DAEDALUS at RepLab 2012: Polarity Classification and Filtering on Twitter Data

Julio Villena-Román[1,2], Sara Lana-Serrano[3,1], Cristina Moreno-García[1],
Janine García-Morera[1], José Carlos González-Cristóbal[3,1]

[1] DAEDALUS - Data, Decisions and Language, S.A.
[2] Universidad Carlos III de Madrid
[3] Universidad Politécnica de Madrid
jvillena@daedalus.es, slana@diatel.upm.es, cmoreno@daedalus.es,
jgarcia@daedalus.es, josecarlos.gonzalez@upm.es

Abstract. This paper describes our participation at the RepLab 2012 profiling scenario, in both polarity classification and filtering subtasks. Our approach is based on 1) the information provided by a semantic model that includes rules and resources annotated for sentiment analysis, 2) a detailed morphosyntactic analysis of the input text that allows to lemmatize and divide the text into segments to be able to control the scope of semantic units and perform a fine-grained detection of negation in clauses, and 3) the use of an aggregation algorithm to calculate the global polarity value of the text based on the local polarity values of the different segments, which includes an outlier filter. The system, experiments and results are presented and discussed in the paper.

Keywords: RepLab, CLEF, reputation analysis, profiling scenario, filtering, polarity classification, sentiment analysis, STILUS.

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. 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 increasingly complex marketplace of personal and business relationships among people and companies. From the technology perspective, the first step towards the automatic reputation analysis is a *sentiment analysis*, i.e., the application of natural language processing and text analytics to identify and extract subjective information from texts about the sentiments, emotions or opinions contained.

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

Reputation 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.

RepLab [1] is a competitive evaluation exercise for reputation analysis, launched in 2012 edition of CLEF campaign, which focuses on two scenarios: *profiling* and *monitoring* scenario. For both scenarios, systems are provided with a set of tweets in Spanish and English related to several companies. The profiling scenario must annotate two kinds of information in those tweets: 1) filtering information, i.e., whether the tweets are related or not to the company, and 2) polarity classification of the tweet, i.e., if the tweet content has positive or negative implications for the company's reputation. The monitoring scenario consists of clustering a given stream of tweets, assigning relative priorities.

This paper describes our participation at the RepLab 2012 profiling scenario, in both polarity classification and filtering subtasks. We are a research group led by DAEDALUS[2], a leading provider of language-based solutions in Spain, and research groups of Universidad Politécnica and Universidad Carlos III of Madrid. We are long-time participants in CLEF, in many different tracks and tasks since 2003.

RepLab is a new task within CLEF. There was a related task in NTCIR three years ago called Multilingual Opinion Analysis Task [2], active for two editions, focused on sentiment analysis. Another somewhat related task in CLEF was Web People Search [3], focusing on the problem of ambiguity for organization names and the relevance of web data for reputation management purposes. We took part in both initiatives as participant research groups [4] [5].

Our approach to the polarity classification is based on 1) the information provided by a semantic model that includes rules and resources (polarity units, modifiers, stopwords) annotated for sentiment analysis, 2) a detailed morphosyntactic analysis of the input text that allows to lemmatize and split the text into segments in order to be able to control the scope of semantic units and perform a fine-grained detection of negation in clauses, and 3) the use of an aggregation algorithm to calculate the global polarity value of the text based on the local polarity values of the different segments, which includes an outlier detection. Our system, experiments and results achieved are presented and discussed in the following sections.

2 Profiling Scenario

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

---

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

companies and their customers' sentiments towards their brands and products. And this creates new market opportunities for the linguistic technology industry.

Thus the main goal behind our participation was to evaluate, in a multilingual scenario and using social media data, the software and resources for sentiment analysis and named entity detection that have been developed by our company in the last year.

This year we focused on the *profiling* scenario, which includes two subtasks: polarity classification and filtering. The following sections give more in-depth details about our work in both subtasks.

2.1 Polarity Classification Subtask

2.1.1 Overview

The polarity classification is based on our software for multilingual sentiment analysis [6], which is available through a web API offered through a REST-based web service. This component performs an in-depth analysis of the input text to determine if it expresses a positive/negative/neutral sentiment or else no sentiment at all.

First the local polarity of the different clauses in the text ("segments") is identified and then the relation among them is evaluated in order to obtain a global polarity value for the whole given text. The output for both the local and global polarity is encoded with a real number ranging from -1 (*strong negative*) to +1 (*strong positive*) and also a set of labels representing 5 discrete levels to simplify the post-processing: *strong positive* (*P+*), *positive* (*P*), *neutral* (*NEU*), *negative* (*N*), *strong negative* (*N+*), and one additional no-sentiment tag (*NONE*).

Apart from the text itself, which can be encoded in plain text, HTML or XML, another required input parameter is the semantic model to use in the sentiment evaluation. This semantic model defines the domain of the text (the analysis scenario) and is mainly based on an extensive set of dictionaries and rules that incorporate both the well-known "domain-independent" polarity values (for instance, in general, in all contexts, "good" is positive and "awful" is negative) and also the specificities of each analysis scenario (for instance, an "increase" in the "interest rate" is probably positive for financial companies but negative for the rest of the people). The semantic model also encodes implicitly the language of text.

Furthermore, the component is able to identify named entities and concepts, referred to as attributes, and assign a specific polarity value to them, depending on the selected semantic model and the context in which the attributes appear. In this case, this information has been used for the second subtask (identifying whether tweets are related or not to the companies).

The component makes an internal call to another software component [7], also accessible through a REST-based web service, in order to split the text into segments, perform the POS tagging and the extraction of their morphosyntactic structure to be used in the sentiment analysis, and identify the named entities and concepts.

The sentiment analysis process is described in detail in the next section.

2.1.2 Sentiment Analysis Process

The sentiment analysis is carried out in the following sequence of steps:

1. **Segment detection**. The text is parsed and split into segments. Although most times these segments are full sentences in "usual" texts (well-written news articles, blog posts, etc.), this is not the case in tweet messages, as the analysis depends on the presence of punctuation marks and correct capitalization of words. Figure 1 and Figure 2 show two examples of tweets in the test set.

Tweet id [entity]: `194453767528259584 [RL2012E06]`
```
RT @elpais_inter: Egipto cancela el acuerdo de gas con Israel. El
suministro egipcio suponía un 40% del consumo israelí de gas natural
```
Segment 1: `RT @elpais_inter:`
Segment 2: `Egipto cancela el acuerdo de gas con Israel.`
Segment 3: `El suministro egipcio suponía un 40% del consumo israelí de gas natural`

Figure 1. Example of segment detection (example 1).

Tweet id [entity]: `200623340069732352 [RL2012E35]`
```
the thing is, apple OS is neat and tidy. microsoft win is much more
harder to see everything that you need to maximize them
```
Segment 1: `the thing is, apple OS is neat and tidy.`
Segment 2: `microsoft win is much more harder to see everything that you need to maximize them`

Figure 2. Example of segment detection (example 2).

2. **Linguistic processing** (lemmatization, morphosyntactic analysis and entity and concept detection)**.** First each segment is tokenized (considering multiword units) and then each token is analyzed to extract its lemma(s).

   In addition, a morphosyntactic analysis is performed to divide the segment into proposition or clauses. This division is useful, as described later, for detecting the negation and analyzing the effect of modifiers on the polarity values. Focusing on a given clause, it is assigned a "clause level" equal to 0, and any step into/out a subordinated clause adds/subtracts 1 from that clause level.

   Last but not least, a named entity and concept recognition step is carried out, based in multilingual linguistic resources and heuristics for detecting unknown *PERSON*, *LOCATION* and/or *ORGANIZATION* entities.

   Next Figure 3 and Figure 4 show the output of this step corresponding to the previous examples.

```
{ { { RT { @elpais_inter } } : } }
{ { { republica_arabe_de_egipto|egipto } { cancelar|cancela } { el
acuerdo { de { gas } } } { con { estado_de_israel|israel } } } . }
{ { { el suministro egipcio } { suponer|suponía } { uno|1|un 40% { del
{ consumo israelí { de { gas_natural } } } } } } }
```

Figure 3. Example of linguistic processing (example 1).

```
{ { { the thing } be|is } , { { apple open_source } be|is { neat } }
and { tidy } . }

{ { { Microsoft win } be|is { much_more hard|harder } see|to_see {
everything } } that { { you } need_to_maximize { them }
```

Figure 4. Example of linguistic processing (example 2).

A visual representation of the syntactic structure is shown in Figure 9 and Figure 10 in the Appendix.

3. **Detection of negation.** The next step is to iterate over every token of each segment to tag whether the token is affected by negation or not. If a given token is affected by negation, the eventual polarity level is reversed (turns from positive to negative and the other round).

   For this purpose, the semantic model includes a list of negation units (NEG), such as the obvious negation particles (adverbs) such as "*no*", "*ni*" (in Spanish) or "*not*" (and its contracted form without/with the auxiliary verbs), "*neither*" (in English) but also words or expressions such as "*carecer*", "*dejar de*", "*bajo ningún concepto*" (in Spanish) or "*against*", "*far from*", "*no room for*" (in English).

   Each NEG unit is considered to affect clauses with a relative (to the NEG unit) clause level up to a given threshold (*NEGATION_LEVEL*) and tokens separated a relative distance up to another threshold (*NEGATION_MAXDISTANCE*), excluding certain punctuation marks (brackets, quotes, colon and semicolon). For Twitter messages, the level threshold is -1 – thus a NEG unit affects to its own clause (group level = 0), any subordinate clause (group level > 0) and its parent clause (group level = -1) –, and the maximum distance threshold is 20.

   The information of negation is stored (as true or false) in each token to be used in the next step.

   The previous examples do not include any negation unit, so all tokens are marked as positive.

4. **Detection of modifiers**. Some special units (MOD units) do not assign a specific polarity value but operate as modifiers of this value, incrementing or decrementing it.

   MOD units included in the semantic model can be assigned a + (positive), ++ (strong positive), - (negative) or -- (strong negative) value. For instance, "if "*good*" is positive (*P*), "*very good*" is be strong positive (*P+*), thus "*very*" would be a positive modifier (+); the opposite is the case of "*less*", which would be a negative modifier (-) ("*less good*" would be *P-*). Some other examples of modifiers are "*adicional*", "*ampliación*", "*principal*" (all positive) or "*apenas*", "*medio*" (negative) (in Spanish) or "*additional*", "*a lot*", "*completely*" (positive) or "*descend*", "*almost*" (negative) (in English).

   Similarly to the negation detection, modifiers are considered to affect clauses with a relative level (*MODIFIER_LEVEL*) and tokens separated a relative distance (*MODIFIER_MAXDISTANCE*) up to a defined threshold values. For this task, the level threshold is 0 (only the clause itself and subordinated clauses) and the maximum distance threshold is 5.

The second previous example includes two positive (+) modifiers, “*much*” and “*more*”.

5. **Polarity tagging**. The next step is to detect polarity units (POL units) in the segments. The POL units in the semantic model can be assigned one of the following values, ranging from the most positive to the most negative: *P++*, *P+*, *P*, *P-*, *P--*, *N--*, *N-*, *N*, *N+* and *N++*.

   To help to avoid false positives, the semantic model also includes stopword units (SW units).

   Moreover, POL units can include a context filter, i.e., one or several words or expressions that must appear or not in the segment so that the unit is considered in the sentiment analysis. Obviously, context filters highly depend on the analysis domain. For example, there are many concepts that are positive (*P*) when increased (such as reputation, employment...) and negative (*N*) when decreased; this could be represented by the following set of rules (including macros):

```
#INCREASE#  increase|increment|grow|growth|gain|rise|go_up|climb
#DECREASE#  decrease|decrement|reduce|loss|do_down|descent

reputation/#INCREASE#   P
reputation/#DECREASE#   N
```

   or else, to increase the recall in the case of missing expressions:

```
reputation/#INCREASE#   P
reputation              N
```

   The final value for each POL unit is calculated from the polarity value of the POL unit in the semantic model, adding or subtracting the polarity value of the modifier (if the thresholds are fulfilled) and considering the negation (again, if the thresholds are fulfilled).

   The previous examples are tagged as shown in Figure 5 and Figure 6.

```
{ { { RT { @elpais_inter } } : } }
   @elpais_inter              entity

{ { { republica_arabe_de_egipto|egipto } { cancelar|cancela } { el
acuerdo { de { gas } } } { con { estado_de_israel|israel } } } . }
   República_Árabe_de_Egipto  entity
   Estado_de_Israel           entity
   cancelar/acuerdo           POL (N+)

{ { { el suministro egipcio } { suponer|suponía } { uno|1|un 40% { del
{ consumo israelí { de { gas_natural } } } } } } }
   gas_natural                SW
```

Figure 5. Example of polarity tagging (example 1).

```
{ { { the thing } be|is } , { { apple open_source } be|is { neat } }
and { tidy } . }
   Open_source                   entity
   neat                          POL (P-)
   tidy                          POL (P-)

{ { { Microsoft win } be|is { much_more hard|harder } see|to_see {
everything } } that { { you } need_to_maximize { them }
   Microsoft                     entity
   win                           POL (P)
   much                          MOD (+)
   more                          MOD(+)
   hard_to_see                   POL (N)
     (much_more) hard_to_see     (N++)
```

Figure 6. Example of polarity tagging (example 2).

6. **Segment scoring**. To calculate the overall polarity of each segment, an aggregation algorithm is applied to the set of polarity values given by the POL units detected in the segment. The aggregation algorithm performs an outlier filtering to try to reduce the effect of miss detections of NEG, MOD or POL units, based on a threshold over the standard deviation from the average of values. The aggregation algorithm finally calculates the *average* and the *standard deviation* of the set of accepted values, which is assigned as the score of the segment.

   In addition to this numeric score, to simplify the post-processing, discrete nominal values are also assigned to each segment: *N+* if score < -0.6, *N* if score < -0.2, *NEU* (neutral) if score < +0.2, *P* if score < 0.6 or else *P+*. If there is no POL unit, the segment is assigned with a polarity value of *NONE*.

   The standard deviation is an indication of the level of agreement within the segment. With this value, we can differentiate for instance whether a segment has a *NEU* score (near 0) because all present POL units or modifiers have a neutral sentiment, so the standard deviation is low, or else there are positive and negative units that lead to a low average but a high standard deviation value. The first case would be detected as *AGREEMENT* (standard deviation < 0.2) and the second as *DISAGREEMENT*.

   In the previous first example, all segments have one POL unit at maximum, so the segment average has the same value and an *AGREEMENT* label. The second example contains a segment with two POL units, “*neat*” and “*tidy*”, which have the same score, so the segment has the same average value and an *AGREEMENT* label. The other segment has a *DISAGREEMENT* label because it contains one positive and one negative POL unit.

7. **Global text scoring.** The same aggregation algorithm is applied to the local polarity values of each segment to calculate the global polarity value of the text, represented by an average value (both numeric and nominal values) that indicate the actual value and a standard deviation that indicates the level of agreement or disagreement within the different segments of the text.

Again, if there is no segment with polarity information (i.e., different from *NONE*), the text is assigned with a global polarity value of *NONE*.

In the first example, the global score has the same value as the only segment that has a sentiment score. In the second example, the global polarity turns to be *NEU* (neutral) with a *DISAGREEMENT* between the two segments.

8. **Attribute scoring.** Additionally, a similar process is applied to the named entities and concepts (the attributes) that have been detected in the segments during the morphosyntactic analysis to calculate their polarity, in this case, considering which POL unit (along with its modifier(s) and possible negation) is affecting each attribute, and using the same aggregation algorithm.

Figure 7 and Figure 8 show the final output in XML of the sentiment analysis.


```
- <report>
    <status code="0">OK</status>
    <model>es-general</model>
    <score>-0.80</score>
    <scoretag>N+</scoretag>
    <sd>0.00</sd>
    <sdtag>AGREEMENT</sdtag>
  - <segments>
    - &lt;segment&gt;
        <inip>0</inip>
        <endp>16</endp>
        <text>RT @elpais_inter:</text>
      - <entities>
          <entity>@elpais_inter</entity>
        </entities>
      &lt;/segment&gt;
    - &lt;segment&gt;
        <inip>18</inip>
        <endp>61</endp>
        <text>Egipto cancela el acuerdo de gas con Israel.</text>
        <score>-0.80</score>
        <scoretag>N+</scoretag>
        <sd>0.00</sd>
        <sdtag>AGREEMENT</sdtag>
      - <keywords>
        - <keyword>
            <text>cancelar/acuerdo</text>
            <score>-0.80</score>
            <scoretag>N+</scoretag>
          - <entities>
              <entity>República Árabe de Egipto</entity>
              <entity>Estado de Israel</entity>
            </entities>
          </keyword>
        </keywords>
      &lt;/segment&gt;
    </segments>
    <keywords>cancelar/acuerdo/N+</keywords>
  - <entities>
    - <entity>
        <text>República Árabe de Egipto</text>
        <score>-0.80</score>
        <scoretag>N+</scoretag>
        <sd>0.00</sd>
        <sdtag>AGREEMENT</sdtag>
      </entity>
    - <entity>
        <text>Estado de Israel</text>
        <score>-0.80</score>
        <scoretag>N+</scoretag>
        <sd>0.00</sd>
        <sdtag>AGREEMENT</sdtag>
      </entity>
    - <entity>
        <text>@elpais_inter</text>
      </entity>
    </entities>
</report>
```


Figure 7. Final output (example 1).

- <report>
<status code="0">OK</status>
<model>en-general</model>
<score>0.10</score>
<scoretag>NEU</scoretag>
<sd>0.70</sd>
<sdtag>DISAGREEMENT</sdtag>
- <segments>
- <segment>
<inip>0</inip>
<endp>39</endp>
<text>the thing is, apple OS is neat and tidy.</text>
<score>0.40</score>
<scoretag>P</scoretag>
<sd>0.00</sd>
<sdtag>AGREEMENT</sdtag>
- <keywords>
- <keyword>
<text>neat</text>
<score>0.40</score>
<scoretag>P-</scoretag>
- <entities>
<entity>Open source</entity>
</entities>
</keyword>
- <keyword>
<text>tidy</text>
<score>0.40</score>
<scoretag>P-</scoretag>
- <entities>
<entity>Open source</entity>
</entities>
</keyword>
</keywords>
</segment>
- <segment>
<inip>45</inip>
<endp>126</endp>
- <text>
Microsoft win is much more harder to see everything that you need to maximize them
</text>
<score>-0.20</score>
<scoretag>NEU</scoretag>
<sd>0.80</sd>
<sdtag>DISAGREEMENT</sdtag>
- <keywords>
- <keyword>
<text>win</text>
<score>0.60</score>
<scoretag>P</scoretag>
- <entities>
<entity>Microsoft</entity>
</entities>
</keyword>
- <keyword>
<text>(much_more) hard_to_see</text>
<score>-1.00</score>
<scoretag>N++</scoretag>
- <entities>
<entity>Microsoft</entity>
</entities>
</keyword>
</keywords>
</segment>
</segments>
- <keywords>
neat/P-, tidy/P-, win/P, (much_more) hard_to_see/N++
</keywords>
- <entities>
- <entity>
<text>Open source</text>
<score>0.40</score>
<scoretag>P</scoretag>
<sd>0.00</sd>
<sdtag>AGREEMENT</sdtag>
</entity>
- <entity>
<text>Microsoft</text>
<score>-0.20</score>
<scoretag>NEU</scoretag>
<sd>0.80</sd>
<sdtag>DISAGREEMENT</sdtag>
</entity>
</entities>
</report>

Figure 8. Final output (example 2).

2.1.3 Semantic Models

Currently there are several semantic models available, some of them developed for general-purpose sentiment analysis and some other for specific cases such as the financial, telecommunications and tourism domains. For the RepLab tasks, the general-purpose models for Spanish and English have been used. Those models where initially inspired in the linguistic resources provided by General Inquirer [7] in English, specifically, terms extracted from the "*Positive*", "*Negative*", "*Strong*" and "*Weak*" categories of Harvard IV-4 dictionary (included in the General Enquirer).

The following Table 1 presents some information about these models.

Table 1. Contents of the semantic models.

Type of unit	**Spanish**	**English**
Negation (NEG units)	59	28
Modifiers (MOD units)	372	107
--	5	3
-	106	12
+	255	72
++	6	20
Polarity (POL units)	3 139	4 226
N++	10	78
N+	340	285
N	1 309	2 106
N-	206	209
N--	11	10
P--	15	6
P-	15	72
P	978	1 113
P+	248	325
P++	7	22
Stopwords (SW units)	91	33
Macros	27	10
TOTAL	3 688	4 404

2.1.4 Submissions

To perform the experiments of the polarity classification subtask, a client was developed for that web service. This client reads each tweet in the test corpus along with the language, makes a call to the web service and parses the response to adapt the returned values to the ones required in the task: *P* and *P+* are "*positive*", *N* and *N+* are "*negative*" and the rest (whether *NEU* or *NONE*) are tagged as "*neutral*".

Just one submission for the polarity classification subtask was made: "*replab2012_polarity_Daedalus_1*". Results are discussed in the corresponding section.

2.2 Filtering Subtask

2.2.1 Overview

Our approach to the filtering subtask is to reuse the result of the named entity recognition step in the linguistic previous analysis, which was performed by calling another software component through a REST-based web service [7].

The difficulty of the detection arises from the fact that entities may appear in different forms: for instance, "*Banco Santander Central Hispano*" may appear as "*BSCH*", "*Banco Santander*", "*Banco de Santander*", "*Santander*", etc. In addition, once detected, there is the problem of ambiguity, both among different categories and even within the same category: for instance, "*Seville*" may be the well-known city in Spain, the soccer team, etc.

2.2.2 Named Entity Detection Process

The software uses the widely-adopted approach based on knowledge, i.e., manually-developed dictionaries and rule sets are used to perform the detection and classification. The main drawback of this approach is the high costs to develop and maintain the resources, as they are highly dependent on language and domain.

The current multilingual entity dictionaries include over 41 000 persons, 17 000 organizations and 45 000 locations. Apart from these common dictionaries, our software allows to include user dictionaries that are specific for a given domain and complement the common dictionaries.

In addition, rules apply regular expression patterns to the entities in the dictionaries to generate a set of possible variants in which that entity might occur, for instance:

```
(N)ame (S)urname :- Name / Surname / N. Surname / Name S. / N. S.
```

Fernando Alonso → Fernando / Alonso / F. Alonso / Fernando A. / F. A.

```
(A)aaa (of|the)? (B)bbb(of|the)? (C)ccc (of|the)? (D)ddd :- ABCD
```

Organization of the Petroleum Exporting Countries → OPEC

Thus our system allows the advanced recognition of unknown entities that are proposed as suggested entities: for instance, "*Mr. Aaaaa Bbbbb*" could be a PERSON name, "*Bank of Ddddd*" an *ORGANIZATION*, "*Eeeee Square*" a *LOCATION*, etc.

The process is as follows:

1. Text is segmented into units (words or multiword expressions).

2. Those units that are contained in any of the entity dictionaries are marked as candidate entities, no matter if they occur in the exact form or in a variant (alias).

3. If any unit matches more than one candidate entity, an heuristic-based disambiguation is carried out, using for instance the frequency of that unit in the text ("*Castro*" will be selected as "*Fidel Castro*" if that name is present in the

text and not "*Raul Castro*"), the presence of discursive clues (for instance, *towards+LOCATION* and *article+ORGANIZATION*: "*towards Madrid*" is disambiguated as the city and "*this Madrid*" as the soccer team), disambiguation based on geographical context (depending on the georeferences in the text), etc.

As a result, entities that appear in the text are returned, along with their class and position in the text.

2.2.3 Submissions

For carrying out the filtering task, three different specific dictionaries ("*user dictionaries*") have been defined, as described in Table 2. Although it is possible to make those dictionaries language-specific, we mixed entries in both Spanish and English to simplify the processing.

Table 2. Description of dictionaries.

Dictionary	Contents
Dictionary 1	• List of entities in the test corpus, along with their well-known variants and aliases extracted from Wikipedia pages. • Products and services from those companies. • A list of stopwords for some very ambiguous entities (for instance, "*BME*" also means "*Boston Most Elite*" and "ING" is the abbreviation for "*ingeniero*" *-engineer-* in Spanish).
Dictionary 2	• The previous dictionary plus variants and aliases extracted from the company web sites. • Email addresses, usernames, hashtags used for those companies in social networks. • Stopwords now include references to foundations, external activities of the companies as sponsoring sporting events or competitions (to avoid positives, for instance, for "*Liga BBVA*", "*Regata Mapfre*", "*Ferrari team*").
Dictionary 3	• Stopwords now include an extensive list of car models (to avoid positives, for instance, for "*Chevrolet Camaro*" or "*VW Golf*").

Similarly to the polarity classification subtask, a client was developed for the web service to perform the experiments. This client again reads each tweet in the test corpus along with the language, makes a call to the web service indicating one of the three different dictionaries at one time, and parses the response. If the expected entity is detected in the text, "yes" is assigned to the tweet and "no" otherwise.

We submitted three experiments corresponding to each dictionary: "*replab2012_related_Daedalus_1*", "*replab2012_related_Daedalus_2*" and "*replab2012_related_Daedalus_3*". Results are described in the next section.

3 Results

Results achieved by the top ranked experiments in the polarity classification subtask are shown in Table 3. The columns in the table are Accuracy (A), Reliability (R), Sensitivity (S) and the typical F-measure calculated over Reliability and Sensitivity.

Table 3. Polarity classification results.

	A	**R**	**S**	**F(R,S)**
All				
replab2012_polarity_Daedalus_1	0.4796	0.3924	0.4491	**0.4018**
replab2012_profiling_uned_5	0.4495	0.3402	0.3747	0.3419
replab2012_profiling_BMedia_2	0.4090	0.3315	0.3651	0.3351
replab2012_profiling_uiowa_2	0.3462	0.3070	0.3899	0.3343
replab2012_profiling_uned_2	0.4866	0.3255	0.3147	0.3078
English				
replab2012_polarity_Daedalus_1	0.4013	0.3452	0.3668	0.3349
replab2012_profiling_uned_5	0.4680	0.3692	0.3496	**0.3483**
replab2012_profiling_BMedia_2	0.4428	0.3421	0.3729	0.3473
replab2012_profiling_uiowa_2	0.4011	0.3180	0.3839	0.3334
replab2012_profiling_uned_2	0.5378	0.2683	0.1967	0.2141
Spanish				
replab2012_polarity_Daedalus_1	0.4802	0.4144	0.4497	**0.4143**
replab2012_profiling_uned_5	0.4269	0.3130	0.3127	0.2961
replab2012_profiling_BMedia_2	0.4182	0.2968	0.3053	0.2839
replab2012_profiling_uiowa_2	0.2948	0.2897	0.3390	0.3011
replab2012_profiling_uned_2	0.4267	0.2926	0.2825	0.2803

The only experiment submitted achieved the best performance of all participants for all languages in general and specifically for Spanish. The difference between Spanish and English, though not very high, is probably because the linguistic processing modules (the tokenizer, stemmer and specially the morphosyntactic analyzer) and the resources included in the semantic model are better for the case of Spanish, the main target language of our market.

The different entities have been organized into a set of sectors of economic activity. Results achieved per sector by our experiment for all languages in general are shown in Table 4.

This table gives an idea of the domains that are best covered by our semantic models. In this case, the "*Banking and Insurance*", "*Audiovisual*" and "*Telecommunications*" sectors are the best covered, whereas the "*Transport and Infrastructure*" (corresponding to "*International Consolidated Airlines Group*" entity) is by large the worst covered.

Table 4. Polarity classification results per activity sector (all languages).

Activity Sector	A	R	S	F(R,S)
Audiovisual	0.5900	0.4300	0.4900	0.4580
Automotive	0.4020	0.3300	0.3920	0.3530
Banking and Insurance	0.4000	0.4500	0.5483	**0.4909**
Energy	0.4780	0.4380	0.4060	0.4133
Personal care	0.4833	0.3750	0.3900	0.3619
Technology and Software	0.5000	0.3160	0.5000	0.3572
Telecommunications	0.5300	0.4200	0.4300	0.4249
Textile	0.5000	0.4500	0.5300	0.4867
Transport and Infrastructure	0.7300	0.4200	0.1300	*0.1985*

Entities that have been marked with “no samples” in the “Sensitivity over Polarity” column of the result spreadsheet, listed in Table 5, are not included in the calculations.

Table 5. Entities marked with “no samples” in the “Sensitivity over Polarity” column.

Entity	Entity Name	Activity Sector
RL2012E12	Indra Sistemas, S. A.	Technology and Software
RL2012E15	ING Group	Banking and Insurance
RL2012E16	Bolsas y Mercados Españoles	Banking and Insurance
RL2012E32	Wilkinson Sword	Personal care

A similar analysis per entity is included in Table 9 in the Appendix. This table may help to improve our semantic model with specific resources for the companies involved.

Next Table 6 shows the results achieved by the top ranked experiments in the filtering subtask. The columns are the same as in previous tables.

Table 6. Filtering results.

	A	R	S	F(R,S)
All				
replab2012_related_Daedalus_2	0.7228	0.2435	0.4330	**0.2639**
replab2012_related_Daedalus_3	0.7022	0.2352	0.4221	0.2535
replab2012_related_Daedalus_1	0.7180	0.2397	0.4037	0.2506
replab2012_related_CIRGDISCO_1	0.7019	0.2179	0.3364	0.2276
replab2012_profiling_kthgavagai_1	0.7741	0.2534	0.3576	0.2228
English				
replab2012_related_Daedalus_2	0.6689	0.3007	0.4427	**0.3161**
replab2012_related_Daedalus_3	0.6477	0.2862	0.4276	0.2997
replab2012_related_Daedalus_1	0.5320	0.2361	0.3336	0.2325

replab2012_related_CIRGDISCO_1	0.7161	0.3002	0.3810	0.2858
replab2012_profiling_kthgavagai_1	0.7164	0.2813	0.3814	0.2705
Spanish				
replab2012_related_Daedalus_2	0.7104	0.1989	0.3386	0.2064
replab2012_related_Daedalus_3	0.6892	0.1988	0.3323	0.2062
replab2012_related_Daedalus_1	0.7947	0.2466	0.3777	0.2540
replab2012_related_CIRGDISCO_1	0.7151	0.3064	0.4630	**0.3241**
replab2012_profiling_kthgavagai_1	0.8252	0.3139	0.3718	0.2776

Again, in general for both languages, our experiments achieve the best results in terms of F-measure of all participants. However, in this case, the performance for English is considerably better for English than for Spanish, which is quite surprising for us. This issue has to be further analyzed.

In any case, the best result is obtained by the "replab2012_related_Daedalus_2" experiment, the one that includes stopwords to avoid matches for external activities (sponsoring, foundations) but does not include the list of car models. So that means that tweets talking about "*Chevrolet Camaro*" are considered to refer to "*Chevrolet*" but "*Ferrari Team*" does not refer to "*Ferrari*". This turns to be a bit inconsistent and raises some doubts about the criteria that have been used for the gold standard.

Results for filtering achieved per sector by our experiments for all languages in general are shown in Table 7.

Table 7. Filtering results per activity sector (all languages).

Activity Sector	A	R	S	F(R,S)
Automotive				
replab2012_related_Daedalus_1	0.7460	0.1880	0.2620	0.2550
replab2012_related_Daedalus_2	0.6620	0.1460	0.2620	0.1668
replab2012_related_Daedalus_3	0.5360	0.1080	0.2120	0.1192
Banking and Insurance				
replab2012_related_Daedalus_1	0.7663	0.0300	0.3567	0.0824
replab2012_related_Daedalus_2	0.7788	0.1033	0.7333	0.1772
replab2012_related_Daedalus_3	0.7788	0.1033	0.7333	0.1772
Energy				
replab2012_related_Daedalus_1	0.7680	0.2275	0.4625	0.3423
replab2012_related_Daedalus_2	0.7640	0.2475	0.5000	0.3804
replab2012_related_Daedalus_3	0.7640	0.2475	0.5000	0.3804
Personal care				
replab2012_related_Daedalus_1	0.7400	0.2900	0.4200	0.2793
replab2012_related_Daedalus_2	0.5233	0.2000	0.2300	0.1960
replab2012_related_Daedalus_3	0.5233	0.2000	0.2300	0.1960
Technology and Software				
replab2012_related_Daedalus_1	0.6217	0.2483	0.4567	0.2679
replab2012_related_Daedalus_2	0.7067	0.2717	0.4283	0.2870

replab2012_related_Daedalus_3	0.7067	0.2717	0.4283	0.2870
Telecommunications				
replab2012_related_Daedalus_1	0.6700	0.4400	0.4900	0.4637
replab2012_related_Daedalus_2	0.7400	0.4900	0.4800	0.4849
replab2012_related_Daedalus_3	0.7400	0.4900	0.4800	0.4849
Transport and Infrastructure				
replab2012_related_Daedalus_1	0.8300	0.7800	0.5600	0.6519
replab2012_related_Daedalus_2	0.8900	0.8400	0.7200	0.7754
replab2012_related_Daedalus_3	0.8900	0.8400	0.7200	0.7754

Again, entities that have been marked with “no samples” in the “Sensitivity over Filtering” column of the result spreadsheet, listed in Table 8, are not included in the calculations.

Table 8. Entities marked with “no samples” in the “Sensitivity over Filtering” column.

Entity	Entity Name	Activity Sector
RL2012E08	Banco Bilbao Vizcaya Argentaria, S.A.	Banking and Insurance
RL2012E16	Bolsas y Mercados Españoles	Banking and Insurance
RL2012E17	Bankia	Banking and Insurance
RL2012E18	Iberdrola	Energy
RL2012E20	Mediaset S.p.A.	Audiovisual
RL2012E22	Industria de Diseño Textil, S.A.	Textile
RL2012E24	Bank of America Corporation	Banking and Insurance
RL2012E36	CaixaBank	Banking and Insurance

Table 7 and the same analysis per entity included in Table 10 in the Appendix again give insights of the sectors that are best covered by our resources and indicate the areas where to invest further efforts.

4 Conclusions and Future work

The significant differences in the results for English and Spanish in both tasks show that there is still much to do in both the enlargement of the semantic resources and also the improvement of the linguistic processing (specially the morphosyntactic analysis), in a general domain or may be focusing on different activity sectors. Future work must be oriented to those aspects.

However, figures show that, despite of the difficulty of the tasks, results are quite acceptable and somewhat validate the fact that this technology may be already included into an automated workflow process for social media mining.

Regarding the polarity classification task, we think that possible future editions should consider the inclusion of a *no-polarity* label, in addition to *positive*, *negative* and *neutral*, to allow to differentiate whether the text has a neutral polarity (neither positive nor negative) or has no polarity at all.

Furthermore, the addition of more levels such as *strong positive* and *strong negative* could also be interesting for the analysis scenario, although this obviously would increase the difficulty of tasks to a great extend.

On the other hand, the filtering task has some points of ambiguity and disagreement regarding the consideration of whether a tweet is related or not to a given company for the case of brand names of products or services, or sponsoring activities. We would thank the elaboration of clear guidelines with the annotation criteria in function of the context.

Acknowledgements

This work has been partially supported by several Spanish research projects: 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) and BUSCAMEDIA: Towards a semantic adaptation of multi-network-multiterminal digital media (CEN-20091026). Authors would like to thank all partners for their knowledge and support.

References

1. Enrique Amigó, Adolfo Corujo, Julio Gonzalo, Edgar Meij and Maarten de Rijke. *Overview of RepLab 2012: Evaluating Online Reputation Management Systems.* CLEF 2012 Labs and Workshop Notebook Papers, 2012.
2. Yohei Seki, David Kirk Evans, Lun-Wei Ku, Le Sun, Hsin-Hsi Chen and Noriko Kando. *Overview of Multilingual Opinion Analysis Task at NTCIR-7*, National Institute of Informatics, October 24, 2008.
3. J. Artiles, A. Borthwick, J. Gonzalo, S. Sekine, and E. Amigó. *WePS-3 Evaluation Campaign: Overview of the Web People Search Clustering and Attribute Extraction Task.* Proceeding of the Conference on Multilingual and Multimodal Information Access Evaluation (CLEF), 2010.
4. Julio Villena-Román, Sara Lana-Serrano, José Carlos González-Cristóbal. *DAEDALUS at WebPS-3 2010: k-Medoids Clustering using a Cost Function Minimization*. CLEF 2010 Labs and Workshops, Notebook Papers. 22-23 September, Padua Italy.
5. Julio Villena-Román, Sara Lana-Serrano and José C. González-Cristóbal. *MIRACLE at NTCIR-7 MOAT: First Experiments on Multilingual Opinion Analysis*. 7th NTCIR Workshop Meeting. Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access. Tokyo, Japan, December 2008.

6. STILUS Sentiment API web page. Available online at http://api.daedalus.es/stilussentiment-info.
7. STILUS Core API web page. Available online at http://api.daedalus.es/stiluscore-info.
8. General Inquirer. Available online at http://www.wjh.harvard.edu/~inquirer.

Appendix

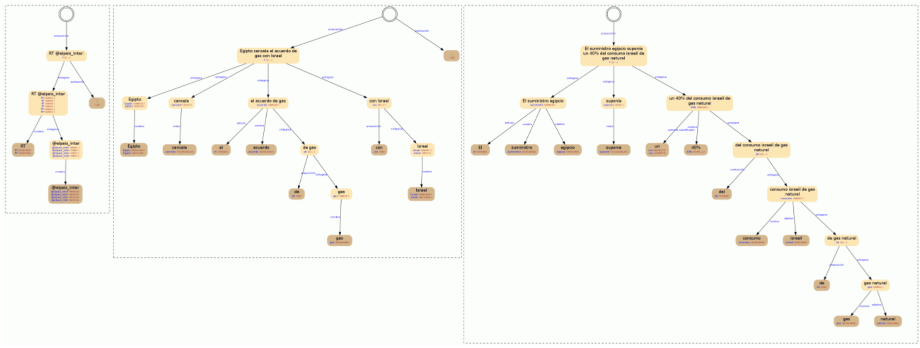

Figure 9. Visual representation of the syntactic structure (example 1).

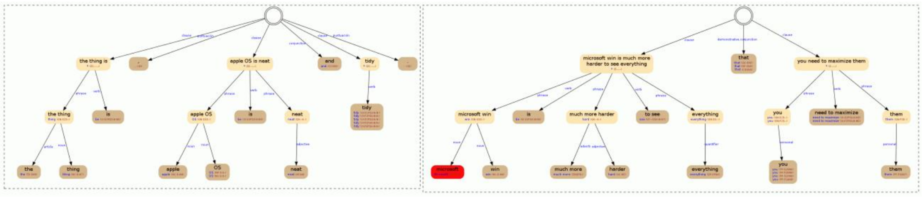

Figure 10. Visual representation of the syntactic structure (example 2).

Table 9. Polarity classification results per activity sector and entity (all languages).

Activity Sector / Entity	A	R	S	F(R,S)
Audiovisual				
Mediaset S.p.A.	0.5900	0.4300	0.4900	0.4580
Automotive				
Bayerische Motoren Werke AG (BMW)	0.3400	0.2900	0.4700	0.3587
Chevrolet	0.3900	0.1300	0.1200	0.1248
Ferrari S.p.A.	0.3600	0.3500	0.3700	0.3597
Fiat S.p.A.	0.5300	0.5900	0.5500	0.5693

Volkswagen	0.3900	0.2900	0.4500	0.3527
Banking and Insurance				
Banco Bilbao Vizcaya Argentaria, S.A.	0.5200	0.3900	0.4400	0.4135
Banco Santander, S.A.	0.6600	0.6700	0.7000	0.6847
Bank of America Corporation	0.4000	0.3700	0.5700	0.4487
Bankia	0.7900	0.4900	0.5000	0.4949
CaixaBank	0.3900	0.4600	0.5800	0.5131
MAPFRE	0.4400	0.3200	0.5000	0.3902
Energy				
BP p.l.c.	0.7500	0.6700	0.4200	0.5163
Endesa, S.A.	0.4000	0.4900	0.5500	0.5183
Gas Natural SDG, S.A.	0.2500	0.1600	0.1500	0.1548
Iberdrola	0.4800	0.5700	0.4900	0.5270
Repsol S. A.	0.5100	0.3000	0.4200	0.3500
Personal care				
Gillette	0.3000	0.6100	0.4700	0.5309
Nivea	0.4000	0.1400	0.3100	0.1929
Technology and Software				
Bing	0.4500	0.3200	0.3400	0.3297
BlackBerry	0.5700	0.3600	0.5200	0.4255
Google Inc.	0.4300	0.2000	0.3900	0.2644
Indra Sistemas, S. A.	0.5000			
Microsoft Corporation	0.6100	0.6200	0.6300	0.6250
Yahoo! Inc.	0.4400	0.0800	0.6200	0.1417
Telecommunications				
Telefónica, S.A.	0.5300	0.4200	0.4300	0.4249
Textile				
Industria de Diseño Textil, S.A.	0.5000	0.4500	0.5300	0.4867
Transport and Infrastructure				
International Consolidated Airlines Group	0.7300	0.4200	0.1300	0.1985

Table 10. Filtering results per activity sector and entity (all languages).

Activity Sector	A	R	S	F(R,S)
Automotive				
Bayerische Motoren Werke AG (BMW)				
replab2012_related_Daedalus_1	0.8500	0.0900	0.2200	0.1277
replab2012_related_Daedalus_2	0.7200	0.0200	0.0900	0.0327
replab2012_related_Daedalus_3	0.5200	0.0200	0.1300	0.0347
Chevrolet				
replab2012_related_Daedalus_1	0.7600	0.1100	0.3100	0.1624

replab2012_related_Daedalus_2	0.8600	0.2400	0.5300	0.3304
replab2012_related_Daedalus_3	0.6300	0.1000	0.3800	0.1583
Ferrari S.p.A.				
replab2012_related_Daedalus_1	0.7800	0.3000	0.4700	0.3662
replab2012_related_Daedalus_2	0.6900	0.1900	0.3300	0.2412
replab2012_related_Daedalus_3	0.6500	0.1600	0.3100	0.2111
Fiat S.p.A.				
replab2012_related_Daedalus_1	0.7700	0.0000	0.0000	
replab2012_related_Daedalus_2	0.4600	0.0900	0.3000	0.1385
replab2012_related_Daedalus_3	0.3000	0.0700	0.1800	0.1008
Volkswagen				
replab2012_related_Daedalus_1	0.5700	0.4400	0.3100	0.3637
replab2012_related_Daedalus_2	0.5800	0.1900	0.0600	0.0912
replab2012_related_Daedalus_3	0.5800	0.1900	0.0600	0.0912
Banking and Insurance				
Banco Santander, S.A.				
replab2012_related_Daedalus_1	0.7200	0.0500	0.4300	0.0896
replab2012_related_Daedalus_2	0.7400	0.0700	0.5900	0.1252
replab2012_related_Daedalus_3	0.7400	0.0700	0.5900	0.1252
ING Group				
replab2012_related_Daedalus_1	0.9700	0.0000	0.0000	
replab2012_related_Daedalus_2	0.9600	0.2000	0.9600	0.3310
replab2012_related_Daedalus_3	0.9600	0.2000	0.9600	0.3310
MAPFRE				
replab2012_related_Daedalus_1	0.6400	0.0400	0.6400	0.0753
replab2012_related_Daedalus_2	0.6600	0.0400	0.6500	0.0754
replab2012_related_Daedalus_3	0.6600	0.0400	0.6500	0.0754
Energy				
BP p.l.c.				
replab2012_related_Daedalus_1	0.5400	0.0400	0.4000	0.0727
replab2012_related_Daedalus_2	0.6900	0.0800	0.6900	0.1434
replab2012_related_Daedalus_3	0.6900	0.0800	0.6900	0.1434
Endesa, S.A.				
replab2012_related_Daedalus_1	0.7500	0.0000	0.0000	
replab2012_related_Daedalus_2	0.4300	0.0000	0.0000	
replab2012_related_Daedalus_3	0.4300	0.0000	0.0000	
Gas Natural SDG, S.A.				
replab2012_related_Daedalus_1	0.9100	0.8000	0.8600	0.8289
replab2012_related_Daedalus_2	0.9200	0.8100	0.8600	0.8343
replab2012_related_Daedalus_3	0.9200	0.8100	0.8600	0.8343
Repsol S. A.				
replab2012_related_Daedalus_1	0.7900	0.0700	0.5900	0.1252

replab2012_related_Daedalus_2	0.8900	0.1000	0.4500	0.1636
replab2012_related_Daedalus_3	0.8900	0.1000	0.4500	0.1636
Personal care				
Gillette				
replab2012_related_Daedalus_1	0.6800	0.3400	0.2900	0.3130
replab2012_related_Daedalus_2	0.7200	0.4100	0.3400	0.3717
replab2012_related_Daedalus_3	0.7200	0.4100	0.3400	0.3717
Nivea				
replab2012_related_Daedalus_1	0.6600	0.2900	0.1000	0.1487
replab2012_related_Daedalus_2	0.4800	0.1600	0.1700	0.1648
replab2012_related_Daedalus_3	0.4800	0.1600	0.1700	0.1648
Wilkinson Sword				
replab2012_related_Daedalus_1	0.8800	0.2400	0.8700	0.3762
replab2012_related_Daedalus_2	0.3700	0.0300	0.1800	0.0514
replab2012_related_Daedalus_3	0.3700	0.0300	0.1800	0.0514
Technology and Software				
Bing				
replab2012_related_Daedalus_1	0.6300	0.4500	0.3500	0.3938
replab2012_related_Daedalus_2	0.6500	0.4300	0.4100	0.4198
replab2012_related_Daedalus_3	0.6500	0.4300	0.4100	0.4198
BlackBerry				
replab2012_related_Daedalus_1	0.4700	0.1600	0.3900	0.2269
replab2012_related_Daedalus_2	0.8800	0.3500	0.2800	0.3111
replab2012_related_Daedalus_3	0.8800	0.3500	0.2800	0.3111
Google Inc.				
replab2012_related_Daedalus_1	0.8900	0.7700	0.8100	0.7895
replab2012_related_Daedalus_2	0.8700	0.7100	0.7900	0.7479
replab2012_related_Daedalus_3	0.8700	0.7100	0.7900	0.7479
Indra Sistemas, S. A.				
replab2012_related_Daedalus_1	0.5000	0.0200	0.5000	0.0385
replab2012_related_Daedalus_2	0.5700	0.0100	0.2800	0.0193
replab2012_related_Daedalus_3	0.5700	0.0100	0.2800	0.0193
Microsoft Corporation				
replab2012_related_Daedalus_1	0.8600	0.0700	0.5800	0.1249
replab2012_related_Daedalus_2	0.9100	0.1000	0.6100	0.1718
replab2012_related_Daedalus_3	0.9100	0.1000	0.6100	0.1718
Yahoo! Inc.				
replab2012_related_Daedalus_1	0.3800	0.0200	0.1100	0.0338
replab2012_related_Daedalus_2	0.3600	0.0300	0.2000	0.0522
replab2012_related_Daedalus_3	0.3600	0.0300	0.2000	0.0522
Telecommunications				
Telefónica, S.A.				

replab2012_related_Daedalus_1	0.6700	0.4400	0.4900	0.4637
replab2012_related_Daedalus_2	0.7400	0.4900	0.4800	0.4849
replab2012_related_Daedalus_3	0.7400	0.4900	0.4800	0.4849
Transport and Infrastructure				
International Consolidated Airlines Group, S.A,				
replab2012_related_Daedalus_1	0.8300	0.7800	0.5600	0.6519
replab2012_related_Daedalus_2	0.8900	0.8400	0.7200	0.7754
replab2012_related_Daedalus_3	0.8900	0.8400	0.7200	0.7754