TASS 2012

Taller de Análisis de Sentimientos en la SEPLN
Workshop on Sentiment Analysis at SEPLN

Julio Villena-Román
Daedalus/THALES UPM
jvillena@daedalus.es
TASS 2012

- Experimental evaluation workshop
- Satellite event of the SEPLN 2012 Conference, held on September 7th, 2012 in Jaume I University at Castellón de la Plana, Spain
- [http://www.daedalus.es/TASS/](http://www.daedalus.es/TASS/)
- Organization

Julio Villena-Román – Daedalus
Sara Lana-Serrano – GSI-UPM
José Carlos González-Cristóbal – GSI-UPM
L. Alfonso Ureña-López – SINAI-UJAEN
Miguel Ángel García-Cumberras – SINAI-UJAEN
María-Teresa Martín-Valdivia – SINAI-UJAEN
Eugenio Martínez-Cámara – SINAI-UJAEN
Agenda

- Context and objectives
- Corpus
- Tasks
- Participants
- Results
Emerging trend...

- **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 [Merriam-Webster]

- In turn, **reputation analysis** is the process of tracking, investigating and reporting an entity's actions and other entities' opinions about those actions to calculate the market value of reputation

- 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, leading to an emerging trend towards **online reputation analysis**
Emerging trend...

- One step is a **sentiment analysis**, i.e., the application of natural language processing and text analytics to identify and extract subjective information from texts.

- Another step is to perform a **text categorization** process to classify the text into a set of predefined facets (categories), thus allowing a further in-depth multidimensional analysis.
Major challenge

- Sentiment analysis is a major technological challenge
  - The task is so hard and ambiguous that even humans often disagree on the sentiment of a given text
  - Issues that one individual finds acceptable or relevant may not be the same to others
  - Impact of multilingual aspects and different contexts
  - The shorter the text is, the harder the task becomes
TASS

- **Objective:**
  - 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

- **Challenge:**
  - To provide a benchmark forum for comparing the latest approaches in this field
  - To provide a benchmark dataset that enables researchers to compare their algorithms and systems
Corpus

- Over 68,000 tweets
- written in Spanish
- by 150 well-known personalities of the world of politics, economy, communication, mass media and culture,
- between November 2011 and March 2012
- classified by topic

### Topic list

<table>
<thead>
<tr>
<th>Topic</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Politics (política)</td>
<td>3,119</td>
</tr>
<tr>
<td>Other (otros)</td>
<td>2,337</td>
</tr>
<tr>
<td>Entertainment (entretenimiento)</td>
<td>1,677</td>
</tr>
<tr>
<td>Economy (economía)</td>
<td>942</td>
</tr>
<tr>
<td>Music (música)</td>
<td>566</td>
</tr>
<tr>
<td>Soccer (fútbol)</td>
<td>252</td>
</tr>
<tr>
<td>Films (cine)</td>
<td>245</td>
</tr>
<tr>
<td>Technology (tecnología)</td>
<td>217</td>
</tr>
<tr>
<td>Sports (deportes)</td>
<td>113</td>
</tr>
<tr>
<td>Literature (literatura)</td>
<td>99</td>
</tr>
</tbody>
</table>
Corpus

- Each message was tagged with
  - its **global polarity**, indicating whether the text expresses a positive, negative or neutral sentiment, or no sentiment at all
    - 5 levels have been defined: strong positive (P+), positive (P), neutral (NEU), negative (N), strong negative (N+) and one additional no sentiment label (NONE)
  - in those cases where applicable, the same polarity related to the **entities** mentioned in the text
  - **level of agreement** or disagreement of the expressed sentiments within the text
  - one or several topics (**thematic areas**)
Corpus

- tweet ID
- user ID
- content*
- creation date
- language
- global polarity, in 5 levels: P+, P, NEU, N, N+ plus NONE
- agreement level: AGR, DISAGR
- when applicable, polarity and agreement level related to each entity
- topics

* subject to restrictions in the Twitter API Terms of Service
Corpus

- Randomly divided into training (10%) and test set (90%)
- **Training corpus**: manually revised and released including tags so that participants could train and validate their experiments
- **Test corpus**: provided without any tag to be used to run and evaluate results submitted by participants

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Value</th>
<th>Training set</th>
<th>Test set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tweets</td>
<td>7 219</td>
<td>60 798</td>
<td></td>
</tr>
<tr>
<td>Topics</td>
<td>10</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Tweet languages</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Users</td>
<td>154</td>
<td>158</td>
<td></td>
</tr>
<tr>
<td>User types</td>
<td>3</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>User languages</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Date start</td>
<td>2011-12-02 00:47:55</td>
<td>2011-12-02 00:03:32</td>
<td></td>
</tr>
<tr>
<td>Date end</td>
<td>2012-04-10 23:40:36</td>
<td>2012-04-10 23:47:55</td>
<td></td>
</tr>
</tbody>
</table>
Tasks

- **Task 1: Sentiment analysis**
  Perform an automatic sentiment analysis to determine the polarity of each text in the test corpus

- **Task 2: Trending topic coverage**
  Identify the topic of the text

- Submission of runs + report
- Groups could participate in one or both tasks
Evaluation

- **Evaluation metrics** (for both tasks)

\[
\text{Precision} = \frac{N (\text{Correct classifications})}{N (\text{all classifications})} \quad \text{Recall} = \frac{N (\text{retrieved documents})}{N (\text{all documents})}
\]

\[
F = (1 + \beta^2) \frac{\text{precision} \cdot \text{recall}}{\beta^2 \cdot \text{precision} + \text{recall}}
\]

- **The gold standard** was generated by
  - pooling all submissions,
  - applying a voting schema
  - an extensive human review of the ambiguous decisions (thousands)

- Both tasks evaluated as an strict single label classification: “success” achieved only when all the test labels are returned
Participants

- 15 groups registered, 9 submitted runs and 8 reports

- Saralegi Urizar, Xabier; San Vicente Roncal, Iñaki. Elhuyar Fundazioa. TASS: Detecting Sentiments in Spanish Tweets.
- Trilla, Alexandre; Alías, Francesc. Grup de Recerca en Tecnologies Mèdia - La Salle, Universitat Ramon Llull. Sentiment Analysis of Twitter messages based on Multinomial Naive Bayes.
- Martín-Wanton, Tamara; Carrillo de Albornoz, Jorge. UNED NLP & IR Group. UNED at TASS 2012: Polarity Classification and Trending Topic System.
- Castellano González, Angel; Cigarrán Recuero, Juan; García Serrano, Ana. UNED. UNED @ TASS: Using IR techniques for topic-based sentiment analysis through divergence models.
- Moreno-Ortiz, Antonio; Pérez-Hernández, Chantal. Universidad de Málaga. Lexicon-Based Sentiment Analysis of Twitter Messages in Spanish.
Results

- **Task 1: Sentiment analysis**
  
  5 levels + NONE

<table>
<thead>
<tr>
<th>Run Id</th>
<th>Group</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>pol-elhuyar-1-5l</td>
<td>Elhuyar Fund.</td>
<td>65.29%</td>
</tr>
<tr>
<td>pol-l2f-1-5l</td>
<td>L2F - INESC</td>
<td>63.37%</td>
</tr>
<tr>
<td>pol-l2f-3-5l</td>
<td>L2F - INESC</td>
<td>63.27%</td>
</tr>
<tr>
<td>pol-l2f-2-5l</td>
<td>L2F - INESC</td>
<td>62.16%</td>
</tr>
<tr>
<td>pol-atrilla-1-5l</td>
<td>La Salle - URL</td>
<td>57.01%</td>
</tr>
<tr>
<td>pol-sinai-4-5l</td>
<td>SINAI - UJAEN</td>
<td>54.68%</td>
</tr>
<tr>
<td>pol-uned1-2-5l</td>
<td>LSI UNED</td>
<td>53.82%</td>
</tr>
<tr>
<td>pol-uned1-1-5l</td>
<td>LSI UNED</td>
<td>52.54%</td>
</tr>
<tr>
<td>pol-uned2-2-5l</td>
<td>LSI UNED 2</td>
<td>40.41%</td>
</tr>
<tr>
<td>pol-uned2-1-5l</td>
<td>LSI UNED 2</td>
<td>39.98%</td>
</tr>
<tr>
<td>pol-uned2-3-5l</td>
<td>LSI UNED 2</td>
<td>39.47%</td>
</tr>
<tr>
<td>pol-uned2-4-5l</td>
<td>LSI UNED 2</td>
<td>38.59%</td>
</tr>
<tr>
<td>pol-imdea-1-5l</td>
<td>IMDEA</td>
<td>36.04%</td>
</tr>
<tr>
<td>pol-sinai-2-5l</td>
<td>SINAI - UJAEN</td>
<td>35.65%</td>
</tr>
<tr>
<td>pol-sinai-1-5l</td>
<td>SINAI - UJAEN</td>
<td>35.28%</td>
</tr>
<tr>
<td>pol-sinai-3-5l</td>
<td>SINAI - UJAEN</td>
<td>34.97%</td>
</tr>
<tr>
<td>pol-uma-1-5l</td>
<td>UMA</td>
<td>16.73%</td>
</tr>
</tbody>
</table>
Results

- **Task 1: Sentiment analysis**

  3 levels + NONE

<table>
<thead>
<tr>
<th>Run Id</th>
<th>Group</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>pol-elhuyar-1-3l</td>
<td>Elhuyar Fund.</td>
<td>71.12%</td>
</tr>
<tr>
<td>pol-l2f-1-3l</td>
<td>L2F - INESC</td>
<td>69.05%</td>
</tr>
<tr>
<td>pol-l2f-3-3l</td>
<td>L2F - INESC</td>
<td>69.04%</td>
</tr>
<tr>
<td>pol-l2f-2-3l</td>
<td>L2F - INESC</td>
<td>67.63%</td>
</tr>
<tr>
<td>pol-atrilla-1-3l</td>
<td>La Salle - URL</td>
<td>61.95%</td>
</tr>
<tr>
<td>pol-sinai-4-3l</td>
<td>SINA - UJAEN</td>
<td>60.63%</td>
</tr>
<tr>
<td>pol-uned1-1-3l</td>
<td>LSI UNED</td>
<td>59.03%</td>
</tr>
<tr>
<td>pol-uned1-2-3l</td>
<td>LSI UNED</td>
<td>58.77%</td>
</tr>
<tr>
<td>pol-uned2-1-3l</td>
<td>LSI UNED 2</td>
<td>50.08%</td>
</tr>
<tr>
<td>pol-imdea-1-3l</td>
<td>IMDEA</td>
<td>45.95%</td>
</tr>
<tr>
<td>pol-uned2-2-3l</td>
<td>LSI UNED 2</td>
<td>43.61%</td>
</tr>
<tr>
<td>pol-uned2-3-3l</td>
<td>LSI UNED 2</td>
<td>41.20%</td>
</tr>
<tr>
<td>pol-uned2-4-3l</td>
<td>LSI UNED 2</td>
<td>40.43%</td>
</tr>
<tr>
<td>pol-uma-1-3l</td>
<td>UMA</td>
<td>37.61%</td>
</tr>
<tr>
<td>pol-sinai-2-3l</td>
<td>SINA - UJAEN</td>
<td>35.83%</td>
</tr>
<tr>
<td>pol-sinai-1-3l</td>
<td>SINA - UJAEN</td>
<td>35.58%</td>
</tr>
<tr>
<td>pol-sinai-3-3l</td>
<td>SINA - UJAEN</td>
<td>35.11%</td>
</tr>
</tbody>
</table>
Results

- **Task 2: Topic detection**

<table>
<thead>
<tr>
<th>Run Id</th>
<th>Group</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>top-l2f-2</td>
<td>L2F - INESC</td>
<td>65.37%</td>
</tr>
<tr>
<td>top-l2f-1y3</td>
<td>L2F - INESC</td>
<td>64.92%</td>
</tr>
<tr>
<td>top-atrilla-1</td>
<td>La Salle - URL</td>
<td>60.16%</td>
</tr>
<tr>
<td>pol-uned2-5a8</td>
<td>LSI UNED 2</td>
<td>45.26%</td>
</tr>
<tr>
<td>top-imdea-1</td>
<td>IMDEA</td>
<td>45.24%</td>
</tr>
<tr>
<td>pol-uned2-9a12</td>
<td>LSI UNED 2</td>
<td>42.24%</td>
</tr>
<tr>
<td>pol-uned2-1a4</td>
<td>LSI UNED 2</td>
<td>40.51%</td>
</tr>
<tr>
<td>top-sinaí-5</td>
<td>SINAÍ - UJAEN</td>
<td>39.37%</td>
</tr>
<tr>
<td>top-sinaí-4</td>
<td>SINAÍ - UJAEN</td>
<td>37.79%</td>
</tr>
<tr>
<td>top-sinaí-2</td>
<td>SINAÍ - UJAEN</td>
<td>34.76%</td>
</tr>
<tr>
<td>top-sinaí-3</td>
<td>SINAÍ - UJAEN</td>
<td>34.06%</td>
</tr>
<tr>
<td>top-sinaí-1</td>
<td>SINAÍ - UJAEN</td>
<td>32.34%</td>
</tr>
<tr>
<td>pol-uned1-1y2</td>
<td>LSI UNED</td>
<td>30.98%</td>
</tr>
</tbody>
</table>
Saralegi Urizar, Xabier; San Vicente Roncal, Iñaki.
Elhuyar Fundazioa

- Supervised approach that includes some linguistic knowledge-based processing using Freeling for preparing the features, comprising:
  - lemmatization and POS tagging
  - tagging of polarity words,
  - treatment of emoticons
  - treatment of negation,
  - weighting of polarity words depending on syntactic nesting level
  - treatment of spell-errors
- Detection of polarity words is done according to a polarity lexicon built in two ways:
  - projection to Spanish of an English lexicon
  - extraction of divergent words of positive and negative tweets of training corpus
Fernández Anta, Antonio; Morere, Philippe; Núñez Chiroque, Luis; Santos, Agustín. Institute IMDEA Networks

- Comparative analysis of different approaches and classification techniques using NPL and supervised machine learning

- The data is preprocessed using basic well-known techniques and tools proposed in the literature

- Then, popular classifiers were used (in particular, all classifiers of WEKA have been evaluated)
Both tasks are considered as classification tasks
Their most successful experiment cast the problem as a binary classification problem
  - Binary classifiers are easier to develop, offer faster convergence ratios, and can be executed in parallel
  - Final results are generated by combining all the different binary classifiers
Specifically, an approach based on logistic regression classification models, which corresponds to the maximum entropy (ME) classification for independent events
Basic preprocessing
Adaptation of a text classification scheme based on Multinomial Naive Bayes (MNB)

- 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
Martín-Wanton, Tamara; Carrillo de Albornoz, Jorge.
UNED NLP & IR Group

- For polarity classification, they propose an emotional concept-based method based on:
  - an affective lexicon to represent the text as the set of emotional meanings it expresses
  - advanced syntactic techniques to identify negations and intensifiers, their scope and their effect on the emotions affected by them
  - a word sense disambiguation algorithm to take the contextual meaning of terms into account

- For topic detection, their system is based on a probabilistic model (Twitter-LDA):
  - each tweet/topic is represented as a ranking of discriminative words
  - a set of events is retrieved based on a probabilistic approach using the training corpus
  - the output is the topic with the highest statistical correlation
Tasks addressed from an Information Retrieval (IR) perspective based on language divergences.

Kullback-Liebler Divergence (KLD) is used to generate both, polarity and topic models.

To improve the accuracy, several approaches are proposed comparing language models that consider:
- the textual content of each tweet
- named entities detected
- adjectives
For task 1, they use a supervised machine learning approach, based on SVM using different features:
  - unigrams
  - emoticons
  - positive and negative words
  - intensity markers

For task 2, SVM is also used along with different bags of words (BoW)
  - The first BoW was obtained using Google AdWordsKeyWordTool (given a term, directly returns the top N related concepts)
  - The second was generated based on the hash tags of the training tweets, per each category
Moreno-Ortiz, Antonio; Pérez-Hernández, Chantal.
Universidad de Málaga

- A lexicon-based approach using Sentitext, a sentiment analysis tool for Spanish.
- Sentitext’s Global Sentiment Value is strongly affected by the number of lexical units available in the text (or the lack of them, rather)
- Confirmation of Sentitext’s tendency to assign middle-of-the-scale ratings, or at least avoid extreme values
  - 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, drawn from their analysis of the average number of polarity lexical segments and Affect Intensity, is that Twitter users employ highly emotional language
Conclusions

- 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 has shown that the sentiment analysis is becoming a trending topic within the information technology field.

- Corpus released to the community.
Future work/Open questions

- More effort in filtering errors and improving the annotation
- Biased corpus
  - Texts in the corpus are quite well-written, so no need for an spellchecker
  - Negation is present in a few cases, so can be practically ignored
- Define more/less levels of polarity
- Study the difference NEUTRAL vs NONE
- Detection of disagreement
- Other evaluation metrics:
  - Error cost matrix
  - Other than P, R, F1
- Polarity at entity level