UNED2 at TASS: Using IR techniques for topic-based sentiment analysis through divergence models

Angel Castellanos, Juan Cigarrán, Ana García-Serrano

UNED2 at TASS 2012

Language Models and IR-based Approach:

- To capture language models based on language divergences
  - Traditional models (TF-IDF) are suitable to capture GENERAL ISSUES not SPECIFIC ONES
  - Models based on divergences take into account class information → Better capture of SPECIFIC ISSUES [1]

- Generation of language models (ML) or polarity as well as for topics

- Polarity and topic identification trough IR approach on the indexed ML.

- Initial step of preprocessing: More important that always! A challenge for Linguistic resource.
UNED2 at TASS 2012

\[ KLD_{pD,pC} = pD(t) \cdot \log \left( \frac{pD(t)}{pC(t)} \right) \]

\( pD(t)/pC(t) \): probability of the presence of \( t \) in set of Documents/tweets or in the rest of them (C)

<table>
<thead>
<tr>
<th>Término</th>
<th>0.024684</th>
</tr>
</thead>
<tbody>
<tr>
<td>garofalo</td>
<td>0.024675</td>
</tr>
<tr>
<td>mafia</td>
<td>0.020647</td>
</tr>
<tr>
<td>narcotrafico</td>
<td>0.019549</td>
</tr>
<tr>
<td>acido</td>
<td>0.014470</td>
</tr>
<tr>
<td>cosco</td>
<td>0.012388</td>
</tr>
<tr>
<td>cocaina</td>
<td>0.011065</td>
</tr>
<tr>
<td>ndrangheta</td>
<td>0.010325</td>
</tr>
<tr>
<td>dinero</td>
<td>0.010292</td>
</tr>
<tr>
<td>empresarios</td>
<td>0.010214</td>
</tr>
</tbody>
</table>

---

UNED2 at TASS 2012

Polarity KLD Language Models (training set)

- Generation of 5/3 LM for the different polarities from the set of tweets of a concrete polarity (and the rest altogether):
  - MP1: using the terms in the content
  - MP2: using only the adjectives (the polarity is highly related to the lexical ways to express it, that is the adjectives at least)
  - MP3: refined MP2 by deleting all related to the neutral/none?

*Label P+: 'Buen día todos! Lo primero mandar un abrazo grande a Miguel y a su familia @libertadmontes Hoy podría ser un día para la grandeza humana.*
UNED2 at TASS 2012

Topic KLD Language Models (training set)

• Generation of 10 LM for the different topics from the set of tweets of a concrete topic (and the rest altogether):
  • MT1: using the terms in the content
  • MT2: using only the named entities

• “Representativity” of the training corpus for the topic?

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

UNED2 at TASS 2012

Some “Technicalities”:

• All the five models are indexed using Solr
  • Normalization of KLD weights

• Retrieval using Lucene with ranking generated by BM25

• The classification of the tweet is the first polarity/topic retrieved
**UNED2 at TASS 2012**

**Preprocesing**

- **Limpieza del contenido de los tweets:** Consistente en: eliminación de caracteres especiales (puntos, comas, etc…) eliminación de palabras vacías y **eliminación de términos propios de Twitter** (menciones, hashtags y retweets).

- **Etiquetado POS de los Tweets:** para identificar las entidades nombradas y los adjetivos presentes.
  - Stilus desarrollada por Daedalus.
  - [http://www.daedalus.es/productos/stilus/](http://www.daedalus.es/productos/stilus/)

**Experiments short description (Opinion)**

### Sentiment Analysis (5 niveles)

<table>
<thead>
<tr>
<th>Run</th>
<th>Index and query</th>
<th>Precisión</th>
</tr>
</thead>
<tbody>
<tr>
<td>TASK1_RUN_01</td>
<td>MP1 (content) and tweet content</td>
<td>0.3998</td>
</tr>
<tr>
<td>TASK1_RUN_02</td>
<td>MP2 (adjectives) and tweet adjectives/NONE</td>
<td>0.4041</td>
</tr>
<tr>
<td>TASK1_RUN_03</td>
<td>MP3 (refined neutral) and tweet adjectives/NONE</td>
<td>0.3947</td>
</tr>
<tr>
<td>TASK1_RUN_04</td>
<td>MP3 and tweet content</td>
<td>0.3859</td>
</tr>
<tr>
<td>Best at TASS</td>
<td></td>
<td><strong>0.6529</strong></td>
</tr>
</tbody>
</table>

### Sentiment Analysis (3 niveles)

<table>
<thead>
<tr>
<th>Run</th>
<th>Index and query</th>
<th>Precisión</th>
</tr>
</thead>
<tbody>
<tr>
<td>TASK1_RUN_01</td>
<td>MP1 (content) and tweet content</td>
<td>0.4043</td>
</tr>
<tr>
<td>TASK1_RUN_02</td>
<td>MP2 (adjectives) and tweet adjectives/NONE</td>
<td>0.4361</td>
</tr>
<tr>
<td>TASK1_RUN_03</td>
<td>MP3 (refined neutral) and tweet adjectives/NONE</td>
<td>0.5008</td>
</tr>
<tr>
<td>TASK1_RUN_04</td>
<td>MP3 and tweet content</td>
<td>0.4120</td>
</tr>
<tr>
<td>Best at TASS</td>
<td></td>
<td><strong>0.7112</strong></td>
</tr>
</tbody>
</table>
Experiments short description (trending topic)

The tweet to be labeled is the query to the retrieval on the index MT1 (all the content)/ MT2 (named entities).

### Trending Topic Coverage

<table>
<thead>
<tr>
<th>Run</th>
<th>Description</th>
<th>Precisión</th>
</tr>
</thead>
<tbody>
<tr>
<td>TASK2.RUN.01-04</td>
<td>MT1 (content) and tweet content</td>
<td>0.4051</td>
</tr>
<tr>
<td>TASK2.RUN.05-08</td>
<td>MT2 (NE) and tweet NEs/OTHER</td>
<td>0.4526</td>
</tr>
<tr>
<td>TASK2.RUN.09-12</td>
<td>MT2 (NE) and content of tweet without NEs</td>
<td>0.4224</td>
</tr>
<tr>
<td>Best at TASS</td>
<td></td>
<td>0.6537</td>
</tr>
</tbody>
</table>

UNED2 at TASS 2012

Experiments Results (Opinion)

• Polarity signals/Sentiment adjectives:
  • El uso de adjetivos en la generación del modelo (RUN 02, 03 y 04) obtiene mejores resultados que el uso del contenido completo de los tweets (RUN 01)
  • Adjectives of the training set only (use of resource?)

• NONE/NEUTRAL challenge