e., shown with) news that arouse positive emotions.

The objective of Task-4 is to encourage the development of systems that can classify a news article into SAFE (positive emotions, so safe for ads) or UNSAFE (negative emotions, so better avoid ads). This task could be considered as a kind of stance classification, on the positioning of readers of news contents. The task is a strong challenge because it has to deal with the polarity of feeling (safe vs. unsafe) and to work in combination with a (pseudo) thematic classification to be able to determine the meaning of the news. For example, the reduction of traffic accidents has a negative feeling because of the accidents, but the context of reducing the numbers of accidents makes those bad news good news, hence SAFE news.

2.4.1 Corpora

The Spanish brANd Safe Emotion corpus (SANSE) corpus was specifically built for this task. RSS feeds of different online newspapers written in different varieties of Spanish (Argentina, Chile, Colombia, Cuba, Spain, USA, Mexico, Peru and Venezuela) were collected for over a month. Finally 15,152 articles were captured, containing the URL, the publication date and the headline. News summaries were also collected for several sources, but finally they were discarded to make the dataset consistent and homogeneous.

Then 2,000 articles (L1 subset) were randomly selected and were manually annotated into an emotional categorisation of SAFE or UNSAFE, from the point of view of the general public of each corresponding country. The other 13,152 articles (L2 subset) were not annotated.

As the datasets were annotated with two levels of safety: SAFE and UNSAFE, the task can be considered as a binary classification task.

The annotation was carried out by two human annotators (the two organisers of the task), and, for those cases with no agreement between the two annotators, a third annotator undid the tie. A safe headline of a news was defined as an utterance that arises a positive or neutral emotion in the reader and is not related to a controversial topics: religion, extreme wing political topics, or controversial topics (those that arise strong positive emotions to some readers but strong negative emotions to other ones). An unsafe headline was defined as an utterance that arises negative emotions on the reader.

Some examples in Spanish:

Así será el nuevo pan integral en España, según una nueva ley en marcha. → SAFE

This will be the new integral bread in Spain, according to a new law underway.

Casi 300 municipios de Colombia en riesgo electoral. → UNSAFE

Almost 300 municipalities in Colombia at electoral risk.

The agreement of the annotation was 0.58 according to π (Scott, 1955) and κ (Cohen, 1960), which may consider moderate according to Landis and Koch (1977). Although the agreement is moderate, it is close to be considered substantial, and we have also to take into account that it is a new classification task that works with a strong subjective content. We will work in making the annotations guidelines more precise in order to improve the agreement of the annotators. Besides, we hope that the participants will give us insights with the aim of improving the annotation of the data.

The L1 subset was then again divided in three subsets, specifically: *training*, *development* and *test*. The statistics of the three subsets are in Table 15.

2.4.2 Tasks

Two subtasks were proposed. Subtask 1 (S1) consists of the classification of headlines into SAFE or UNSAFE for incorporating an ad of a brand. The evaluation of the systems does not take into account the cultural varieties of the Spanish language, it thus a monolingual evaluation. In this task, datasets are composed of headlines of news written in different version of the Spanish language, but the country of the text is not relevant for this task.

Subset	Size
Training	1250
Development	250
Test	500

Table 15: Statistics of the SANSE corpus

Subset	Size
Training (Spain)	300
Dev. (Spain)	48
Test (Mexico)	144
Test (Cuba)	194
Test (Chile)	194
Test (Colombia)	195
Test (Argentina)	198
Test (Venezuela)	233
Test (Peru)	234
Test (USA)	260

Table 16: Caption

Participants were provided with the *training* and *development* subsets of L1 SANSE corpus for building the systems, and two *test* sets for the evaluation: the test subset of L1 SANSE corpus and the L2 SANSE corpus.

The systems presented were evaluated using the measures of Macro-Precision (M. P.), Macro-Recall (M. R.), Macro-F1 (M. F1) and Accuracy (Acc.).

Subtask 2 (S2) is similar to S1, but in this case the aim is to evaluate the generalisation capacity of the submitted systems. For training their systems, participants were provided with SANSE subsets with headlines written only in the Spanish language spoken in Spain. The *test* set was composed of headlines written in the Spanish language spoken in different countries of America. The statistics of SENSE corpus for S2 are shown in the Table 16.

2.4.3 Results

Task 4 attracted the attention of seven teams, and most of them participated in both levels of evaluation of the S1 and in S2. Table 2.4.3 shows the participation of the teams in each Subtask. Five groups of the seven ones

submitted a system description paper, whose main features will be detailed as what follows.

INGEOTEC. Moctezuma et al. (2018) propose an ensemble classification system (EvoMSA), which is composed of several and heterogeneous base systems and a genetic programming system (EvoDAG, (Graff et al., 2017)) that optimises the contribution of each base system in the final classification. The authors combined supervised and unsupervised system as base classification systems. The supervised ones are based on the use of the algorithm SVM with different representations of the input text, namely TF-IDF and pre-trained word vectors. The system reached the best results in the monolingual and the multilingual evaluations, however the performance of the system dropped a bit in S1 L2. Since the annotation *test* set of S1 L2 was conducted by a voting system of the all the submitted systems, the lower performance in S1 L2 may be caused by a different error distribution between INGEOTEC and the systems submitted by the other groups.

ELiRF_UPV. González, Hurtado, and Pla (2018a) propose a deep neural network, specifically the model Deep Averaging Networks (DAN) (Iyyer et al., 2015). The authors used a set of pre-trained word embeddings for representing the news headlines. The set of pre-trained word embeddings were prepared by the authors and built upon a corpus of tweets (Hurtado, Pla, and González, 2017). The high performance reached by a set of pre-trained word embeddings built upon tweets with news headlines stands out, because the genre of news headlines and tweets are different. However, it may mean that the use of language in tweets and news headlines is similar.

Team	S1_L1	S1_L2	S2
INGEOTEC†	✓	✓	✓
ELiRF-UPV†	✓	✓	✓
rbnUGR†	✓	✓	✓
MeaningCloud†	✓	✓	✓
SINAI†	✓	✓	-
lone_wolf	✓	✓	-
TNT-UA-WFU	✓	✓	-

Table 17: Participation of each team on each Subtask. The symbol † means that the group submitted a system description paper

rbnUGR. Rodríguez Barroso, Martínez-Cámara, and Herrera (2018) submitted three systems grounded in deep learning. Although the three systems are based on Long Short-Term Memory (LSTM) Recurrent Neural Network (RNN), they have several differences:

Run_1. It uses a LSTM layer as encoding layer, and its output is the last vector state of the LSTM layer.

Run_2. It uses a BiLSTM[8] layer as encoding layer, and its output is the concatenation of the last vector state of the two LSTM layers.

Run_3. It uses a LSTM layer as encoding layer, and its output is the concatenation of the corresponding output state vector of each input token.

The results show that the systems based on one single LSTM layer perform better than the one based on BiLSTM. Regarding the different results in S1 and S2 indicate that the use the entire output of the encoding layer allow to improve the generalisation capacity of the model, because the multilingual evaluation requires a higher generalisation capacity.

MeaningCloud. Herrera-Planells and Villena-Román (2018) propose three supervised systems, two of them are lineal classification systems and the other one a non-lineal classification system. The linear classification systems use XGBoost (Chen and Guestrin, 2016) as classification system. They differ in the set of features used to represent the news headlines, which are mainly built using the public APIs of the text analytic platform of MeaningCloud. The non lineal classification system is a neural network based on the use of a CNN layer. The proposal that reached higher results was the one grounded in a CNN (Run_3).

SINAI. Plaza del Arco et al. (2018) propose to represent the news headlines as a vector of unigrams weighted with TF-IDF, and the number of positive and negative words according to three list of opinion bearing words. The authors used SVM as classification algorithm.

The evaluation measures in the two Subtasks were accuracy and the macro-average

[8] A BiLSTM is an elaboration of two LSTM layers.

System	S1 L1				S1 L2				S2			
	M. P.	M. R.	M. F1	Acc.	M. P.	M. R.	M. F1	Acc.	M. P.	M. R.	M. F1	Acc.
INGEOTEC_run1	0.794	0.795	0.795[1]	0.802	0.853	0.880	0.866[4]	0.871	0.722	0.715	0.719[1]	0.737
ELiRF_UPV_run2	0.787	0.794	0.790[2]	0.794	0.850	0.884	0.867[3]	0.865	0.747	0.657	0.699[2]	0.722
ELiRF_UPV_run1	0.795	0.784	0.790[3]	0.800	0.878	0.889	0.883[1]	0.893	0.736	0.649	0.690[3]	0.715
rbnUGR_run1	0.784	0.764	0.774[4]	0.786	0.880	0.867	0.873[2]	0.888	0.683	0.661	0.672[6]	0.700
MEANING-CLOUD_run3	0.767	0.767	0.767[5]	0.776	0.781	0.804	0.793[7]	0.801	0.647	0.654	0.651[7]	0.658
rbnUGR_run3	0.763	0.765	0.764[6]	0.772	0.838	0.870	0.853[6]	0.853	0.687	0.678	0.683[4]	0.631
rbnUGR_run2	0.774	0.752	0.763[7]	0.776	0.868	0.857	0.863[5]	0.878	0.679	0.672	0.676[5]	0.698
SINAI	0.733	0.722	0.728[8]	0.742	0.769	0.777	0.773[8]	0.793	-	-	-	-
MEANING-CLOUD_run2	0.723	0.727	0.725[9]	0.732	-	-	-	-	-	-	-	-
MEANING-CLOUD_run1	0.713	0.722	0.717[10]	0.714	-	-	-	-	-	-	-	-

Table 18: Macro precision (M. P.), macro recall (M. R.), macro f1 (M. F1) and accuracy (Acc.) reached by each submitted system to each Subtask of the groups that submitted a system description paper

of precision, recall and F1, and the systems were ranked according to the value of macro-F1. Table 18 show the results reached by each group that submitted the description of their systems in S1_L1, S1_L2 and S2 respectively.

3 Conclusions

The edition of TASS 2018 has attracted the participation of 16 systems, and the submission of 15 system description papers. Moreover, we have proposed two new challenges to the international reserach community, which are in line to the requirements of the Industry.

The submitted systems are in the line of the state-of-the-art in other similar workshops, and most of them are grounded in Deep Learning and the use of hand-crafted linguistic features. Therefore, TASS may be considered as a reference forum for setting up the state-of-the-art in semantic analysis in Spanish.

As future work, we plan to enlarge the coverage of the Spanish language of the corpus InterTASS, as well as consolidating the new challenges (Task 3 and Task 4). Moreover, we will keep working in the development of new corpora and linguistic resources for the research community.

Acknowledgments

This work has been partially supported by a grant from the Fondo Europeo de Desarrollo Regional (FEDER), the projects REDES (TIN2015-65136-C2-1-R, TIN2015-65136-C2-2-R) and SMART-DASCI (TIN2017-89517-P) from the Spanish Government, and "Plataforma Inteligente para Recuperación, Análisis y Representación de la Información Generada por Usuarios en Internet" (GRE16-01) from University of Alicante. Eugenio Martínez Cámara was supported by the Spanish Government Programme Juan de la Cierva Formación (FJCI-2016-28353).

References

Augenstein, I., M. Das, S. Riedel, L. Vikraman, and A. McCallum. 2017. Semeval 2017 task 10: Scienceie-extracting keyphrases and relations from scientific publications. *arXiv preprint arXiv:1704.02853*.

Chen, T. and C. Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. ACM.

Chiruzzo, L. and A. Rosá. 2018. Retuytinco at tass 2018: Sentiment analysis in spanish variants using neural networks and svm. 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, and J. Villena-Román, editors, *Proceedings of TASS 2018: Workshop on Semantic Analysis at SEPLN (TASS 2018)*, volume 2172 of *CEUR*

Workshop Proceedings, Sevilla, Spain, September. CEUR-WS.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Denecke, K. 2015. *Health Web Science*. Springer International Publishing.

Doing-Harris, K. M. and Q. Zeng-Treitler. 2011. Computer-assisted update of a consumer health vocabulary through mining of social network data. *Journal of medical Internet research*, 13(2).

Estevez-Velarde, S., Y. Gutiérrez, A. Montoyo, A. Piad-Morffis, R. M. noz, and Y. Almeida-Cruz. 2018. Gathering object interactions as semantic knowledge. In *Proceedings of the 2018 International Conference on Artificial Intelligence (ICAI'18)*.

García-Cumbreras, M. A., J. Villena-Román, E. Martínez-Cámara, M. C. Díaz-Galiano, M. T. Martín-Valdivia, and L. A. Ureña López. 2016. Overview of tass 2016. In *TASS 2016: Workshop on Sentiment Analysis at SEPLN*, pages 13–21.

González, J.-A., L.-F. Hurtado, and F. Pla. 2018a. ELiRF-UPV en TASS 2018: Categorización emocional de noticias. In *Proceedings of TASS 2018: Workshop on Semantic Analysis at SEPLN (TASS 2018)*, volume 2172, September.

González, J.-A., L.-F. Hurtado, and F. Pla. 2018b. Elirf-upv en tass 2018: Análisis de sentimientos en twitter basado en aprendizaje profundo. 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, and J. Villena-Román, editors, *Proceedings of TASS 2018: Workshop on Semantic Analysis at SEPLN (TASS 2018)*, volume 2172 of *CEUR Workshop Proceedings*, Sevilla, Spain, September. CEUR-WS.

González, J.-A., L.-F. Hurtado, and F. Pla. 2018c. Elirf-upv en tass 2018: Análisis de sentimientos en twitter basado en aprendizaje profundo. In *Proceedings of TASS 2018: Workshop on Semantic Analysis at SEPLN (TASS 2018)*.

Gonzalez-Hernandez, G., A. Sarker, K. O'Connor, and G. Savova. 2017. Capturing the patient's perspective: a review of advances in natural language processing of health-related text. *Yearbook of medical informatics*, 26(01):214–227.

Graff, M., E. S. Tellez, H. Jair Escalante, and S. Miranda-Jiménez, 2017. *Semantic Genetic Programming for Sentiment Analysis*, pages 43–65. Springer International Publishing, Cham.

Herrera-Planells, J. and J. Villena-Román. 2018. MeaningCloud at TASS 2018: News headlines categorization for brand safety assessment. In *Proceedings of TASS 2018: Workshop on Semantic Analysis at SEPLN (TASS 2018)*, volume 2172, September.

Hurtado, L.-F., F. Pla, and J.-A. González. 2017. Elirf-upv en tass 2017: Análisis de sentimientos en twitter basado en aprendizaje profundo. In *Proceedings of TASS 2017: Workshop on Sentiment Analysis at SEPLN co-located with 33nd SEPLN Conference (SEPLN 2017)*.

Iyyer, M., V. Manjunatha, J. Boyd-Graber, and H. Daumé III. 2015. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1681–1691. Association for Computational Linguistics.

Landis, J. R. and G. G. Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Liu, H., S. J. Bielinski, S. Sohn, S. Murphy, K. B. Wagholikar, S. R. Jonnalagadda, K. Ravikumar, S. T. Wu, I. J. Kullo, and C. G. Chute. 2013. An information extraction framework for cohort identification using electronic health records. *AMIA Summits on Translational Science Proceedings*, 2013:149.

López-Ubeda, P., M. C. Díaz-Galiano, M. T. Martín-Valdivia, and L. A. Urena-Lopez. 2018. Sinai en tass 2018 task 3. clasificando acciones y conceptos con umls en medline. In *Proceedings of TASS 2018:*

Workshop on Semantic Analysis at SEPLN (TASS 2018).

Luque, F. M. and J. M. Pérez. 2018. Atalaya at tass 2018: Sentiment analysis with tweet embeddings and data augmentation. 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, and J. Villena-Román, editors, *Proceedings of TASS 2018: Workshop on Semantic Analysis at SEPLN (TASS 2018)*, volume 2172 of *CEUR Workshop Proceedings*, Sevilla, Spain, September. CEUR-WS.

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

Medina, S. and J. Turmo. 2018. Joint classification of key-phrases and relations in electronic health documents. In *Proceedings of TASS 2018: Workshop on Semantic Analysis at SEPLN (TASS 2018)*.

Moctezuma, D., J. Ortiz-Bejar, E. S. Tellez, S. Miranda-Jiménez, and M. Graff. 2018. Ingeotec solution for task 4 in tass'18 competition. In *Proceedings of TASS 2018: Workshop on Semantic Analysis at SEPLN (TASS 2018)*, volume 2172, September.

Moctezuma1, D., J. Ortiz-Bejar, E. S. Téllez, S. Miranda-Jiménez, and M. Graff. 2018. Ingeotec solution for task 1 in tass'18 competition. 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, and J. Villena-Román, editors, *Proceedings of TASS 2018: Workshop on Semantic Analysis at SEPLN (TASS 2018)*, volume 2172 of *CEUR Workshop Proceedings*, Sevilla, Spain, September. CEUR-WS.

Montanés, R., R. Aznar, and R. del Hoyo. 2018. Aplicación de un modelo híbrido de aprendizaje profundo para el análisis de sentimiento en twitter. 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, and J. Villena-Román, editors, *Proceedings of TASS 2018: Workshop on Semantic Analysis at SEPLN (TASS 2018)*, volume 2172 of *CEUR Workshop Proceedings*, Sevilla, Spain, September. CEUR-WS.

Palatresi, J. V. and H. R. Hontoria. 2018. Tass2018: Medical knowledge discovery by combining terminology extraction techniques with machine learning classification. In *Proceedings of TASS 2018: Workshop on Semantic Analysis at SEPLN (TASS 2018)*.

Plaza del Arco, F. M., E. Martínez-Cámara, M. T. Martín Valdivia, and A. Ureña López. 2018. SINAI en TASS 2018: Inserción de conocimiento emocional externo a un clasificador lineal de emociones. In *Proceedings of TASS 2018: Workshop on Semantic Analysis at SEPLN (TASS 2018)*, volume 2172, September.

Rodríguez Barroso, N., E. Martínez-Cámara, and F. Herrera. 2018. SCI^{2}S at TASS 2018: Emotion classification with recurrent neural networks. In *Proceedings of TASS 2018: Workshop on Semantic Analysis at SEPLN (TASS 2018)*, volume 2172, September.

Scott, W. A. 1955. Reliability of content analysis: The case of nominal scale coding. *Public opinion quarterly*, pages 321–325.

Suarez-Paniagua, V., I. Segura-Bedmar, and P. Martínez. 2018. Labda at tass-2018 task 3: Convolutional neural networks for relation classification in spanish ehealth documents. In *Proceedings of TASS 2018: Workshop on Semantic Analysis at SEPLN (TASS 2018)*.

Villena-Román, J., J. García-Morera, M. A. García-Cumbreras, E. Martínez-Cámara,

M. T. Martín-Valdivia, and L. A. Ureña López. 2015. Overview of tass 2015. In *TASS 2015: Workshop on Sentiment Analysis at SEPLN*, pages 13–21.

Zavala, R. M. R., P. Martínez, and I. Segura-Bedmar. 2018. A hybrid bi-lstm-crf model for knowledge recognition from ehealth documents. In *Proceedings of TASS 2018: Workshop on Semantic Analysis at SEPLN (TASS 2018)*.