SENTIMENT ANALYSIS FOR NEWS

COMPARISON OF MODELING STRATEGIES AND LINGUISTIC APPROACHES

Tània Arnau Serra

24 d'Octubre de 2013

Table of contents

Introduction

- Text Mining
- Natural language processing (NLP)
- Sentiment analysis

Methods for language processing

- Corpus
- Initial dataset
  - Publication, edition, section, page, date…
  - Gold standard.
- Scenarios

Example

El problema no es que los mayas se equivocaran de fecha al pronosticar el fin del mundo tal como lo conocíamos hasta ahora. El problema es que ninguno de ellos llegó siquiera a imaginar que el Armagedón definitivo llegaría en realidad un día como hoy, 7 de enero de 2013, y que vendría de la mano de ese hiperdinámico ente mediático en constante necesidad de merecer que es Leticia Sabater. ***Sí, amigos y amigas del telerreciclaje a fondo perdido: si nadie lo remedia, la polifacética creadora del Leti-Rap volverá esta misma tarde al Sálvame (y nunca mejor dicho) de Telecinco dispuesta a poner en busca y captura, y mirada periférica mediante, a cuanto maromo con pinta de noviete de usar y tirar se cruce en su camino.*** *Así que vayan preparándose, porque apocalíptica o no, tan temible reaparición se promete especialmente escatológica...*

Transformations

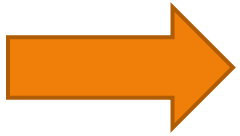
- Convert to lower case
- Remove numbers
- Remove punctuation
- Remove stopwords
- Eliminating extra whitespace

Words of one news

News 1				
Amigos	amigos	telereciclaje	fondo	perdido
Remedia	polifacética	creadora	tira	volverá
Tarde	sálvame	**telecinco**	dispuesta	busca
Captura	mirada	periférica	maromo	pinta
Noviete	usar	tirar	cruce	camino

All words

News 1				
Amigos	amigos	telereciclaje	fondo	perdido
Remedia	polifacética	creadora	tira	volverá
Tarde	sálvame	**telecinco**	dispuesta	busca
Captura	mirada	periférica	maromo	pinta
Noviete	usar	tirar	cruce	camino



All News
Amigos
…
…
…
…
…
…
…
…
…

Scenarios

- Scenario 1 : classify_polarity.
- Scenario 2 : the most frequents words.
- Scenario 3: the most relevants words.
- Scenario 4: the most frequents and the most relevants.

Scenarios

- Scenario 1 : classify_polarity
- Scenario 2 : the most frequents words.
- Scenario 3: the most words.
- **Scenario 4: the most frequents and the most relevants.**

Objectives

- To develop a classifier to evaluate the positivity of news about a particular brand.
  - To compare the efficiency between the classification obtained when the corpus is external or internal.
  - To compare models that use internal corpus using different lexicons.
  - To compare different statistical models.

Statistical methods

- Linear model
- Naive Bayes classifier
- Tree based methods

- Performance assessment

Performance assessment

- Internal validation
- External validation

Results

Internal validation:

Models	Linear model	Naive Bayes	Classification Tree
	89.27%	75.7%	73.32%

External validation:

Models	Linear model	Naive Bayes	Classification Tree
	77.28%	75.86%	60.88%

Conclusions

- The classification obtained when **the corpus is internal** is better than when is external.
- The better lexicons are taking into account **the more relevant and frequently words.**
- The **Naive Bayes classifier** and the **linear model** are the best models because resists the external validation.
- The classification tree gives a generous results in internal validation due to over-fitting.

Discussion

- Improve the models: (Neural Networks, k-nearest neighbours…)
- Improve qualitative approach.