Procesamiento del Lenguaje Natural, Revista nº 52 marzo de 2014, pp 37-44 recibido 15-11-13 revisado 06-02-14 aceptado 07-02-14

TASS 2013 - A Second Step in Reputation Analysis in Spanish

TASS 2013 - Un Segundo Paso en Análisis de Reputación en Español

Julio Villena-Román
Janine García-Morera
Daedalus, S.A.
Av. de la Albufera 321
28031 Madrid, Spain
{jvillena, jgarcia}@daedalus.es

Sara Lana-Serrano
José Carlos González-Cristóbal
Universidad Politécnica de Madrid
E.T.S.I. Telecomunicación
Ciudad Universitaria s/n
28040 Madrid, Spain
jgonzalez@dit.upm.es, slana@diatel.upm.es

Resumen: TASS 2013 es la segunda edición del taller de evaluación experimental en el congreso anual de la SEPLN dedicado al análisis de reputación en español. El principal objetivo es fomentar la investigación en técnicas y algoritmos avanzados para realizar análisis de sentimientos y clasificación automática de opiniones extraídas de mensajes cortos en medios sociales en español. Este artículo describe en profundidad, en comparación con la edición anterior, las tareas propuestas este año, el contenido, formato y las estadísticas principales de los corpus generados, los participantes y los diferentes enfoques planteados, así como los resultados generales obtenidos y las lecciones aprendidas en estos dos años.
Palabras clave: TASS 2013, análisis de reputación, análisis de sentimientos, clasificación automática de texto, medios sociales, español.

Abstract: TASS 2013 is the second edition of the experimental evaluation workshop within the SEPLN annual Conference focused on reputation analysis in Spanish language. The main objective is to foster the research on advanced algorithms and techniques for performing sentiment analysis and automatic text categorization on opinions extracted from short social media messages in Spanish. This paper fully describes the proposed tasks, the contents, format and main figures of the generated corpus, the participant groups and their different approaches, and, finally, the overall results achieved and lessons learned in these two years.
Keywords: TASS 2013, reputation analysis, sentiment analysis, text categorization, social media, Spanish.

1 Introduction

TASS is an experimental evaluation workshop on reputation analysis focused on Spanish language, organized as a satellite event of the SEPLN Conference. After a successful first edition in 2012 (Villena-Román et al., 2013), TASS 2013[1] was held on September 20th, 2013 at Universidad Complutense de Madrid, Spain.

The long-term objective of TASS is to foster research in the field of reputation analysis, i.e., the process of tracking, investigating and reporting an entity's actions and other entities' opinions about those actions, in Spanish language. As a first approach, reputation analysis has at least two technological aspects: sentiment analysis and text classification.

Sentiment analysis is the application of natural language processing and text analytics to identify and extract subjective information from texts. It is a major technological challenge and the task is so hard that even humans often disagree on the sentiment of a given text, as issues that one individual may find acceptable or relevant may not be the same to others. And the shorter the text is (for instance, Twitter messages or short comments in Facebook), the harder the task becomes.

On the other hand, automatic text classification (or categorization) is used to guess the topic of the text, among those of a predefined set of categories, so as to be able to assign the reputation level into different axis or

[1] http://www.daedalus.es/TASS2013

points of view of analysis. Text classification techniques, although studied for a long time, still need more research effort to be able to build complex models with many categories with less workload and increase the precision and recall of the results. In addition, these models should deal with specific text features in social media messages (such as spelling mistakes, abbreviations, etc.).

Within this context, the aim of TASS is to provide a forum for discussion where the latest research work in these fields can be discussed by scientific and business communities. The setup is based on a series of challenge tasks intended to provide a benchmark forum for comparing different approaches. In addition, with the creation and open release of the fully tagged corpus, the aim is to provide a common reference dataset for the research community.

The rest of the paper is organized as follows. Section 2 describes the corpus provided to participants and used for the challenge tasks. The third section describes the different tasks proposed this edition. Section 4 describes the participants and the overall results are presented in Section 5. The last section draws some conclusions and future directions.

2 Corpus

Experiments were based on two corpus. After the workshop, both were made freely available for research purposes to the community. The only requirement was to make a request to tass@daedalus.es with the email, affiliation and a brief description of the research objectives, and include a proper citation in publications.

2.1 General corpus

The general corpus, the same used in 2012, contains over 68000 Twitter messages, written in Spanish by about 150 well-known personalities and celebrities of the world of politics, economy, communication, mass media and culture, between November 2011 and March 2012. Each message includes its ID (`tweetid`), the creation date (`date`) and the user ID (`user`). According to the Twitter API Terms of Service[2], text contents and user information had to be removed for the corpus distribution.

The general corpus was divided into two sets: *training* (about 10%) and *test* (90%). The training set was released so that participants could train and validate their models. The test corpus was provided without any tagging and was used to evaluate the results provided by the different systems. Table 1 shows a summary of the training and test sets.

Attribute	Value
Tweets	68 017
Tweets (train)	7 219 (11%)
Tweets (test)	60 798 (89%)
Topics	10
Users	154
Date start	2011-12-02 T00:03:32
Date end	2012-04-10 T23:47:55

Table 1: General corpus statistics

Each message in both the training and test set was tagged with its *global polarity*, indicating whether the text expresses a positive, negative or neutral sentiment, or no sentiment at all. 5 levels have been defined: strong positive (`P+`), positive (`P`), neutral (`NEU`), negative (`N`), strong negative (`N+`) and one additional no sentiment tag (`NONE`).

In addition, the *level of agreement* of the expressed sentiment within the text was also included, to make out whether a neutral sentiment comes from neutral keywords (`AGREEMENT`) or else the text contains positive and negative sentiments at the same time (`DISAGREEMENT`).

Moreover, the *polarity at entity level*, i.e., the polarity values related to the entities that are mentioned in the text, was also included for those cases when applicable. These values were similarly divided into 5 levels and include the level of agreement as related to each entity.

On the other hand, a selection of a set of *topics* was made based on the thematic areas covered by the corpus, such as politics, literature or entertainment. Each message in both the training and test set was assigned to one or several of these topics. The list of selected topics is shown later in Table 7.

All tagging was carried out semi automatically: a baseline machine learning model was first run (Villena-Román et al., 2011) and then all tags were manually checked by two human experts. For polarity at entity level, due to the high volume of data to check, this tagging was done just for the training set.

Figure 1 shows the information of two sample tweets. The first tweet is only tagged

[2] https://dev.twitter.com/terms/api-terms

with the global polarity (P+) and the agreement level (AGREEMENT), as it contains no mentions to any entity, but the second one is tagged with both the global polarity (P), the agreement level (AGREEMENT) and the polarity associated to each of the entities that appear in the text (UPyD and Foro Asturias, both tagged as P).

```
<tweet>
    <tweetid>0000000000</tweetid>
    <user>usuario0</user>
    <content>
        <![CDATA['Conozco a alguien q es adicto al drama! Ja ja ja te suena d algo!]]>
    </content>
    <date>2011-12-02T02:59:03</date>
    <lang>es</lang>
    <sentiments>
        <polarity>
            <value>P+</value>
            <type>AGREEMENT</type>
        </polarity>
    </sentiments>
    <topics>
        <topic>entretenimiento</topic>
    </topics>
</tweet>
<tweet>
    <tweetid>0000000001</tweetid>
    <user>usuario1</user>
    <content>
        <![CDATA['UPyD contará casi seguro con grupo gracias al Foro Asturias.]]>
    </content>
    <date>2011-12-02T00:21:01</date>
    <lang>es</lang>
    <sentiments>
        <polarity>
            <value>P</value>
            <type>AGREEMENT</type>
        </polarity>
        <polarity>
            <entity>UPyD</entity>
            <value>P</value>
            <type>AGREEMENT</type>
        </polarity>
        <polarity>
            <entity>Foro_Asturias</entity>
            <value>P</value>
            <type>AGREEMENT</type>
        </polarity>
    </sentiments>
    <topics>
        <topic>política</topic>
    </topics>
</tweet>
```

Figure 1: Sample tweets (General corpus)

2.2 Politics corpus

The Politics corpus, new in this edition, contains 2500 tweets, gathered[3] during the electoral campaign of the 2011 General Elections in Spain from Twitter messages mentioning any of the four main national-level political parties: PP, PSOE, IU and UPyD.

Similarly to the General corpus, the global polarity and the polarity at entity level for those four entities was manually tagged for all messages. However, in this case, due to the lack of time and the high amount of work that the tagging required, only 3 levels were used: positive (P), neutral (NEU), negative (N), and one additional no sentiment tag (NONE).

The format was the same as the General corpus, but the entity element includes a source attribute indicating the political party the entity refers to.

[3] This corpus was completely built by E. Martínez-Cámara (SINAI group, Universidad de Jaen), member of the organization of TASS 2013.

The following figure shows the information of one sample tweet. The global polarity is N with AGREEMENT, and the polarity at entity level for the entity *@marianorajoy* whose source is *PP* is also N with AGREEMENT.

```
<tweet>
    <tweetid>137231808990412800</tweetid>
    <user>marianarajay</user>
    <content>
        <![CDATA['@marianorajoy Por favor, amigosh, no me votéish
        que me lo he penshado mejor con este tshunami que se me viene encima.]]>
    </content>
    <date>2011-10-17T19:13:07</date>
    <lang>es</lang>
    <sentiments>
        <polarity>
            <value>N</value>
            <type>AGREEMENT</type>
        </polarity>
        <polarity>
            <entity source="PP">@marianorajoy</entity>
            <value>N</value>
            <type>AGREEMENT</type>
        </polarity>
    </sentiments>
    <topics>
        <topic>política</topic>
    </topics>
</tweet>
```

Figure 2: Sample tweet (Politics corpus)

3 Tasks

This year four tasks were proposed, extending the two tasks that were offered in TASS 2012, covering different aspects of sentiment analysis and text classification.

3.1 Task 1: Sentiment Analysis at Global Level

This task consisted on performing an automatic sentiment analysis to determine the global polarity of each message in the test set of the General corpus. Participants were provided with the training set of the General corpus so that they could train and validate their models.

There are two different evaluation criteria: i) fine-grained polarity using 5 levels, and ii) coarse-grained polarity with just 3 levels.

The standard metrics of precision, recall and F-measure calculated over the test set are used to evaluate and compare the different systems.

3.2 Task 2: Topic Classification

The challenge of this task was to automatically identify the topic of each message in the test set of the General corpus. Participants could use the training set of the General corpus to train and validate their models.

3.3 Task 3: Sentiment Analysis at Entity Level

This task was similar to Task 1, but sentiment polarity (using 3 levels) should be determined

at entity level of each message in the Politics corpus. In this case, the polarity at entity level included in the training set of the General corpus could be used by participants to train and validate the models (converting from fine-grained to coarse-grained polarity).

Entities were tagged in the corpus to make participant focus on the sentiment analysis and not on entity recognition. The difficulty of the task arises from the fact that messages can contain more than one sentence with more than one entity per sentence, so more advanced text processing techniques are needed.

3.4 Task 4: Political Tendency Identification

This task moves one step forward towards reputation analysis and the objective is to estimate the political tendency of each user in the test set of the General corpus, in four possible values: `LEFT`, `RIGHT`, `CENTRE` and `UNDEFINED`. Participants could use whatever strategy they decide, but a first approach could be to aggregate the results of the previous tasks by author and topic.

4 Participants

31 groups registered (compared to 15 groups last year) and finally 14 groups (9 last year) sent their submissions. The list of active groups is shown in Table 2, including the tasks in which they have participated.

Group	1	2	3	4
CITIUS-Cilenis	X		X	
DLSI-UA	X			
Elhuyar	X			
ETH-Zurich	X	X	X	X
FHC25-IMDEA		X		
ITA	X			
JRC	X			
LYS	X	X		X
SINAI-EMML	X			
SINAI-CESA	X	X	X	X
Tecnalia-UNED	X			
UNED-JRM	X	X		
UNED-LSI	X	X		
UPV	X	X	X	X
Total groups	13	7	4	4

Table 2: Participant groups (Díaz Esteban, Alegría y Villena Román, 2013)

Along with the experiments, all participants were invited to submit a paper to describe their experiments and discuss the results with the audience in the workshop session. These papers should follow the usual SEPLN template and could be written in Spanish or English. Papers were reviewed by the program committee and were included in the workshop proceedings (Díaz Esteban, Alegría y Villena Román, 2013).

In these two years, the trend has been to adopt a machine learning supervised approach to sentiment analysis, mainly using Weka (Hall et al., 2009), with a text processing often using Freeling (Padró and Stanilovsky, 2012).

For instance, CITIUS-Cilenis runs achieved a good performance using a Naive-Bayes binary classifier to distinguish between just two sharp polarity categories (positive and negative) and used experimentally set thresholds for detecting the fine grain polarity values. Another supervised approach based in SVM is used by Elhuyar, including linguistic knowledge-based processing with Freeling and tagging of polarity words, emoticons, negation and spelling errors.

Similarly, UPV used a SVM approach (based on libSVM library for Weka) and submitted runs for all tasks that are often at the top results. In addition, as the type of language used in social networks (non-grammatical phrases, lack or misuse of punctuation symbols, specific terminology, etc.) is not covered by the standard publicly available tools, they made specific adaptations for improving the tokenization. Both Freeling and Tweetmotif[4] adapted to Spanish were used.

JRC also adapted a supervised approach based on different feature combinations, originally designed for English to Spanish, using several in-house built dictionaries and machine translated data. UNED-JRM also deals with both Task 1 and 2 as purely-classification tasks, developing a classifier indifferently for both tasks, with similar results.

Tecnalia-UNED also rely on advanced linguistic process (again based on Freeling) to deal with complex issues such as negation detection and emphatiser treatment (aiming at distinguishing the range of polarity levels).

LYS present the best-performing approach in the topic classification task. In addition to an ad-hoc normalization process, POS tagging and dependency parsing algorithms are applied and psychological resources are used to exploit the psychometric properties of human language (Vilares, Alonso and Gómez-Rodríguez, 2013).

[4] https://github.com/brendano/tweetmotif

Quite differently, SINAI-EMML group opted for a completely unsupervised strategy based on the combination of three linguistic resources, SentiWordNet, Q-WordNet and iSOL. The polarity value is calculated with the normalized addition of the differences between the positive and negative values of each term.

Sentiment lexicons are also present in most systems. For instance, the contribution from DLSI-UA consisted of two different graph-based approaches: a modified version of a ranking algorithm (RA-SR) using bigrams, used on the Task 2 of the Semeval 2013 competition[5], and a new proposal using a skipgrams scorer. Both approaches create sentiment lexicons able to retain the context of the terms, and employ machine learning techniques to detect the polarity of a text. All their runs appear in the top 10 best results and their combination reaches the first position.

Another graph-based approach for topic classification is presented by FHC25-IMDEA. They used a technique based on graph similarity to classify Twitter messages. Their assumption is that any text can be represented as a graph. For a given text, their system places the terms (actually the stems) in the vertexes of a graph and creates links with a given weight among them. Then their hypothesis is that graphs belonging to texts of the same topic usually form unique structures (i.e., a topic graph). Thus, a metric is used for calculating the similarity between the text graph to classify and the different topic graphs.

Other interesting approaches are based on Information Retrieval (IR) techniques. For instance, SINAI-CESA propose a solution using Latent Semantic Analysis. Train data is taken from the continuous stream of posts from Twitter, capturing those that are likely to include affective expressions and generating a corpus of "feelings" labeled according to their polarity, and without using any training data from controlled corpora to avoid suffering from domain related limitations.

Similarly, UNED-LSI adopt an IR approach where the classes are modeled according to the textual information of the tweets belonging to each class, and used as queries (Castellanos, Cigarrán and García-Serrano, 2012).

ITA group made some experiments with the Non-Axiomatic Reasoning System[6], a general-purpose reasoning system, as a tool to dynamically discover content words and phrases with opinion. Their idea is to use a seed dictionary to look for similar polarity words.

Last but not least, ETH-Zurich present an interesting study of political discourse and emotional expression by analyzing the political position of four major parties through their Twitter activity, revealing that Twitter political discourse depends on subjective perception, and resembles the political space of Spain.

5 Results

Participants were expected to submit one or several runs for one or several of the tasks. Results should be submitted in a plain text file with the following format:

```
id \t output \t confidence
```

where:

- `id` is the tweet ID for Tasks 1 and 2, the combination of tweet ID and entity for Task 3, and the user ID for Task 4.
- `output` refers to the expected output of each task (polarity, topic, political tendency).
- `confidence` is a number ranging [0, 1] that indicates the confidence as assigned by the system (not currently used for evaluation).

After the submission deadline, runs were collected by the organization and the evaluation results were made available to the participants to allow them to prepare their reports. Results included a spreadsheet with the overall evaluation figures for each task, and also detailed results per experiment for all the 5 evaluations (as explained before, Task 1 was evaluated using both 5-level and 3-level setups), the confusion matrix with all labels to allow error analysis, and finally the gold standard for the task itself. The PHP script used for the evaluation of each submission was also included for their convenience.

5.1 Task 1: Sentiment Analysis at Global Level

56 runs (10 of them specific for 3-level evaluation) were submitted by 13 different groups. Results for the best ranked experiment from each groups are listed in the tables below. All tables show the precision (P), recall (R) and F1 value achieved in each experiment. Table 3 considers 5 polarity levels. Precision values range from 61.6% to 12.6%. The average values are 43.3% for all metrics.

[5] http://www.cs.york.ac.uk/semeval-2013/
[6] https://sites.google.com/site/narswang/

Run Id	P	R	F1
DLSI-UA-pol-dlsiua3-3-5l	0.616	0.616	0.616
Elhuyar-TASS2013 _run1	0.601	0.601	0.601
UPV_ELiRF_task1_run2	0.576	0.576	0.576
CITIUS-task1_ 1	0.558	0.558	0.558
lys_global_sentiment_task_6c	0.553	0.553	0.553
JRC-tassTrain-base-DICT-5way	0.519	0.519	0.519
ITA_ResultadosAnalisisOpiniónAlg	0.439	0.439	0.439
LSI_UNED_2_TASK1_RUN_09	0.402	0.402	0.402
UNED-JRM-task1-run2	0.393	0.393	0.393
TECNALIA-UNED	0.348	0.344	0.346
ETH-task1-Warriner	0.328	0.328	0.328
sinai_emml_task1_6classes	0.314	0.314	0.314
sinai_cesa-task1_raw	0.135	0.134	0.134

Table 3: Results for Task 1 with 5 levels

Table 4 gives results considering the classification only in 3 levels. In this case, precision values improve, as expected as the task seems to be easier. The precision obtained now ranges from 68.6% to 23.0%. The average values for all metrics in this case is 53.0%.

Run Id	P	R	F1
Elhuyar-TASS2013 _run1	0.686	0.686	0.686
UPV_ELiRF_task1_run2	0.674	0.674	0.674
CITIUS-task1_ 1	0.668	0.668	0.668
DLSI-UA-pol-dlsiua3-3-5l	0.663	0.663	0.663
lys_global_sentiment_task_6c	0.657	0.657	0.657
JRC-tassTrain-base-DICT-3way	0.612	0.612	0.612
ITA_ResultadosAnalisisOpiniónAlg	0.543	0.543	0.543
TECNALIA-UNED	0.496	0.490	0.493
UNED-JRM-task1-run2	0.496	0.496	0.496
LSI_UNED_2_TASK1_RUN_06	0.479	0.479	0.479
ETH-task1-Warriner	0.466	0.466	0.466
sinai_emml_task1_3classes	0.409	0.409	0.409
sinai_cesa-task1_raw	0.389	0.388	0.388

Table 4: Results for Task 1 with 3 levels

Initially a gold standard was generated by pooling all submissions with a voting scheme and then an extensive human review of the ambiguous decisions was carried out. However, as some groups had submitted many runs and other groups had only submitted a few, some concern arose about a possible bias. To avoid any systemic problem, the gold standard creation should be repeated or at least carefully evaluated for correctness. Due to the summer holidays and lack of human resources for the task, finally the gold standard of TASS 2012, which was not subject to this bias as the number of submissions was balanced, was used to evaluate the submissions.

The distribution of labels in both the training and test corpus is shown in Table 5. Obviously, the distribution is not evenly balanced in both corpus, i.e., the gold standard may be not well built. This fact causes that, for example, given a system that is able to correctly classify `P+` and `NONE` with a high precision (both count 70% of tweets in test corpus), and maybe, not so good at classifying the other labels, may achieve better results on the test corpus than the training corpus, as it is actually reported by some participants (CITIUS-Cilenis and Elhuyar). Obviously this has to be taken into account for future initiatives.

Label	Frequency (Train)	Frequency (Test)
P+	22.44%	34.12%
P	4.12%	2.45%
NEU	8.45%	2.15%
N	16.91%	18.56%
N+	12.51%	7.5%
NONE	23.58%	35.22%

Table 5: Sentiment distribution

This is for example the case of *CITIUS-task1_1* run, which achieves better results than *lys_global_sentiment_task_6c*, but is worse balanced (Table 6).

Label	CITIUS	LYS	DLSI	Elhuyar
P+	0.791	0.578	0.705	0.638
P	0.363	0.569	0.263	0.661
NEU	0.022	0.195	0.108	0.185
N	0.289	0.548	0.586	0.583
N+	0.533	0.526	0.390	0.427
NONE	0.546	0.557	0.649	0.631
all	*0.558*	*0.553*	*0.616*	*0.601*

Table 6: Precision per sentiment label

Another interesting comparison is the top ranked run, *DLSI-UA-pol-dlsiua3-3-5l*, vs the second ranked, *TASS2013_Elhuyar_run1*. Results from Elhuyar are quite balanced and can be compared to the LYS run, but they are better ranked as they achieve greater precision for all labels but `N+` and `NEU`. In turn, results

from DLSI are better than Elhuyar run because their system performs better for P+ and NONE that are the most frequent labels. This issue must be studied for eventual future editions.

5.2 Task 2: Topic Classification

This task was evaluated as a single label classification. The most restrictive criterion has been applied: a "success" is achieved only when all the test labels have been returned. As in Task 1, the gold standard finally considered was the one used in TASS 2012.

The distribution of topics in both the train and test corpus is shown in Table 7. The total count is greater than the number of tweets as several topics could be assigned per tweet.

Topic	Frequency (Train)	Frequency (Test)
Politics	3 120 (33%)	30 067 (43%)
Other	2 337 (24%)	28 191 (40%)
Entertainment	1 678 (17%)	5 421 (8%)
Economy	942 (10%)	2 549 (3%)
Music	566 (6%)	1 498 (2%)
Soccer	252 (3%)	823 (1%)
Films	245 (3%)	596 (1%)
Technology	217 (2%)	287 (0%)
Sports	113 (1%)	135(0%)
Literature	103 (1%)	93(0%)
all	*9 573*	*69 660*

Table 7: Topic distribution

20 experiments were submitted in all. Table 8 shows the results for this task. The average values are 62.4% precision, 44.4% recall and 49.6 F1. Precision ranges from 80.4% to 16.1%. As in Task 1, different submissions from the same group usually have similar values. No approach (learning, graph or IR-based) clearly stand out among the others.

Run Id	P	R	F1
lys_topic_task_with_user_info	0.804	0.804	0.804
LSI_UNED_2_TASK2_RUN_07	0.777	0.184	0.298
UPV_ELiRF_task2_run2	0.756	0.756	0.756
ETH-task2	0.734	0.455	0.562
FHC25-IMDEAults_PR_GD_TT	0.719	0.702	0.710
FHC25-IMDEAults_PR_TT	0.705	0.688	0.696
UNED-JRM-task2-run2	0.479	0.479	0.479
sinai_cesa-task2_normalized	0.161	0.159	0.160

Table 8: Results for Task 2

Some participants such as FHC25-IMDEA pointed out that, as shown in Table 7, the distribution is quite balanced between both corpus but not on different topics. This may cause that the trained systems tend to be biased towards the most frequent topics (politics and other). Systems that are optimized for those categories, even at the cost of a low performance in the less frequent topics, will seem to achieve a better overall result than a system that is more balanced system.

5.3 Task 3: Sentiment Analysis at Entity Level

The evaluation was made over the Politics corpus, which was tagged manually, so the gold standard was created with no pooling. Finally 6 runs were submitted for this task.

Results are shown in Table 9. Average precision is 37.2%, recall is 36.5% and F1 is 36.9%. These figures are much lower than in Task 1. This is because this task is harder than Task 1 and systems do not reach the adequate level of development as learning-based approaches are not able to represent enough knowledge about the text semantic contents.

Run Id	P	R	F1
CITIUS-task3_CITIUS.txt	0.411	0.378	0.394
UPV_ELiRF_task3_run0.txt	0.395	0.395	0.395
sinai_cesa-task3_normalized.tsv	0.384	0.384	0.384
ETH-task3.txt	0.307	0.307	0.307

Table 9: Results for Task 3

5.4 Task 4: Political Tendency Identification

The gold standard was built manually by reviewing each user's political tendency, as defined by himself/herself, or assigning UNDEFINED if not stated or unknown. 11 runs were submitted. Results are shown in Table 10.

Run Id	P	R	F1
ETH-Task4-Crowdsource.txt [MANUAL]	0.734	0.734	0.734
UPV_ELiRF_task4_run1.txt	0.703	0.703	0.703
sinai_cesa-task4_nound_raw.tsv	0.583	0.399	0.474
lys_political_tendency_task_model2	0.424	0.424	0.424

Table 10: Results for Task 4

Average values for precision, recall and F1 are 57.7%, 51.7% and 54.1% respectively. Run from ETH is based on a manual assignment of political tendency to each user, made with crowdsourcing, so it is supposed to achieve the best result in the gold standard, as it happens.

6 Conclusions and Future Work

TASS is the first workshop about reputation analysis specifically focused on Spanish. This second edition of TASS has been even more successful than the first one, as the number of participants has increased up to 31 groups registered (15 groups last year) and 14 groups (9 last year) sent their submissions. The number of participants and the quality of their work has met and gone beyond all our expectations.

It is still necessary to perform a more detailed analysis of the results, which is in our short-term roadmap. However, reports from participants and the developed corpora are already valuable available resources, helpful for other research groups approaching these tasks.

Furthermore, the reuse of the General corpus in these two years allows to analyze the evolution in the field and provides a benchmark for future research. TASS 2012 corpus has been downloaded by more than 50 research groups, 20 out of Spain. We hope to reach a similar impact with the new corpus.

Some ideas for future editions gathered during the workshop involve solving the corpus uneven distribution, the inclusion of text normalization issues, the development of new corpus with different varieties of Spanish, and some tasks related to irony detection, mixed sentiments (disagreement within the text), subjectivity and the speaker point of view (first person vs eyewitness vs hearsay witness).

Acknowledgements

This work has been supported by several Spanish R&D projects: *Ciudad2020: Hacia un nuevo modelo de ciudad inteligente sostenible* (INNPRONTA IPT-20111006), *MA2VICMR: Improving the access, analysis and visibility of the multilingual and multimedia information in web for the Region of Madrid* (S2009/TIC-1542) and *MULTIMEDICA: Multilingual Information Extraction in Health domain and application to scientific and informative documents* (TIN2010-20644-C03-01).

References

Castellanos, A., J. Cigarrán, y A. García-Serrano. 2012. Generación de un corpus de usuarios basado en divergencias del Lenguaje. *II Congreso Español de Recuperación de Información*. Valencia, June 2012.

Díaz Esteban, A., I. Alegría, y J. Villena Román (eds). 2013. *Actas del XXIX Congreso de la Sociedad Española de Procesamiento de Lenguaje Natural.* IV Congreso Español de Informática. 17-20 September 2013, Madrid, Spain. ISBN: 978-84-695-8349-4.

Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, Volume 11, Issue 1.

Padró, L. and E. Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pp 2473-2479, Istanbul, Turkey.

Vilares, D., M.A. Alonso, y C. Gómez-Rodríguez. 2013. Una aproximación supervisada para la minería de opiniones sobre tuits en español en base a conocimiento lingüístico. *Revista de Procesamiento del Lenguaje Natural*, 51, pp 127-134, sep 2013. ISSN 1135-5948.

Villena-Román, J., S. Collada-Pérez, S. Lana-Serrano, and J.C. González-Cristóbal. 2011. Hybrid Approach Combining Machine Learning and a Rule-Based Expert System for Text Categorization. In *Proceedings of the 24th International Florida Artificial Intelligence Research Society Conference (FLAIRS-11),* May 18-20, 2011, Palm Beach, Florida, USA. AAAI Press 2011.

Villena-Román, J., S. Lana-Serrano, E. Martínez-Cámara, and J.C. González-Cristobal. 2013. TASS - Workshop on Sentiment Analysis at SEPLN. *Revista de Procesamiento del Lenguaje Natural,* 50, pp 37-44, mar 2013. ISSN 1135-5948.