Procesamiento del Lenguaje Natural, Revista nº 50 marzo de 2013, pp 37-44 recibido 28-11-12 revisado 14-02-13 aceptado 19-02-13

TASS - Workshop on Sentiment Analysis at SEPLN

TASS - Taller de Análisis de Sentimientos en la SEPLN

Julio Villena-Román
DAEDALUS
jvillena@daedalus.es

Sara Lana-Serrano
DIATEL - Universidad Politécnica de Madrid
slana@diatel.upm.es

Eugenio Martínez-Cámara
SINAI - Universidad de Jaén
emcamara@ujaen.es

José Carlos González-Cristóbal
GSI - Universidad Politécnica de Madrid
jgonzalez@dit.upm.es

Resumen: Este artículo describe el desarrollo de TASS, taller de evaluación experimental en el contexto de la SEPLN para fomentar la investigación en el campo del análisis de sentimiento 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 complejos sistemas capaces de realizar un análisis de sentimientos basados en opiniones de textos cortos extraídos de medios sociales (concretamente Twitter). Este artículo describe las tareas propuestas, el contenido, formato y las estadísticas más importantes del corpus generado, los participantes y los diferentes enfoques planteados, así como los resultados generales obtenidos.
Palabras clave: TASS, análisis de reputación, análisis de sentimientos, medios sociales.

Abstract: This paper describes TASS, 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) published by representative personalities. The paper presents the proposed tasks, the contents, format and main statistics of the generated corpus, the participant groups and their different approaches, and, finally, the overall results achieved.
Keywords: TASS, reputation analysis, sentiment analysis, social media.

1 Introduction

According to Merriam-Webster dictionary,[1] **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.

For the economic implications, reputation is especially important in business, where it refers to the perception or attitudes that customers, stakeholders, employees, competitors and any other agent exhibit about that organization. Reputation includes aspects such as customer satisfaction about the company's product and services, commitment and loyalty from employees, partners' trust on agreements and obligations, support from investors, etc.

In turn, **reputation analysis** is the process of tracking, investigating and reporting other entities' opinions about the entity's actions. It covers many factors to calculate the market value of reputation. Reputation analysis can be used by companies as a tool to improve competitiveness in the complex marketplace of 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, etc., has led to

[1] http://www.merriam-webster.com/

ISSN 1135-5948

creation of an emerging trend towards the use of online reputation analysis.

The so-called **sentiment analysis**, i.e., the application of natural language processing and text analytics to identify and extract subjective information from texts, which is the first step towards online reputation analysis, is becoming a promising topic in the field of 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 about their customers' sentiments towards their products.

Sentiment analysis is a major technological challenge. The task is so hard that even humans often disagree on the categorization on the positive or negative sentiment that is supposed to be expressed on a given text, either in a specific segment of the text or as a global property of the full text. The fact that issues that one individual finds acceptable may not be the same to others, along with multilingual aspects, cultural factors and different contexts make it very hard to categorize a text written in a natural language into a positive or negative sentiment, even with a training based on a given user model and a context for analysis. And the shorter the text is, for example, when analyzing Twitter messages or short comments in Facebook, the harder the task becomes.

Within this context, TASS,[2] which stands for *Taller de Análisis de Sentimientos en la SEPLN* (*Workshop on Sentiment Analysis at SEPLN*, in English) is an experimental evaluation workshop, organized as a satellite event of the SEPLN 2012 Conference, held on September 7th, 2012 in Jaume I University at Castellón de la Plana, Comunidad Valenciana, Spain, to promote the research in the field of sentiment analysis in social media, initially focused on Spanish though it could be extended to any language.

The main objective was to encourage participants to improve the existing techniques and algorithms and even design new ones in order to perform a sentiment analysis in short text opinions extracted from social media messages (specifically Twitter) published by a series of important personalities. Moreover, the sentiment extraction is complemented with a text categorization, thus researching on the whole process of reputation analysis.

[2] http://www.daedalus.es/TASS

The challenge task is intended to provide a benchmark forum for comparing the latest approaches in this field. 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.

2 Description of tasks

Two tasks were proposed for the participants in this first edition: a first task focused on **sentiment analysis** and a second task about text categorization, which was called **trending topic coverage**. Groups could participate in both tasks or just in one of them.

Along with the submission of the results of their experiments, participants were encouraged to submit a paper to the workshop in order to describe their systems to the audience in a regular workshop session. Submitted papers were reviewed by the program committee.

2.1 Task 1: Sentiment Analysis

This task consists on performing an automatic sentiment analysis to determine the polarity of each message in the test corpus.

The evaluation metrics to evaluate and compare the different systems are the usual measurements of precision (1), recall (2) and F-measure (3) calculated over the full test set.

$$\text{precision} = \frac{\text{N(correct classifications)}}{\text{N(all classifications)}} \quad (1)$$

$$\text{recall} = \frac{\text{N(retrieved documents)}}{\text{N(all documents)}} \quad (2)$$

$$F(\beta) = (1+\beta^2)\frac{\text{precision} \cdot \text{recall}}{\beta^2 \cdot \text{precision} + \text{recall}} \quad (3)$$

2.2 Task 2: Trending topic coverage

In this case, the technological challenge is to build a classifier to identify the topic of the text, and then apply the polarity analysis to get the assessment for each topic.

The evaluation metrics are the same as in Task 1 (precision, recall and F-measure).

3 Corpus

The corpus provided to participants contains over 70,000 tweets, written in Spanish by nearly 200 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 origin 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.

Due to restrictions in the Twitter API Terms of Service,[3] 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 Tweet ID. In addition, using the user ID, it is possible to extract information about the user's name, registration date, geographical information of their location, and many other fields, which may allow to perform experiments for instance on the different varieties of Spanish.

Thus each Twitter message includes its Tweet ID (`twitid`), the user ID (`user`) and the creation date (`date`). Each message is annotated with its global polarity for sentiment, i.e., an indication of whether the text expresses a positive, negative or neutral sentiment, or no sentiment at all. 5 polarity levels have been defined: *strong positive* (`P+`), *positive* (`P`), *neutral* (`NEU`), negative (`N`), strong negative (`N+`) and one additional *no sentiment* label (`NONE`).

Moreover, in those cases where applicable, this same polarity levels are annotated but related to the entities that are mentioned in the text. This is done when the sentiment is referring to the identified entity but not in cases where an entity appears in the text but is not involved in the expressed sentiment.

There is also an indication of the level of *agreement* or *disagreement* of the expressed sentiment within the content. This is especially useful to make out whether a neutral sentiment comes from a neutral set of keywords (i.e., slightly positive or negative) or else the text contains positive and negative sentiments at the same time. For example, "*Peter is a very good friend but I cannot stand John*" could be considered `NEU` with `DISAGREEMENT` where *Peter* is regarded as `P+` and *John* as `N+`.

On the other hand, a selection of a set of 10 topics has been made based on the thematic areas covered by the corpus, such as *politics* (*política*), *soccer* (*fútbol*), *literature* (*literatura*) or *entertainment* (*entretenimiento*).

Each message of the corpus has been semiautomatically assigned to one or several of these topics. A baseline model (Villena-Román *et al.*, 2011) was used to obtain the candidate topics that then were manually revised.

This tagged corpus has been divided into two sets: training and test. The *training corpus* was released along with the corresponding tags so that participants may train and validate their models for classification and sentiment analysis. The *test corpus* was provided without any tag and was used to evaluate the results provided by the different systems.

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

Attribute	Train corpus	Test corpus
Tweets	7 219	60 798
Topics	10	10
Tweet languages	1	1
Users	154	158
User types	3	3
User languages	1	1
Date start	2011-12-02 T00:47:55	2011-12-02 T00:03:32
Date end	2012-04-10 T23:40:36	2012-04-10 T23:47:55

Table 1: Train and test corpus

There were 3 user types: journalists (*periodistas*), politicians (*políticos)* or celebrities (*famosos*). The only language involved this year was Spanish (*es)*.

The list of selected topics is shown in Table 2, sorted by frequency in the test corpus.

Topic	Frequency
Politics (*política*)	3 119
Other (*otros*)	2 337
Entertainment (*entretenimiento*)	1 677
Economy (*economía*)	942
Music (*música*)	566
Soccer (*fútbol*)	252
Films (*cine*)	245
Technology (*tecnología*)	217
Sports (*deportes*)	113
Literature (*literatura*)	99

Table 2: Topic list

The corpus is encoded in XML in which the text of the content entity has been removed to

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

follow the Twitter restrictions. Two sample tweets are shown in Figure 1. The second one is tagged with both the global polarity of the message and the polarity associated to each one of the entities that appears in the text ("*UPyD*" and "*Foro Asturias*"), whereas the first tweet is only tagged with the global polarity as the text contains no mentions to any entity.

```
<twit>
    <twitid>0000000000</twitid>
    <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>
</twit>

<twit>
    <twitid>0000000001</twitid>
    <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>politica</topic>
    </topics>
</twit>
```

Figure 1: Sample tweet messages

The full corpus was made public after the workshop[4] so that any group interested in the field of sentiment analysis in Spanish could make use of it in their own research.

4 Participants

Participants were required to register for the task(s) in order to obtain the corpus.

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

```
twitid \t polarity \t topic
```

Where `twitid` is the Tweet ID for every message in the test corpus, the `polarity` contains one of the 6 valid tags (`P+`, `P`, `NEU`, `N`, `N+` and `NONE`), and the same for `topic`.

Although the polarity level should be classified into those 5 levels and results were primarily evaluated for them, the evaluation also included metrics with just 3 levels (*positive*, *neutral* and *negative*).

Participants could submit results for one or both tasks. Several results for the same task were allowed too.

15 groups registered and finally 8 groups sent their submissions for one of the two tasks. The list of active groups is shown in Table 3. All of them submitted results for the sentiment analysis tasks and most of them (6 out of 8, 75%) participated in both tasks.

Group	Task 1	Task 2
Elhuyar Fundazioa	Yes	No
IMDEA	Yes	Yes
L2F - INESC	Yes	Yes
La Salle - URL	Yes	Yes
LSI UNED	Yes	Yes
LSI UNED 2	Yes	Yes
SINAI - UJAEN	Yes	Yes
UMA	Yes	No

Table 3: Participant groups

There was another group at Delft University of Technology that submitted experiments for both tasks but finally did not submit a report for the workshop, so their results are not included.

The next sections briefly describe the approaches for the different groups.

4.1 Elhuyar Fundazioa

In their paper, Saralegi and San Vicente (2012) describe their supervised approach that includes some linguistic knowledge-based processing for preparing the features.

The processing comprises lemmatization, part-of-speech tagging, tagging of polarity words, treatment of emoticons, negation, and weighting of polarity words depending on syntactic nesting level. A pre-processing step for handling spelling errors was also performed.

Detection of polarity words is done according to a polarity lexicon built in two ways: projection to Spanish of an English lexicon, and extraction of divergent words of positive and negative tweets of training corpus.

Evaluation results show a good performance and also good robustness of the system both for

[4] http://www.daedalus.es/TASS/corpus.php

the fine granularity (65% of accuracy) as well as for coarse granularity polarity detection (71% of accuracy).

4.2 IMDEA

The IMDEA (Instituto Madrileño de Estudios Avanzados) team state that sentiment analysis and topic detection are new problems that are at the intersection of natural language processing and data mining (Fernandez Anta *et al.*, 2012).

An interesting comparative analysis of different approaches and classification techniques for these problems is presented.

The data is preprocessed using well-known techniques and tools proposed in the literature, together with others specifically proposed here that take into account the characteristics of Twitter. Then, popular classifiers have been used, in particular, most popular classifiers of WEKA (Hall *et al.*, 2009) were evaluated. Their report describes some of the results obtained in their preliminary research.

4.3 L2F – INESC

The strategy used by the L2F (Laboratório de sistemas de Língua Falada) team at INESC (Instituto de Engenharia de Sistemas e Computadores) for performing automatic sentiment analysis and topic classification over Spanish Twitter data is described by Batista and Ribeiro (2012).

They have decided to consider both tasks as classification tasks, thus sharing the same method. Their most successful and recent experiments in this field cast the problem as a binary classification problem, which aims at discriminating between two possible classes. Binary decisions are stated as "document matches/does not match category A", and a binary classifier exists for each polarity level and topic. Binary classifiers are easier to develop, offer faster convergence ratios, and can be executed in parallel. The final results are then generated by combining all the different binary classifiers.

Specifically, they have adopted an approach based on logistic regression classification models, which corresponds to the maximum entropy classification for independent events.

As described in their paper, the L2F system achieved the best results for the topic classification contest, and the second place in terms of sentiment analysis.

4.4 La Salle – Universitat Ramon Llull

Trilla and Alías (2012) describe how they adapt a text classification scheme based on multinomial naive Bayes. The multinomial naive Bayes is a probabilistic generative approach that builds a language model assuming conditional independence among the linguistic features. Therefore, no sense of history, sequence nor order is introduced in this model. This approach achieves a good result in terms of the evaluation metrics.

4.5 LSI – UNED

Martín-Wanton and Carrillo de Albornoz, (2012) present the participation of the LSI (Lenguajes y Sistemas Informáticos) group at UNED (Universidad Nacional de Educación a Distancia) in TASS. For polarity classification, they propose an emotional concept-based method. The original method makes use of an affective lexicon to represent the text as the set of emotional meanings it expresses, along with advanced syntactic techniques to identify negations and intensifiers, their scope and their effect on the emotions affected by them.

Besides, the method addresses the problem of word ambiguity, taking into account the contextual meaning of terms by using a word sense disambiguation algorithm. On the other hand, for topic detection, their system is based on a probabilistic model called Twitter-LDA, based on Latent Dirichlet Allocation technique. They first build for each topic of the task a lexicon of words that best describe it, thus representing each topic as a ranking of discriminative words. Moreover, a set of events is retrieved based on a probabilistic approach adapted to the characteristics of Twitter.

To determine which of the topics corresponds to each event, the topic with the highest statistical correlation was obtained comparing the ranking of words of each topic and the ranking of words most likely to belong to the event.

The experimental results achieved show the adequacy of their approach for the task.

4.6 LSI – UNED 2

Castellano, Cigarrán and García-Serrano (2012a) describe the research done for the workshop by the second team component of the second group from LSI at UNED.

Their proposal addresses the sentiment and topic detection from an information retrieval perspective, based on language divergences. Kullback-Liebler divergence (computed against the testing corpus) is used to generate both, polarity and topic models, which will be used in the information retrieval process (Castellano, Cigarrán and García-Serrano, 2012b).

In order to improve the accuracy of the results, they propose several approaches focused on carry out language models, not only considering the textual content associated to each tweet but, as an alternative, the named entities or adjectives detected as well.

Results show that modeling the tweets set using named entities and adjectives improves the final precision results and, as a consequence, their representativeness in the model compared with the use of common terms.

General results are promising (fifth and fourth position in each of the proposed tasks), indicating that an IR and language models based approach may be an alternative to other classical proposals focused on the application of classification techniques.

4.7 SINAI – Universidad de Jaén

The participation of the SINAI (Sistemas Inteligentes de Acceso a la Información) research group of the University of Jaén is described by Martínez Cámara *et al.* (2012).

For the first task, they have chosen a supervised machine learning approach, in which they have used support vector machines (SVM) for classifying the polarity. Text features included are unigrams, emoticons, positive and negative words and intensity markers.

In the second task, they have also used SVM for the topic classification but several bags of words (BoW) have been used with the goal of improving the classification performance.

One BoW has been obtained using Google Adwords Keyword Tool,[5] which allows to enter a term and directly returns the top *N* related concepts. The second BoW has generated based on the hash tags of the training tweets, per each category.

4.8 Universidad de Málaga

Moreno-Ortiz and Pérez-Hernández (2012) describe the participation of the group at Facultad de Filosofía y Letras in Universidad de Málaga (UMA). They use a lexicon-based approach to sentiment analysis. These approaches differ from the more common machine-learning based approaches in that the former rely solely on previously generated lexical resources that store polarity information for lexical items, which are then identified in the texts, assigned a polarity tag, and finally weighed, to come up with an overall score for the text.

Such systems have been proved to perform on par with supervised, statistical systems, with the added benefit of not requiring a training set. However, it remains to be seen whether such lexically-motivated systems can cope equally well with extremely short texts, as generated on social networking sites, such as Twitter.

In their paper they perform such an evaluation using Sentitext, a lexicon-based sentiment analysis tool for Spanish. One conclusion is that Sentitext's Global Sentiment Value is strongly affected by the number of lexical units available in the text (or the lack of them, rather). On the other hand, they also confirm Sentitext's tendency to assign middle-of-the-scale ratings, or at least avoid extreme values, which is reflected on its poor performance for the `N+` and `P+` classes, most of which were assigned to the more neutral `N` and `P` classes.

Another interesting conclusion which is drawn from their analysis of the average number of polarity lexical segments and Sentitext's Affect Intensity (an internal measure similar to the polarity level) is that Twitter users employ highly emotional language.

5 Results

The gold standard has been generated by first pooling all submissions, then a voting scheme has been applied and finally an extensive human review of the ambiguous decisions (thousands of them). Due to the high volume of data, this was the only way to generate a tagged set; unfortunately, this is subject to errors and misclassifications. Obviously, if all annotators are consistently wrong, the gold standard will end up with a wrong label, and accuracy figures will then be an upper bound of actual accuracy.

A manual evaluation of a part of the gold standard to assess its quality has not been done yet, although this task is planned for future editions of the workshop.

[5] https://adwords.google.com/o/KeywordTool

Both tasks have been evaluated as a single label classification. This specifically affects to the topic classification, where the most restrictive criterion has been applied: a "success" is achieved only when all the test labels have been returned. Participants were welcome to discuss and reevaluate their experiments with a less restrictive strategy in their papers.

Regarding Task 1, 17 different experiments were submitted. Results are listed in the tables below. All tables show the precision value achieved in each experiment (recall and F-measure were also evaluated and provided to participants but are omitted here).

Table 4 considers 5 levels of sentiments (P+, P, NEU, N, N+) and no sentiment (NONE).

Precision values range from 65.3% to 16.7%. Only 8 from 20 submissions achieve figures higher than 50% and specifically 5 of the 9 groups have at least one submission above this value. Besides, results for different submissions from the same group are typically very similar except for the SINAI group.

Run Id	Group	Precision
pol-elhuyar-1-5l	Elhuyar Fund.	65.3%
pol-l2f-1-5l	L2F - INESC	63.4%
pol-l2f-3-5l	L2F - INESC	63.3%
pol-l2f-2-5l	L2F - INESC	62.2%
pol-atrilla-1-5l	La Salle - URL	57.0%
pol-sinai-4-5l	SINAI - UJAEN	54.7%
pol-uned1-2-5l	LSI UNED	53.9%
pol-uned1-1-5l	LSI UNED	52.5%
pol-uned2-2-5l	LSI UNED 2	40.4%
pol-uned2-1-5l	LSI UNED 2	40.0%
pol-uned2-3-5l	LSI UNED 2	39.5%
pol-uned2-4-5l	LSI UNED 2	38.6%
pol-imdea-1-5l	IMDEA	36.0%
pol-sinai-2-5l	SINAI - UJAEN	35.7%
pol-sinai-1-5l	SINAI - UJAEN	35.3%
pol-sinai-3-5l	SINAI - UJAEN	35.0%
pol-uma-1-5l	UMA	16.7%

Table 4: Results for task 1 (Sentiment Analysis) with 5 levels + NONE

In order to perform a supplementary evaluation, Table 5 gives results 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.

In this case, precision values improve, as expected. The precision obtained now ranges from 71.1% to 35.1%. In this case, 9 submissions have a precision value over 50% and 6 groups have at least one result over this percent.

Run Id	Group	Precision
pol-elhuyar-1-3l	Elhuyar Fund.	71.1%
pol-l2f-1-3l	L2F - INESC	69.0%
pol-l2f-3-3l	L2F - INESC	69.0%
pol-l2f-2-3l	L2F - INESC	67.6%
pol-atrilla-1-3l	La Salle - URL	62.0%
pol-sinai-4-3l	SINAI - UJAEN	60.6%
pol-uned1-1-3l	LSI UNED	59.0%
pol-uned1-2-3l	LSI UNED	58.8%
pol-uned2-1-3l	LSI UNED 2	50.1%
pol-imdea-1-3l	IMDEA	46.0%
pol-uned2-2-3l	LSI UNED 2	43.6%
pol-uned2-4-3l	LSI UNED 2	41.2%
pol-uned2-3-3l	LSI UNED 2	40.4%
pol-uma-1-3l	UMA	37.6%
pol-sinai-2-3l	SINAI - UJAEN	35.8%
pol-sinai-1-3l	SINAI - UJAEN	35.6%
pol-sinai-3-3l	SINAI - UJAEN	35.1%

Table 5: Results task 1 (Sentiment Analysis) with 3 levels + NONE

Table 6 shows the results for Task 2. 13 experiments were submitted in.

Run Id	Group	Precision
top-l2f-2	L2F – INESC	65.4%
top-l2f-1y3	L2F – INESC	64.9%
top-atrilla-1	La Salle - URL	60.1%
pol-uned2-5a8	LSI UNED 2	45.3%
top-imdea-1	IMDEA	45.2%
pol-uned2-9a12	LSI UNED 2	42.2%
pol-uned2-1a4	LSI UNED 2	40.5%
top-sinai-5	SINAI - UJAEN	39.4%
top-sinai-4	SINAI - UJAEN	37.8%
top-sinai-2	SINAI - UJAEN	34.8%
top-sinai-3	SINAI - UJAEN	34.1%
top-sinai-1	SINAI - UJAEN	32.3%
pol-uned1-1y2	LSI UNED	31.0%

Table 6: Results for task 2 (Trending topic coverage)

In this task, precision ranges from 65.4% to 31.0% and only 4 submissions are above 50% (2 groups). As in task 1, different submissions from the same group usually get a similar precision, thus showing that the variations from each baseline do not affect much.

6 Conclusions and Future Work

TASS has been the first workshop about sentiment analysis in the context of SEPLN. We expected to attract a certain interest in the proposed tasks, as many groups around the world are currently carrying out an intense research in sentiment/opinion analysis in general and using short-texts in particular. We think that the number of participants, the quality of their work and their reports, and the good results achieved in such hard tasks, has met and gone beyond all our expectations.

The diversity of groups coming from different fields and areas of expertise including information retrieval, natural language processing, computational linguistics, machine learning, data or text mining, text analytics, and even semantic web, has shown that the sentiment analysis is becoming a trending topic within the information technology field.

Some participants expressed in their papers and during the workshop some concerns about the quality of both the annotation of the training corpus and also of the gold standard (the test corpus). In case of future editions of TASS and the reuse of the corpus, more effort must be invested in filtering errors and improving the annotation of the corpora.

Furthermore, as expressed by Moreno-Ortiz and Pérez-Hernández (2012), there is a need of further discussion about whether differentiating between neutral and no polarity is the best decision, since it is not always clear what the difference is, and, moreover, if this distinction is interesting from a practical perspective.

In future editions of the workshop, it would be interesting to extend the corpus to other languages to compare the performance of the different approaches on different languages.

References

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

Saralegi Urizar, Xabier; San Vicente Roncal, Iñaki. 2012. *TASS: Detecting Sentiments in Spanish Tweets*. TASS 2012 Working Notes. Castellón, September 2012.

Fernández Anta, Antonio; Morere, Philippe; Núñez Chiroque, Luis; and Santos, Agustín. 2012. *Techniques for Sentiment Analysis and Topic Detection of Spanish Tweets: Preliminary Report*. TASS 2012 Working Notes. Castellón, September 2012.

Hall, Mark; Frank, Eibe; Holmes, Geoffrey; Pfahringer, Bernhard; Reutemann, Peter; Witten, Ian H. 2009. *The WEKA Data Mining Software: An Update*. SIGKDD Explorations, Volume 11, Issue 1.

Batista, Fernando; Ribeiro, Ricardo. 2012. *The L2F Strategy for Sentiment Analysis and Topic Classification*. TASS 2012 Working Notes. Castellón, September 2012.

Trilla, Alexandre; Alías, Francesc. 2012. *Sentiment Analysis of Twitter messages based on Multinomial Naive Bayes*. TASS 2012 Working Notes. Castellón, September 2012.

Martín-Wanton, Tamara; Carrillo de Albornoz, Jorge. 2012. *UNED en TASS 2012: Sistema para la Clasificación de la Polaridad y Seguimiento de Temas*. TASS 2012 Working Notes. Castellón, September 2012.

Castellanos, Ángel; Cigarrán, Juan; García-Serrano, Ana. 2012. *Generación de un corpus de usuarios basado en divergencias del Lenguaje*. II Congreso Español de Recuperación de Información. Valencia, June 2012.

Castellano, Angel; Cigarrán, Juan; García-Serrano, Ana. 2012. *UNED @ TASS: Using Information Retrieval techniques for topic-based sentiment analysis through divergence models*. TASS 2012 Working Notes. Castellón, September 2012.

Martínez Cámara, Eugenio; García Cumbreras, M. Ángel. Martín Valdivia, M. Teresa; Ureña López, L. Alfonso. 2012. *SINAI at TASS 2012*. TASS 2012 Working Notes. Castellón, September 2012.

Moreno-Ortiz, Antonio; Pérez-Hernández, Chantal. 2012. *Lexicon-Based Sentiment Analysis of Twitter Messages in Spanish*. TASS 2012 Working Notes. Castellón, September 2012.