Procesamiento del Lenguaje Natural, Revista nº 62, marzo de 2019, pp. 77-84 recibido 22-11-2018 revisado 09-01-2019 aceptado 11-02-2019

TASS 2018: The Strength of Deep Learning in Language Understanding Tasks

TASS 2018: La Potencia del Aprendizaje Profundo en Tareas de Comprensión del Lenguaje

Manuel Carlos Díaz-Galiano,[1] Miguel Á. García-Cumbreras,[1] Manuel García-Vega,[1] Yoan Gutiérrez,[2] Eugenio Martínez-Cámara,[3] Alejandro Piad-Morffis,[4] Julio Villena-Román[5]
[1]Centro de Estudios Avanzados en Tecnologías de la Información y de la Comunicación (CEATIC). Universidad de Jaén, España
[2]University of Alicante, España
[3]Andalusian Research Institute in Data Science and Computational Intelligence (DaSCI). Universidad de Granada, España
[4]University of Havana, Cuba. [5]MeaningCloud, España.
[1]{mcdiaz,magc,mgarcia}@ujaen.es, [2]ygutierrez@dlsi.ua.es
[3]emcamara@decsai.ugr.es, [4]apiad@matcom.uh.cu
[5]jvillena@meaningcloud.com

Abstract: The edition of TASS in 2018 was the edition of the evolution of TASS to a competitive evaluation workshop on semantic and text understanding tasks. Consequently, TASS has enlarged the coverage of tasks, and it goes beyond sentiment analysis. Thereby, two new tasks focused on semantic relation extraction in the health domain and emotion classification in the news domain were added to the two traditional tasks of TASS, namely sentiment analysis at tweet level and aspect level. Several systems were submitted, and most of them are based on state of the art classification methods, which highlight those ones grounded in Deep Learning. As addition contribution, TASS 2018 released two new corpora, specifically the ones of the two new tasks.
Keywords: Sentiment analysis, emotion classification, digital health

Resumen: La edición de 2018 de TASS ha sido la de la evolución de TASS hacia un taller de evaluación competitiva sobre tareas de análisis semántico y de entendimiento del lenguaje, ampliando así su cobertura de tareas más allá del análisis de opiniones. De este modo, a las dos tareas clásicas de clasificación de la polaridad a nivel de tuit y a nivel de aspecto, se ha añadido una tarea de extracción de relaciones semánticas en el dominio médico, y otra de clasificación de emociones en el dominio periodístico. Son numerosos los sistemas que se evaluaron en TASS 2018, y hay que destacar que la mayoría de ellos están a la vanguardia en el uso de técnicas de clasificación, destacando los sistemas basados en Aprendizaje Profundo. Como contribución adicional, TASS 2018 ha publicado dos nuevos *corpora*, correspondientes a las dos nuevas tareas.
Palabras clave: Análisis de opiniones, análisis de emociones, salud digital

1 Introduction

The edition of TASS 2018 was the edition of the evolution of TASS into a competitive evaluation workshop on semantic analysis and text understanding tasks. Accordingly, TASS has enlarged the coverage of Natural Language Processing (NLP) tasks, and it goes beyond Sentiment Analysis.

In this paper, we describe the edition of TASS 2018, in which four tasks were organized. Two of them were the usual tasks of TASS, namely sentiment analysis at tweet level (Task 1) and sentiment analysis at aspect level (Task 2). Two new tasks were additionally organized, the first one focused on semantic relation extraction in health data (Task 3), and the second one emotion classification in the the news domain (Task 4).

Eighteen research teams participated in the four tasks of TASS 2018, and they sub-

ISSN 1135-5948. DOI 10.26342/2019-62-9

	P	N	NEU	NONE	Total
ES	1,115	1,402	418	474	3,409
PE	756	820	594	758	2,928
CR	677	912	297	447	2,333

Table 1: InterTASS 2.0: tweets subsets

mitted state-of-the-art systems, which is a relevant contribution for the research community of NLP in Spanish. As additional contribution, two new corpora were released, namely the two ones of the two new tasks.

The paper is organized as follows: Section 2 describes the corpora provided in TASS 2018. The four organized tasks are detailed in Section 3. Section 4 exposes the main conclusions of TASS 2018 and the future work related to TASS.

2 Resources

TASS 2018 provided five datasets to the participants for the evaluation of their systems. Only two of them were already used in previous editions.

2.1 InterTASS 2.0 Corpora

In 2018, a new version of InterTASS Corpus arised, with new subsets of training and test data. *InterTASS 2.0* is composed of three subsets, with tweets written in different varieties of Spanish (for Spain, Peru and Costa Rica). It exhibits a large amount of lexical and even structural differences in each variant, and tweets were annotated with 4 different polarity labels POSITIVE, NEGATIVE, NEUTRAL and NONE. Table 1 shows the tweets distribution for each subset (Spain:ES, Peru:PE and Costa Rica:CR).

All subsets are balanced, although they have more positive and negative tweets.

2.2 Social-TV and STOMPOL Corpora

The Social-TV and STOMPOL corpora were released in previous TASS editions (Martínez-Cámara et al., 2017). They have tagged at aspect level, with 3 levels of opinion: positive (P), neutral (NEU) and negative (N). Table 2 shows the tweets distribution for these data in training and test sets.

	training	Test	Total
Social-TV	1,773	1,000	2,773
STOMPOL	1,284	784	2,068

Table 2: Social-TV and STOMPOL Corpora

2.3 eHealth-KD Corpora

For evaluation purposes of the eHealth-KD challenge, 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[1] website. MedlinePlus is the United States National Institutes of Health's website. This platform freely provides large health textual data from which a selection was made for constituting the eHealth-KD corpus in Spanish language. Table 3 shows the distribution of this corpus.

Metric	Total	Trial	Training	Dev.	Test
Files	11	1	6	1	3
Sentences	1173	29	559	285	300
Annotations	13113	254	5976	3573	3310
Entities	7188	145	3280	1958	1805
- Concepts	5366	106	2431	1524	1305
- Actions	1822	39	849	434	500
Roles	3586	71	1684	843	988
- subject	1466	33	693	339	401
- target	2120	38	991	504	587
Relations	2339	38	1012	772	517
- is-a	1057	18	434	370	235
- part-of	393	3	149	145	96
- property-of	836	15	399	244	178
- same-as	53	2	30	13	8

Table 3: Size of the eHealth-KD v1.0 corpus

2.3.1 SANSE corpus

The Spanish brANd Safe Emotion (SANSE) corpus comprises 15,152 news headlines from newspapers of some Spanish speaking countries: Spain, Argentina, Chile, Colombia, Cuba, USA, Mexico, Peru and Venezuela. The aim was to build a representative corpus of the use of Spanish in headlines.

The corpus was randomly splitted into two sets: the L1 subset with 2,000 headlines, and the L2 subset with the rest 13,152 headlines.

L1 was manually annotated by two annotators, and a third annotator undid the tie in those cases with no agreement. A SAFE headline was defined as an utterance that arises a positive or neutral emotion and is not related to any controversial topic such as religion or

[1] `https://medlineplus.gov/xml.html`

extreme wing political news.[2] Otherwise utterances were considered as UNSAFE. The agreement of the annotation was 0.58 according to π and κ (Cohen, 1960), which may be considered as moderate according to Landis and Koch (1977), though close to be substantial. This is justified considering the strong subjective nature of emotions.

Finally the L1 set was splitted into three subsets: *training* (1,250 headlines), *development* (250) and *test* (500). The L2 subset was automatically annotated by a voting system built upon the outputs of the systems submitted for S1 subtask (see Section 3.4.1).

3 Tasks

TASS 2018 organized four tasks, the usual tasks 1 (see Section 3.1) and 2 (see Section 3.2) about sentiment analysis at tweet and aspect levels, and two new tasks. TASK 3 (see Section 3.3) is focused on the extraction of semantic relations on health data. TASK 4 (see Section 3.4) proposes the emotional classification of news headlines in order to identify their level of safety for publishing spot ads.

3.1 Task 1. Sentiment analysis at tweet level

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

The submitted systems had to classify short tweets written in an informal language, many of them with misspelling or emojis, even onomatopoeias. But this year, systems had to solve a new problem: Multilinguality. One of the corpus was expanded with tweets written in Spanish from Peru and Costa Rica.

This extended corpus was the International TASS Corpus (InterTASS), a corpus released in 2017 with text written in Spanish from Spain and its description can be found in (Martínez-Cámara et al., 2017), while the varieties from Peru and Costa Rica have been released this year and their descriptions are shown in (Martínez-Cámara et al., 2018).

Currently, 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.

[2]These topics may arise strong conflicting emotions in some readers.

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

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.

Four sub-tasks were proposed:

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 Spanish of *training* set had to be different from the evaluation one, in order to test the dependency of systems on a language.

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.

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, RETUYT-InCo (Chiruzzo and Rosá (2018)), ITAINNOVA (Montanés, Aznar, and del Hoyo (2018)), ELiRF-UPV. (González, Hurtado, and Pla (2018b)) and ATALAYA (Luque and Pérez (2018)).

The first three teams classified for each monolingual subtask are shown in Table 4.

For the cross-lingual runs, the participants selected an InterTASS dataset to train their systems and a different one to test the dependency of systems on a language. Table 4 shows the results of the first 3 teams classified in these cross-lingual subtasks.

To read a complete information about the systems, runs and results see (Martínez-Cámara et al., 2018).

3.2 Task 2. Aspect-based Sentiment Analysis

Aspect-based Sentiment Analysis challenge was focused on aspect-based polarity classification systems. The datasets to evaluate the approaches were similar to previous editions (Martínez-Cámara et al., 2017): Social-TV and STOMPOL.

This year only one group has participated, ELiRF (González, Hurtado, and Pla, 2018b)

Run	Task	M. F1	Acc.
elirfEsRun1	monoES	0.503	0.612
retuytLstmEs1	monoES	0.499	0.549
atalayaUbav3	monoES	0.476	0.544
retuytLstmCr2	monoCR	0.504	0.537
elirfCrRun2	monoCR	0.482	0.561
atalayaCrLr502	monoCR	0.475	0.582
retuytCnnPe1	monoPE	0.472	0.494
atalayaPeLr502	monoPE	0.462	0.451
ingeotecRun1	monoPE	0.439	0.447
retuytSvmEs2	multiES	0.471	0.555
ingeotecRun1	multiES	0.445	0.530
atalayaMlp300	multiES	0.441	0.485
ingeotecRun1	multiPE	0.447	0.506
retuytSvmPe2	multiPE	0.445	0.514
atalayaMlp300	multiPE	0.438	0.523
retuytSvmCr1	multiCR	0.476	0.569
ingeotecRun2	multiCR	0.454	0.5382
itainnovaClBase	multiCR	0.409	0.440

Table 4: Task 1: Best teams per task

that has submitted three experiments for each collection. 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. Table 5 show the results obtained in their experiments.

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.

Run	Corpus	M. F1	Acc.
ELiRF-run1	Social-TV	0.485	0.627
ELiRF-run3	Social-TV	0.483	0.628
ELiRF-run2	Social-TV	0.476	0.625
ELiRF-run2	STOMPOL	0.526	0.633
ELiRF-run1	STOMPOL	0.490	0.613
ELiRF-run3	STOMPOL	0.447	0.576

Table 5: Macro f1 (M. F1) and accuracy (Acc.) in Task 2 Social-TV corpus and STOMPOL corpus results

To read a complete information about the systems, runs and results see (Martínez-Cámara et al., 2018).

3.3 Task 3. eHealth-KD

eHealth Knowledge Discovery (eHealth-KD) challenge proposed the identification of **Concepts** and **Actions**, for linking them later as a form of capturing the semantics of a broad range of health related text. Concepts are key-phrases that represent actors or entities relevant in a domain, while Actions represent how these Concepts interact with each other. For this challenge two types of relations: **Subject** and **Target**, were used to link Concepts and Actions (an special type of Concept), describing the main roles that a Concept can perform. In addition, other four specific semantic relations between Concepts were defined. A detailed description of the eHealth-KD challenge can be found in its web site[3] and in (Martínez-Cámara et al., 2018).

3.3.1 Subtasks and evaluation scenarios

eHealth-KD proposed two evaluation strategies: at subtask level (i.e. subtasks A, B and C) and per scenario. The subtasks were: **subtask A** concerned with the extraction of the relevant key phrases; **subtask** B concerned with classifying the key phrases identified in subtask A as either **Concept** or **Action**; and **subtask C** concerned with discovering the semantic relations between pairs of entities.

The evaluation scenarios were: **Scenario 1** which involved subtasks A, B and C sequentially; **Scenario 2** which involved subtasks B and C sequentially; and **Scenario 3** which only involved subtask C.

3.3.2 Participants

Six teams evaluated their systems on eHealth-KD 2018 challenge. These are listed next and classified regarding the characteristics that they used, referring the tag labels described in the next paragraph:**[Team UC3M]** [SDEN] (Zavala, Martínez, and Segura-Bedmar, 2018); **[Team SINAI]** [KRN] (López-Ubeda et al., 2018); **[Team UPF-UPC]** [SKN] (Palatresi and Hontoria, 2018); **[Team TALP]** [DEN] (Medina and Turmo, 2018); **[Team LaBDA]** [DE] (Suarez-Paniagua, Segura-Bedmar, and

[3] http://www.sepln.org/workshops/tass/2018/task-3/

Martínez, 2018); and **[Team UH]** [RN], which is described in (Martínez-Cámara et al., 2018). The tag labels designed to provide an overview of the characteristics of each system: **[S]** Used shallow supervised models such as CRF, logistic regression, SVM, decision trees, etc; **[D]** Used deep learning models, such as LSTM, convolutional networks, etc; **[E]** Used word embeddings or other embedding models trained with external corpora; **[K]** Used external knowledge bases, either explicitly or implicitly (i.e, through third-party tools); **[R]** Used hand crafted rules based on domain expertise; and **[N]** Used natural language processing techniques or features, i.e., POS-tagging, dependency parsing, etc.

3.3.3 Results

A variety of approaches dealt effectively with the health knowledge discovery problem. However, there are still issues to resolve. Classic supervised learning, deep learning and knowledge-based techniques were the best performing submissions, in general. The official results can be found in (Martínez-Cámara et al., 2018) and in the TASS 2018 web site[4]. From them the top results per subtask were:

- Subtasks A anb B: Team `UC3M` which was based on a CRF model with pre-trained embeddings as features. This team got F1 87.2% and Acc 95.9% in both subtasks respectively.
- Subtask C: in concordance with Scenario 3, did not exceed 45% in F-score. This can be considered as the most difficult subtask to deal with, even after having applied novel approaches (i.e. `TALP` with F1 44.8% and `LaBDA` with F1 42%) based on convolutional neural networks.

Regarding the top results per scenario. In:

- Scenario 1, the top performing strategy belonged to `UC3M` with an 74.4% of F1, pretty close to `SINAI` with 71.0%. These teams got a high F1 basically because their results in the subtasks A and B were high bringing on advantage in the average measure.
- Scenario 2 and 3: the top performing strategy belonged to `TALP` with an F1 of 72.2% and F1 of 44.8% respectively per scenario. This team reduced its advantage in the overall score, due to this did not submit results for the Scenario 1.

An interesting phenomenon is that the best systems in subtask A were not correlated 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.

3.3.4 Analysis of the results

The analysis of the results revealed that subtasks A and B were easier than subtask C for mostly participant teams. In subtask A, around 70% of the annotations in the test set were correctly identified by at least 3 of the participant systems. Likewise, in subtask B, 71% of the annotations were correctly classified by at least 4 systems. On the contrary, 64% of the relations in subtask C were not recognized by any system.

In general, the most competitive approaches in individual tasks were dominated by state-of-the-art machine learning. In the particular case of subtask C, modern deep learning approaches seemed to outperform classic techniques. However, adding domain-specific knowledge, mostly in the form of knowledge bases with health-related concepts, provided a significant boost, even when less powerful learning techniques were used, particularly for key phrase extraction (subtask A). Most participants used NLP features, either explicitly, or implicitly captured in word embeddings and other representations. The best overall systems did not generalize across the three subtasks, while systems that did generalize did not outperform the baseline in general.

3.4 Task 4: Good or Bad News?

Emotions are usually related to subjective data. However, the reading of facts, like the ones described in news, may also arise emotions. Task 4 is motivated by the industrial interest on the identification of the emotions that news headlines may arise on a reader, as they have an indubitable impact in the perception of ads placed along with those articles. The goal of Task 4 was defined as a binary classification problem: systems had to identify SAFE news (positive emotion) and

[4] `http://www.sepln.org/workshops/tass/2018/task-3/index.html\#results`

UNSAFE news (negative emotions).

3.4.1 Subtasks

Two subtasks were proposed: subtask 1 (S1) was set up as a monolingual classification task, and subtask 2 (S2) as a multilingual classification task.

S1 proposed the classification of headlines into SAFE or UNSAFE without taking into account the Spanish version. Participants were provided with the *training* and *development* subsets of the L1 SANSE corpus, and two *test* sets for the evaluation: the L1 SANSE test subset and the L2 SANSE corpus.

The aim of S2 was to assess the generalization capacity of the submitted systems. Participants were provided with SANSE subsets with headlines written only in the Spanish language spoken in Spain for training. The *test* set was composed of headlines written in the Spanish language spoken in different countries of America. Due to space constraints, the statistics of the SENSE corpus for S2 are shown in the web page of the task.[5]

3.4.2 Participants and Results

Seven research groups participated in TASK 4. Four of them submitted the results of their systems on the two subtasks, and three of them on the two runs of S1.

Each group was allowed to submit up to three systems. From all the submitted systems, we highlight the following: (1) most of the submitted systems were grounded in deep learning methods; (2) although most of the neural network systems were based on the use of Recurrent Neural Networks (RNN), specifically the Long Short-Term Memory (LSTM) architecture, Herrera-Planells and Villena-Román (2018) (MEANINGCLOUD) proposed the use of Convolutional Neural Networks (CNN), which reached good results in S1 L1; (3) the top performance system (INGEOTEC) in S1 and S2 was based on the optimization of a set of base linear classification systems using a genetic programming system; and (4) only one group (Plaza del Arco et al., 2018) (SINAI) proposed the incorporation of external knowledge to represent the headlines and subsequently used a linear classification system.

The evaluation measures used the macro-averaged version of the Precision, Recall and F1, as well as the Accuracy. Systems were ranked according to F1. The results of the best systems in the two subtasks are shown in Table 6. The main features of the three best systems are detailed next.

INGEOTEC. The system by Moctezuma et al. (2018) was the highest ranked one in S1 and S2. The system was an ensemble method built upon a genetic programming method, EvoDAG (Graff et al., 2017), for optimizing the contribution of each base system.

ELiRF_UPV. The system of González, Hurtado, and Pla (2018a) was based on the Deep Averaging Network (DAN) model. The main contributions were (1) the use of a set of pre-trained vectors of word embeddings in Spanish, which was generated from a set of Spanish tweets; and (2) the conclusion that the language used in news headlines and Twitter must be similar, as the use embeddings trained on tweets is not harmful for the system.

rbnUGR. The main contribution of Rodríguez Barroso, Martínez-Cámara, and Herrera (2018) was the comparison of three architectures of RNN for the encoding of the input headlines: (1) taking the last vector state of a LSTM layer; (2) the concatenation of the two last vector states of a Bidirectional LSTM layer; and (3) the concatenation of the output vectors of a LSTM layer per each input token. Results showed that the use of single LSTM layers is more beneficial, and the use of all the output vectors (run_3) allows to improve the generalization capacity, as it reached better results than the other two systems in S2.

3.4.3 Analysis

We conducted an analysis of the difficulty of the subtasks, which consisted on the study of the percentage of headlines correctly classified by the systems of the five groups that submitted a description paper.

We combined the output of the systems of each group[6] by a voting system, which resulted as the overall output of each group. The rate of headlines rightly predicted by the groups in each task is in Table 7. The analysis shows that S1 L2 is the least hard task because all the headlines were at least predicted by one group, as expected due to the fact that the annotation was performed by a voting system built upon the submitted systems.

[5] http://www.sepln.org/workshops/tass/2018/task-4/

[6] Three systems were allowed to submit as utmost.

System	S1 L1				S1 L2				S2			
	P	**R**	**F1**	**Acc**	**P**	**R**	**F1**	**Acc**	**P**	**R**	**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

Table 6: Macro averaged precision (P), recall (R), F1 and accuracy (Acc) reached by each submitted system to each subtask of the groups that submitted a system description paper. The superscripts are the rank order of the submitted systems

In contrast, the 4.40% and 5.57% of the headlines were not predicted by any group in S1 L2 and S2 respectively. Also as expected, the multilingual task (S2) is substantially harder than the monolingual one (S1 L1), because 34.14% of the headlines were classified by as much two groups, whereas only 17% of the headlines in S1.

#	S1 L1	(Acc.)	S1 L2	(Acc.)	S2	(Acc.)
0	4.40	4.40	0	0	5.57	5.57
1	5.00	9.40	0.69	0.69	10.80	16.37
2	7.60	17.00	5.35	6.04	17.77	34.14
3	12.00	29.00	13.96	20.01	30.66	64.80
4	23.00	52.00	26.04	46.05	35.19	100.00
5	48.00	100.00	53.94	100.00	-	-

Table 7: The % of headlines rightly classified by the groups. Only four groups participated in S2. Acc. indicates the accumulative percentage. Column one is the number of groups

Regarding the results shown in Table 6 and the statistics of Table 7, there is still room for improvement. First, only high performance systems can predict the safety meaning of 20% of the headlines, hence more efforts should be done in the design of classification systems to understand the meaning of the headlines. Subsequently, the subtask S2 pointed out the differences among the spoken Spanish versions, showing the need of increasing the generalization capacity of machine learning methods, which is essential for text understanding tasks with documents in different versions of the same language.

4 Conclusions

The edition of 2018 of TASS contributed with the organization of two new tasks and the update of the InterTASS corpus and the release of two new ones (eHealth-KD and SANSE).

The systems presented at Task 1, with the new version of the InterTASS corpora, obtains similar results, with F1 values near to 0.50. This means that systems can achieve better results with this new collection.

Task 3 was mostly dominated machine learning approaches. Nevertheless, adding domain-specific knowledge, mostly in the form of knowledge bases with health-related concepts, provided a significant boost, even when less powerful learning techniques were used. As future works, new semantic relations will be considered.

Task 4 showed an industrial application of emotion classification in a monolingual and multilingual environment. There is room for improvement in both environments, because there are headlines that were not rightly classified by any submitted system. As future work, the annotation of the SANSE corpus will be revised with the aim of improving the agreement score.

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), PROMETEU/2018/089 and SMART-DASCI (TIN2017-89517-P) from the Spanish Government. Eugenio Martínez Cámara was supported by the Spanish Government Programme Juan de la Cierva Formación (FJCI-2016-28353).

References

Chiruzzo, L. and A. Rosá. 2018. RETUYT-InCo at TASS 2018: Sentiment analysis in spanish variants using neural networks and svm. In *Proceedings of TASS 2018*, volume 2172, Sevilla, Spain. CEUR-WS.

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

González, J.-A., L.-F. Hurtado, and F. Pla. 2018a. ELiRF-UPV en TASS 2018: Categorización emocional de noticias. In *Proceedings of TASS 2018*, volume 2172, Sevilla, Spain. CEUR-WS.

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 *Proceedings of TASS 2018*, volume 2172, pages 37–44, Sevilla, Spain. CEUR-WS.

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

Herrera-Planells, J. and J. Villena-Román. 2018. MeaningCloud at TASS 2018: News headlines categorization for brand safety assessment. In *Proceedings of TASS 2018*, volume 2172, Sevilla, Spain. CEUR-WS.

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

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*, volume 2172, Sevilla, Spain. CEUR-WS.

Luque, F. M. and J. M. Pérez. 2018. Atalaya at TASS 2018: Sentiment analysis with tweet embeddings and data augmentation. In *Proceedings of TASS 2018*, volume 2172, Sevilla, Spain. CEUR-WS.

Martínez-Cámara, E., 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, and J. Villena-Román. 2018. Overview of TASS 2018: Opinions, health and emotions. In *Proceedings of TASS 2018*, volume 2172, Sevilla, Spain. 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 *Proceedings of TASS 2017*, volume 1896, Murcia, Spain. CEUR-WS.

Medina, S. and J. Turmo. 2018. Joint classification of key-phrases and relations in electronic health documents. In *Proceedings of TASS 2018*, volume 2172, Sevilla, Spain. CEUR-WS.

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*, volume 2172, Sevilla, Spain. 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 *Proceedings of TASS 2018*, volume 2172, Sevilla, Spain. 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*, volume 2172, Sevilla, Spain. CEUR-WS.

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*, volume 2172, Sevilla, Spain. CEUR-WS.

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*, volume 2172, Sevilla, Spain. CEUR-WS.

Suarez-Paniagua, V., I. Segura-Bedmar, and P. Martínez. 2018. ABDA at TASS-2018 task 3: Convolutional neural networks for relation classification in spanish ehealth documents. In *Proceedings of TASS 2018*, volume 2172, Sevilla, Spain. CEUR-WS.

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*, volume 2172, Sevilla, Spain. CEUR-WS.