TASS 2014 - Workshop on Sentiment Analysis at SEPLN - Overview

TASS 2014 - Taller de Análisis de Sentimientos en la SEPLN

Julio Villena Román
Janine García Morera
César de Pablo Sánchez
Daedalus, S.A.
28031 Madrid, Spain
{jvillena, jgarcia, cdepablo}@daedalus.es

Miguel Ángel García Cumbreras
Eugenio Martínez Cámara
Alfonso Ureña López
María Teresa Martín Valdivia
Universidad de Jaén
23071 Jaén, Spain
{magc, emcamara, laurena, maite}@uja.es

Resumen: Este artículo describe el desarrollo de TASS 2014, la tercera edición del taller de evaluación experimental en el contexto de la SEPLN para fomentar la investigación en el campo del análisis de la opinión en los medios sociales, específicamente centrado en el idioma español. El principal objetivo es promover el diseño de nuevas técnicas y algoritmos y la aplicación de los ya existentes para la implementación de sistemas complejos capaces de realizar un análisis de opiniones basados en opiniones de textos cortos extraídos de medios sociales (concretamente Twitter). Este artículo describe las tareas propuestas en la edición de 2014, el contenido de los corpus generados, la lista de participantes y los resultados generales obtenidos.
Palabras clave: TASS 2014, análisis de opiniones, medios sociales

Abstract: This paper describes TASS 2014, the third edition of an experimental evaluation workshop within SEPLN to foster the research in the field of sentiment analysis in social media, specifically focused on Spanish language. The main objective is to promote the application of existing state-of-the-art algorithms and techniques and the design of new ones for the implementation of complex systems able to perform a sentiment analysis based on short text opinions extracted from social media messages (specifically Twitter). The paper presents the proposed tasks in the 2014 edition, the contents of the generated corpora, the participant groups and the overall results achieved.
Keywords: TASS 2014, sentiment analysis, social media.

1 Introduction

TASS is an experimental evaluation workshop for sentiment analysis and online reputation analysis focused on Spanish language, organized as a satellite event of the annual SEPLN Conference. After two successful editions in 2012[1] (Villena *et al.*, 2013) and 2013[2] (Villena *et al.*, 2014), TASS 2014[3] will be held on September 16th, 2014 at Universitat de Girona, Spain.

According to Merriam-Webster dictionary, reputation is the overall quality or character of a given person or organization as seen or judged by people in general, or, in other words, the general recognition by other people of some characteristics or abilities for a given entity. Specifically, in business, reputation comprises the actions of a company and its internal stakeholders along with the perception of consumers about the business. Reputation affects attitudes like satisfaction, commitment and trust, and drives behavior like loyalty and support. In turn, reputation analysis is the process of tracking, investigating and reporting an entity's actions and other entities' opinions about those actions. It covers many factors to calculate the market value of reputation. Reputation analysis has come into wide use as a major factor of competitiveness in the

[1] http://www.daedalus.es/TASS2012
[2] http://www.daedalus.es/TASS2013
[3] http://www.daedalus.es/TASS2014

increasingly complex marketplace of personal and business relationships among people and companies.

Currently market research using user surveys is typically performed. However, the rise of social media such as blogs and social networks and the increasing amount of user-generated contents in the form of reviews, recommendations, ratings and any other form of opinion, has led to creation of an emerging trend towards online reputation analysis. This analysis has two technological aspects: sentiment analysis and text classification (or categorization).

First, the so-called sentiment analysis, i.e., the application of natural language processing and text analytics to identify, extract and classify subjective information from texts, which is the first step towards the online reputation analysis, is becoming a promising topic in the field of marketing and customer relationship management, as the social media and its associated word-of-mouth effect is turning out to be the most important source of information for companies and their customers' sentiments towards their brands and products.

Then, automatic text classification is used to guess the topic of the text, among those of a predefined set of categories or classes, so as to be able to assign the reputation level of the company into different facets, axis or points of view of analysis.

Sentiment analysis is a major technological challenge. The task is so hard that even humans often disagree on the sentiment of a given text. The fact that issues that one individual finds acceptable or relevant may not be the same to others, along with multilingual aspects, cultural factors and different contexts make it very hard to classify a text written in a natural language into a positive or negative sentiment. And the shorter the text is, for example, when analyzing Twitter messages or short comments in Facebook, the harder the task becomes.

On the other hand, text classification techniques, although studied for a longer 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 work well with short texts and deal with specific text features that are present in social media messages (such as spelling mistakes, abbreviations, SMS language, etc.).

Within this context, the aim of TASS is to provide a forum for discussion and communication where the latest research work and developments in the field of sentiment analysis in social media, specifically focused on Spanish language, can be shown and discussed by scientific and business communities. The main objective is to promote the application of existing state-of-the-art algorithms and techniques and the design of new ones for the implementation of complex systems able to perform a sentiment analysis and text classification on short text opinions extracted from social media messages (specifically Twitter) published by a series of representative personalities.

The challenge task is intended to provide a benchmark forum for comparing the latest approaches in these fields. In addition, with the creation and release of the fully tagged corpus, we aim to provide a benchmark dataset that enables researchers to compare their algorithms and systems.

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

TASS 2014 experiments will be based on two different corpus.

2.1 General corpus

The general corpus contains over 68 000 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. Although the context of extraction has a Spain-focused bias, the diverse nationality of the authors, including people from Spain, Mexico, Colombia, Puerto Rico, USA and many other countries, makes the corpus reach a global coverage in the Spanish-speaking world.

Each Twitter message includes its ID (*tweetid*), the creation date (*date*) and the user ID (*user*). Due to restrictions in the Twitter API

Terms of Service[4]), it is forbidden to redistribute a corpus that includes text contents or information about users. However, it is valid if those fields are removed and instead IDs (including Tweet IDs and user IDs) are provided. The actual message content can be easily obtained by making queries to the Twitter API using the *tweetid*.

The general corpus has been divided into two sets: training (about 10%) and test (90%). The training set will be released so that participants may train and validate their models. The test corpus will be provided without any tagging and will be used to evaluate the results provided by the different systems. Obviously, it is not allowed to use the test data from previous years to train the systems.

Each message in both the training and test set is tagged with its global polarity, indicating whether the text expresses a positive, negative or neutral sentiment, or no sentiment at all. A set of 6 labels has been defined: strong positive (P+), positive (P), neutral (NEU), negative (N), strong negative (N+) and one additional no sentiment tag (NONE).

In addition, there is also an indication of the level of agreement or disagreement of the expressed sentiment within the content, with two possible values: AGREEMENT and DISAGREEMENT. This is especially useful to make out whether a neutral sentiment comes from neutral keywords or else the text contains positive and negative sentiments at the same time.

Moreover, the polarity at entity level, i.e., the polarity values related to the entities that are mentioned in the text, is also included for those cases when applicable. These values are similarly tagged with 6 possible values and include the level of agreement as related to each entity.

On the other hand, a selection of a set of topics has been made based on the thematic areas covered by the corpus, such as *politics*, *soccer*, *literature* or *entertainment*. Each message in both the training and test set has been assigned to one or several of these topics (most messages are associated to just one topic, due to the short length of the text).

All tagging has been done semi automatically: a baseline machine learning model is first run and then all tags are manually checked by human experts. In the case of the polarity at entity level, due to the high volume of data to check, this tagging has just been done for the training set.

Table 1 shows a summary of the training and test corpora provided to participants.

Attribute	Value
Tweets	68 017
Tweets (test)	60 798 (89%)
Tweets (test)	7 219 (11%)
Topics	10
Tweet languages	1
Users	154
User types	3
User languages	1
Date start (train)	2011-12-02 T00:47:55
Date end (train)	2012-04-10 T23:40:36
Date start (test)	2011-12-02 T00:03:32
Date end (test)	2012-04-10 T23:47:55

Table 1: Corpus statistics

Users were journalists (*periodistas*), politicians (*políticos)* or celebrities (*famosos*). The only language involved this year was Spanish (*es)*.

The list of topics that have been selected is shown in Table 2.

Topic
Politics *(política)*
Other *(otros)*
Entertainment *(entretenimiento)*
Economy *(economía)*
Music *(música)*
Soccer *(fútbol)*
Films *(cine)*
Technology *(tecnología)*
Sports *(deportes)*
Literature *(literatura)*

Table 2: Topic list

The corpus is encoded in XML (the XSD schema is provided for validation). Figure 1 shows the information of two sample tweets. The first tweet is only tagged with the global polarity as the text contains no mentions to any entity, but the second one is tagged with both the global polarity of the message and the polarity associated to each of the entities that appear in the text (UPyD and Foro Asturias).

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

```
<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 Social-TV corpus

This corpus was collected during the 2014 Final of Copa del Rey championship in Spain between Real Madrid and F.C. Barcelona, played on 16 April 2014 at Mestalla Stadium in Valencia.

Over 1 million tweets were collected from 15 minutes before to 15 minutes after the match. After filtering useless information, tweets in other languages than Spanish, a subset of 2 773 was selected.

All tweets have been manually tagged with the aspects of the expressed messages and its sentiment polarity. Tweets may cover more than one aspect.

The list of the 31 aspects that have been defined is shown in Table 3.

Aspect
Afición
Arbitro
Autoridades
Entrenador
Equipo - Atlético de Madrid
Equipo - Barcelona
Equipo - Real Madrid
Equipo *(any other team)*
Jugador - Alexis Sánchez
Jugador - Alvaro Arbeloa
Jugador - Andrés Iniesta
Jugador - Angel Di María
Jugador - Asier Ilarramendi
Jugador - Carles Puyol
Jugador - Cesc Fábregas
Jugador - Cristiano Ronaldo
Jugador - Dani Alves
Jugador - Dani Carvajal
Jugador - Fábio Coentrão
Jugador - Gareth Bale
Jugador - Iker Casillas
Jugador - Isco
Jugador - Javier Mascherano
Jugador - Jesé Rodríguez
Jugador - José Manuel Pinto
Jugador - Karim Benzema
Jugador - Lionel Messi
Jugador - Luka Modric
Jugador - Marc Bartra
Jugador - Neymar Jr.
Jugador - Pedro Rodríguez
Jugador - Pepe
Jugador - Sergio Busquets
Jugador - Sergio Ramos
Jugador - Xabi Alonso
Jugador - Xavi Hernández
Jugador *(any other player)*
Partido
Retransmisión

Table 3: Aspect list.

Sentiment polarity has been tagged from the point of view of the person who writes the tweet, using 3 levels: P, NEU and N. No distinction is made in cases when the author does not express any sentiment or when he/she expresses a no-positive no-negative sentiment.

The Social-TV corpus has been randomly divided into two sets: training (1 773 tweets) and test (1 000 tweets), with a similar distribution of both aspects and sentiments. The training set will be released so that participants may train and validate their models. The test corpus will be provided without any tagging and will be used to evaluate the results provided by the different systems.

The following figure shows the information of three sample tweets in the training set.

```
<tweet id="456544898791907328">
  <sentiment aspect="Equipo-Real_Madrid" polarity="P">#HalaMadrid
  </sentiment> ganamos sin <sentiment aspect="Jugador-Cristiano_Ronaldo"
   polarity="NEU">Cristiano</sentiment>. .perdéis con <sentiment aspect=
  "Jugador-Lionel_Messi" polarity="N">Messi</sentiment>. Hala <sentiment
   aspect="Equipo-Real_Madrid" polarity="P">Madrid</sentiment>! !!!!!
</tweet>
<tweet id="456544898942906369">
  @nevermind2192 <sentiment aspect="Equipo-Barcelona" polarity="P">Barça
  </sentiment> por siempre!!
</tweet>
<tweet id="456544898951282688">
  <sentiment aspect="Partido" polarity="NEU">#FinalCopa</sentiment>
  Hala <sentiment aspect="Equipo-Real_Madrid" polarity="P">Madrid
  </sentiment>, hala <sentiment aspect="Equipo-Real_Madrid" polarity="P"
  >Madrid</sentiment>, campeón de la <sentiment aspect="Partido"
  polarity="P">copa del rey</sentiment>
</tweet>
```

Figure 2: Sample tweets (Social-TV corpus)

Both corpus will be made freely available to the community after the workshop. Please send an email to tass@daedalus.es filling in the TASS Corpus License agreement with your email, affiliation (institution, company or any kind of organization) and a brief description of your research objectives, and you will be given a password to download the files in the password protected area. The only requirement is to include a citation to a relevant paper and/or the TASS website.

3 Description of tasks

This edition of TASS has two objectives. First of all, we are interested in evaluating the evolution of the different approaches for sentiment analysis and text classification in Spanish during these years. So, two legacy tasks will be repeated again, reusing the same corpus, to compare results.

Moreover, we want to foster the research in the analysis of fine-grained polarity, i.e., more specific than the global polarity of the text. So two new tasks are proposed related to polarity detection and analysis at aspect level (aspect-based sentiment analysis), one of the new requirements of the market of natural language processing in these areas.

Thus the following four tasks are proposed this year.

Participants are expected to submit up to 3 results of different experiments for one or several of these tasks, in the appropriate format described next.

Along with the submission of experiments, participants were invited to submit a paper to the workshop in order to describe their experiments and discussing the results with the audience in a regular workshop session.

These papers should follow the usual SEPLN template (as given in the author guidelines page). Reports can be written in Spanish or English. All papers were reviewed by the program committee and are included in the proceedings of the workshop.

The four proposed tasks are described next.

3.1 (legacy) Task 1: Sentiment Analysis at Global Level

This task consists on performing an automatic sentiment analysis to determine the global polarity of each message in the test set of the General corpus (see below). This task is a reedition of the task in the previous years. Participants will be provided with the training set of the General corpus so that they may train and validate their models.

There will be two different evaluations: one based on 6 different polarity labels (`P+`, `P`, `NEU`, `N`, `N+`, `NONE`) and another based on just 4 labels (`P`, `N`, `NEU`, `NONE`).

Participants are expected to submit (up to 3) experiments for the 6-labels evaluation, but are also allowed to submit (up to 3) specific experiments for the 4-labels scenario.

Accuracy (correct tweets according to the gold standard) will be used for ranking the systems. Precision, recall and F1-measure will be used to evaluate each individual category.

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

```
tweetid \t polarity
```

where polarity can be:

- `P+`, `P`, `NEU`, `N`, `N+` and `NONE` for the 6-labels case
- `P`, `NEU`, `N` and `NONE` for the 4-labels case.

The same test corpus of previous years will be used for the evaluation, to allow for comparison among systems. Obviously, participants are not allowed to use any test data to train their systems. However, to deal with the problem reported last years of the imbalanced distribution of labels between the training and test set, a new selected test subset containing 1000 tweets with a similar distribution to the training corpus will be extracted and used for an alternate evaluation of the performance of systems.

3.2 (legacy) Task 2: Topic Classification

The challenge is to build a classifier to automatically identify the topic of each message

in the test set of the General corpus. Again, a reedition of the same task in previous years. Participants may use the training set of the General corpus to train and validate their models.

The task is a multi-label classification and tweets can have more than one label.

Participants are expected to submit up to 3 experiments, each one in a plain text file with the following format:

```
tweetid \t topic
```

A given tweet ID can be repeated in different lines if it is assigned more than one topic.

Microaveraged precision, recall and F1-measure calculated over the full test set will be used to evaluate the systems. Systems will be ranked by F1.

To allow for comparison with previous years, the same test corpus will be used for the evaluation. Again, participants are not allowed to use any test data to train their systems.

3.3 (new) Task 3: Aspect Detection

The objective is the automatic identification of the different aspects expressed by users, among a predefined list, in their opinions in Twitter about a given topic. A new Social-TV corpus will be used for the training and evaluation of the systems (see description below).

This is a multi-label classification and tweets can have more than one aspect.

Participants are expected to submit up to 3 experiments, each in a plain text file with the following format:

```
tweetid \t aspect
```

Again, a given tweet ID can be repeated in different lines if it is assigned more than one aspect. Just the aspect must be returned, not the detected terms or fragment in the text nor its offsets.

Microaveraged precision, recall and F1-measure calculated over the full test set will be used to evaluate the systems. Systems will be ranked by F1.

3.4 (new) Task 4: Aspect-based Sentiment Analysis

Systems in this task must identify the polarity of the aspect that was detected in the previous task. Again. participants will be provided with the Social-TV corpus to train and evaluate their models. This task is equivalent to Task 1 but focused on fine-grained polarity detection.

Participants are expected to submit up to 3 experiments, each in a plain text file with the following format:

```
tweetid \t aspect \t polarity
```

Allowed polarity values are `P`, `NEU` and `N`.

Microaveraged precision, recall and F1-measure will be used to evaluate the systems, considering a unique label combining aspect-polarity. Systems will be ranked by F1.

4 Participants and Results

This year 35 groups registered (as compared to 31 groups last year) but unfortunately only 7 groups (14 last year) sent their submissions. The list of active participant groups is shown in Table 4, including the tasks in which they have participated.

Group	1	2	3	4
LyS	X	X	X	X
SINAI-ESMA	X			
Elhuyar	X			
SINAIword2vec	X			
JRC	X			
ELiRF-UPV	X	X	X	X
IPN	X	X		
Total groups	7	3	2	2

Table 4: Participant groups

All participants groups sent a report describing their experiments and results achieved. Papers were reviewed and included in the workshop proceedings. References are listed in Table 5.

Group	Report
LyS	(Vilares et al., 2014)
SINAI-ESMA	(Jiménez Zafra *et al.*, 2014)
Elhuyar	(San Vicente and Saralegi, 2014)
SINAIword2vec	(Montejo-Ráez *et al.*, 2014)
JRC	(Perea-Ortega and Balahur, 2014)
ELiRF-UPV	(Hurtado and Pla, 2014)
IPN	(Hernández Petlachi and Li, 2014)

Table 5: Participant reports

5 Results

After the submission deadline, runs were collected and checked and results were evaluated and made available to the participants using an automated web page in the password protected area in the website.

Results included a spreadsheet with the overall values for each task, including the defined metrics, and also detailed results per experiment for all the evaluations The PHP script used for the evaluation of each submission was also included.

Results for each task are described next.

5.1 Task 1: Sentiment Analysis at Global Level

Submitted runs and results for Task 1, evaluation based on 5 polarity levels with the whole General test corpus, are shown in Table 6. Accuracy (correct tweets according to the gold standard) is used for ranking the systems. Precision, recall and F1-measure could be used to evaluate each individual label, though it is not presented here.

Run Id	Acc
ELiRF-UPV-run3	0.64
ELiRF-UPV-run1	0.63
ELiRF-UPV-run2	0.63
Elhuyar-Run1	0.61
Elhuyar-Run3	0.61
Elhuyar-Run2	0.61
LyS-1	0.58
LyS-2	0.56
SINAIword2vec-1	0.51
SINAI-ESMA-1	0.51
SINAI-ESMA-without_negation	0.51
JRC-run1-ER	0.48
JRC-run2-RPSN-ER-AWM-4-all-2-skipbigrams	0.48
JRC-run3-baseline-stop	0.48
IPN-Linguistic_2	0.37
IPN-1	0.37

Table 6: Results for Task 1, 5 levels, whole test corpus

As previously described, an alternate evaluation of the performance of systems was done using a new selected test subset containing 1000 tweets with a similar distribution to the training corpus. Results are shown in next Table 7. Figures are much lower as compared to the previous evaluation, thus showing a high bias in the semi automatic tagging of the whole test corpus.

Run Id	Acc
ELiRF-UPV-run1-1k	0.48
ELiRF-UPV-run3-1k	0.48
Elhuyar-Run2-1k	0.47
Elhuyar-Run3-1k	0.47
ELiRF-UPV-run2-1k	0.47
Elhuyar-Run1-1k	0.47
SINAIword2vec-1-1k	0.46
LyS-2-1k	0.46
LyS-1-1k	0.45
JRC-run3-baseline-stop-1k	0.42
JRC-run1-ER-1k	0.41
JRC-run2-RPSN-ER-AWM-4-all-2-skipbigrams-1k	0.40
SINAI-ESMA-without_negation-1k	0.37
SINAI-ESMA-1-1k	0.37
IPN-Linguistic_2-1k	0.35
IPN-1-1k	0.33

Table 7: Results for Task 1, 5 levels, selected 1k corpus

In order to perform a more in-depth evaluation, results are calculated considering the classification only in 3 levels (`POS`, `NEU`, `NEG`) and no sentiment (`NONE`) merging `P` and `P+` in only one category, as well as `N` and `N+` in another one. The same double evaluation using the whole test corpus and a new selected corpus have been carried out, shown in next tables.

Run Id	Acc
ELiRF-UPV-run2	0.71
ELiRF-UPV-run1	0.71
ELiRF-UPV-run3	0.70
Elhuyar-Run1	0.70
Elhuyar-Run2	0.70
Elhuyar-Run3	0.70
LyS-1	0.67
LyS-2	0.67
SINAIword2vec-2	0.61
JRC-run2-RPSN-ER-AWM-4-all-2-skipbigrams	0.61

SINAI-ESMA-1	0.61
JRC-run1-ER	0.61
SINAI-ESMA-without_negation	0.60
JRC-run3-baseline-stop	0.60
SINAIword2vec-1	0.59
IPN-Linguistic_2	0.55

Table 8: Results for Task 1, 3 levels, whole test corpus

Run Id	Acc
ELiRF-UPV-run3-1k	0.66
ELiRF-UPV-run1-1k	0.65
LyS-2-1k	0.64
Elhuyar-Run3-1k	0.64
SINAIword2vec-2-1k	0.63
Elhuyar-Run2-1k	0.63
LyS-1-1k	0.63
Elhuyar-Run1-1k	0.62
SINAIword2vec-1-1k	0.61
ELiRF-UPV-run2-1k	0.60
JRC-run1-ER-1k	0.56
JRC-run3-baseline-stop-1k	0.56
JRC-run2-RPSN-ER-AWM-4-all-2-skipbigrams-1k	0.55
SINAI-ESMA-1-1k	0.52
SINAI-ESMA-without_negation-1k	0.52
IPN-Linguistic_2-1k	0.52

Table 9: Results for Task 1, 3 levels, selected 1k corpus

5.2 Task 2: Topic Classification

Microaveraged precision, recall and F1-measure calculated over the full test set are used to evaluate the systems. Systems are ranked by F1. Table 10 shows the results for Task 2.

Run Id	P	R	F1
ELiRF-UPV-run3	0.67	0.75	0.70
ELiRF-UPV-run2	0.70	0.71	0.70
ELiRF-UPV-run1	0.68	0.69	0.69
LyS-1	0.68	0.60	0.64
LyS-2	0.68	0.59	0.63
IPN-2	0.27	0.33	0.30

Table 10: Results for Task 2

5.3 Task 3: Aspect Detection

Microaveraged precision, recall and F1-measure calculated over the full test set are used to evaluate the systems. Systems are ranked by F1. Results for Task 3 are shown in next table.

Run Id	P	R	F1
ELiRF-UPV-run1	0.91	0.91	0.91
LyS-1	0.81	0.90	0.85

Table 11: Results for Task 3

5.4 Task 4: Aspect-based Sentiment Analysis

Last, results for Task 4 are shown in Table 12. Microaveraged Precision, recall and F1-measure are used to evaluate the systems, considering a unique label combining aspect-polarity, and systems are ranked by F1.

Run Id	P	R	F1
ELiRF-UPV-run2	0.58	0.60	0.59
ELiRF-UPV-run1	0.57	0.59	0.58
ELiRF-UPV-run3	0.56	0.58	0.57
LyS-2	0.52	0.58	0.55
LyS-1	0.51	0.57	0.54
LyS-3	0.46	0.51	0.48

Table 12: Results for Task 4

6 Conclusions and Future Work

TASS was the first workshop about sentiment analysis in the context of SEPLN. Clearly this area receives great attraction from research groups and companies, as this third edition of TASS has had a greater impact in terms of registered groups. However, on the other hand, the number of participants that submitted experiments this year for the tasks has decreased even with respect to the first edition. Perhaps the schedule of the tasks has to be rearranged and probably the dissemination activities have to be greatly improved.

Anyway, the developed corpus and gold standards, and the reports from participants will for sure be helpful for other research groups approaching these tasks.

TASS corpora will be released after the workshop for free use by the research

community. Last year's corpora have been downloaded up to date by more than 70 research groups, 23 out of Spain, by groups coming from academia and also from private companies to use the corpus as part of their product development. We expect to reach a similar impact with this year's corpus.

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), *MULTIMEDICA: Multilingual Information Extraction in Health domain and application to scientific and informative documents* (TIN2010-20644-C03-01), ATTOS project (TIN2012-38536-C03-0) from the Spanish Government, and AORESCU project (P11-TIC-7684 MO) from the regional government of Junta de Andalucía.

References

Villena-Román, Julio; Lana-Serrano, Sara; Martínez-Cámara, Eugenio; González-Cristobal, José Carlos. 2013. *TASS - Workshop on Sentiment Analysis at SEPLN*. Revista de Procesamiento del Lenguaje Natural, 50, pp 37-44. http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/4657.

Villena-Román, Julio; García-Morera, Janine; Lana-Serrano, Sara; González-Cristobal, José Carlos. 2014. *TASS 2013 - A Second Step in Reputation Analysis in Spanish*. Revista de Procesamiento del Lenguaje Natural, 52, pp 37-44. http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/4901.

Vilares, David; Doval, Yerai; Alonso, Miguel A.; Gómez-Rodríguez, Carlos. *LyS at TASS 2014: A Prototype for Extracting and Analysing Aspects from Spanish tweets*. In Proc. of the TASS workshop at SEPLN 2014. 16-19 September 2014, Girona, Spain.

Perea-Ortega, José M. Balahur, Alexandra. *Experiments on feature replacements for polarity classification of Spanish tweets*. In Proc. of the TASS workshop at SEPLN 2014. 16-19 September 2014, Girona, Spain.

Hernández Petlachi, Roberto; Li, Xiaoou. *Análisis de sentimiento sobre textos en Español basado en aproximaciones semánticas con reglas lingüísticas*. In Proc. of the TASS workshop at SEPLN 2014. 16-19 September 2014, Girona, Spain.

Montejo-Ráez, A.; García-Cumbreras, M.A.; Díaz-Galiano, M.C. *Participación de SINAI Word2Vec en TASS 2014*. In Proc. of the TASS workshop at SEPLN 2014. 16-19 September 2014, Girona, Spain.

Hurtado, Lluís F.; Pla, Ferran. *ELiRF-UPV en TASS 2014: Análisis de Sentimientos, Detección de Tópicos y Análisis de Sentimientos de Aspectos en Twitter*. In Proc. of the TASS workshop at SEPLN 2014. 16-19 September 2014, Girona, Spain.

Jiménez Zafra, Salud María; Martínez Cámara, Eugenio; Martín Valdivia, M. Teresa.; Ureña López, L. Alfonso. *SINAI-ESMA: An unsupervised approach for Sentiment Analysis in Twitter*. In Proc. of the TASS workshop at SEPLN 2014. 16-19 September 2014, Girona, Spain.

San Vicente Roncal, Iñaki; Saralegi Urizar, Xabier. *Looking for Features for Supervised Tweet Polarity Classification*. In Proc. of the TASS workshop at SEPLN 2014. 16-19 September 2014, Girona, Spain.