Sentiment Analysis on Monolingual, Multilingual and Code-Switching Twitter Corpora

David Vilares, Miguel A. Alonso and Carlos Gómez-Rodríguez
{david.vilares, miguel.alonso, carlos.gomez}@udc.es
What is Twitter about? It is mostly about trends!
The problem

Around 51% of the tweets are written in English

She looked shockin. The only person who can pull off crazy outfits like that is the legend Ms Lady GaGa

but trends might be global (Spanish, Portuguese, ...)

Suena #BadRomance y no puedo ponerme a cantar... args ¡Q temaso @ladygaga!

lady gaga ja nao basta ser maravilhosa ainda tem um mapa astral maravilhoso tambem
The problem

and also exists code-switching tweets:

I hate Katty Perra. Carece de toda la gracia de mi.adorada @ladygaga

Translated from Spanish by Bing
I hate Katty bitch. It lacks all the grace of mi.adorada @ladygaga

Multilingual sentiment analysis becomes a real need!
In this study we evaluate...

Different SA approaches to deal with:

- monolingual texts.
- multilingual texts.
- code-switching texts.

Relying on widely used supervised models, sets of features, ...

Restricting the study to English (EN) and Spanish (ES).
1) Monolingual approach:
Methodology: Approaches

2) Pipeline approach:
Methodology: Approaches

3) Multilingual approach:
Methodology: Training the classifiers

We use a squared-regularised logistic regression:

- EN-model trained on a EN corpus.
- ES-model trained on a ES corpus.
- EN-ES model trained on the union of the EN and ES corpora.

Features:

- Words, lemmas, psychometric properties, part-of-speech tags.
- Contextual features: generalised dependency triplets, ...
Methodology: Training the classifiers

Training the EN-ES model requires a bilingual tagger and parser.

How to train dependency parsers for multilingual and code-switching texts?

David Vilares, Miguel A. Alonso and Carlos Gómez-Rodríguez, One model, two languages: training bilingual parsers with harmonized treebanks, in arXiv:1507.08449 [cs.CL], 2015.
Methodology: Corpora

Three ‘environments’ are considered for evaluation:

- **Monolingual corpora:**
  - TASS 2014: Spanish corpus.

- **EN-ES (synthetic) corpus:** TASS + SEMEVAL.

- **EN-ES Code-Switching corpus.** **NEW!**
Methodology: EN-ES Code-Switching corpus

First code-switching Twitter corpus with sentiment labels.

3,062 tweets manually labelled:

- 3 annotators.
- Originally annotated using the SentiStrength scheme.
- Transformed to trinary scale (positive, negative and objective).

31.45% positive, 25.67% negative and 42.88% objective.

EN is predominant.

Freely available.
Experimental results: EN monolingual corpus

Accuracy

- Words (W)
- Lemmas (L)
- Psychometric (P)
- Triplets of L
- L+P

Languages:
- en
- pipe
- en-es
Experimental results: ES monolingual corpora

<table>
<thead>
<tr>
<th>Words (W)</th>
<th>Lemmas (L)</th>
<th>Psychometric (P)</th>
<th>Triplets of L</th>
<th>L+P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>Accuracy</td>
<td>Accuracy</td>
<td>Accuracy</td>
<td>Accuracy</td>
</tr>
</tbody>
</table>

The chart shows the accuracy of different features across different corpora. The features are Words (W), Lemmas (L), Psychometric (P), Triplets of L, and L+P. The corpora are es, pipe, and en-es.
Experimental results: Multilingual corpus

English is the majority language

Accuracy

<table>
<thead>
<tr>
<th>Words (W)</th>
<th>Lemmas (L)</th>
<th>Psychometric (P)</th>
<th>Triplets of L</th>
<th>L+P</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>en-es</td>
<td>Pipe</td>
<td>Eh-es</td>
<td></td>
</tr>
</tbody>
</table>
Experimental results: Multilingual corpus

Spanish is the majority language

- Words (W)
- Lemmas (L)
- Psychometric (P)
- Triplets of L
- L+P

Accuracy

Pipe

en-es
Experimental results: Code-switching corpus

![Experimental results graph]

Accuracy

Words (W)  Lemmas (L)  Psychometric (P)  Triplets of L  L+P

en  es  pipe  en-es
Conclusions

We proposed a multilingual model:

- No need of any language detection or machine translation step.
- Very robust on monolingual, multilingual and code-switching texts.

The first code-switching sentiment corpus is available:

- You can also ask any of the authors.
Conclusions

Monolingual evaluation
- Expected: EN-model gets best results for the EN corpus.
- Unexpected: EN-ES model gets the best results on the ES corpus.

Multilingual evaluation:
- Unexpected: EN-ES obtain the best performance when the majority language is ES.

Code-switching evaluation:
- Expected: ES and the pipeline model obtain low performance.
- Unexpected: EN-model obtain results close to EN-ES model.
Thank you.