



MediaMiser

Cross-lingual sentiment analysis

David Nadeau, CTO

@pythonner



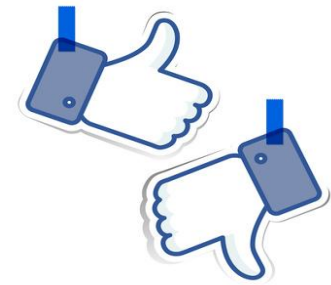
Problem

Bla bla bla bla
bla bla bla bla.



Bla bla bla ?

Bla bla !

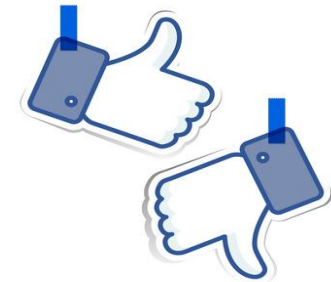


Blah blah blah
blah blah.



Blah blah?

Blah!



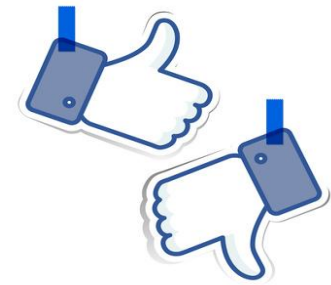


Solution 1

Bla bla bla bla
bla bla bla bla.

Bla bla bla ?

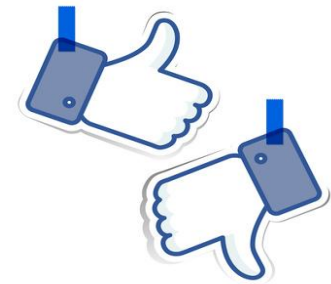
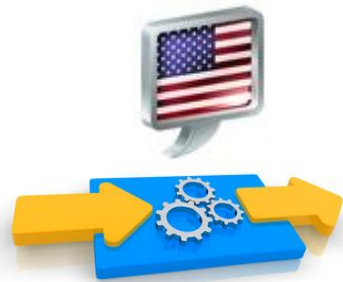
Bla !



Blah blah blah
blah blah.

Blah blah?

Blah!





Solution 2



Bla bla bla bla
bla bla bla bla.

Bla bla bla ?

Bla bla !

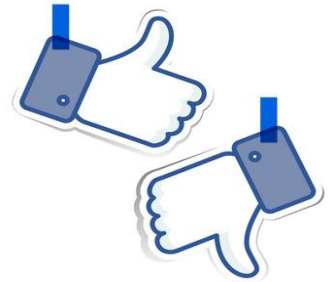
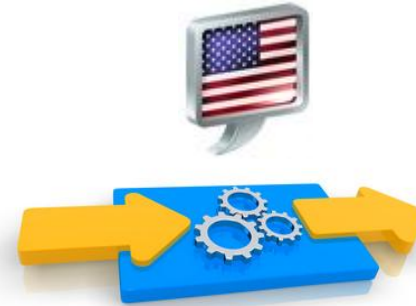
Machine
Translation



Blah blah blah
blah blah.

Blah blah?

Blah!





Experiments

1. Cross-lingual baseline:

English sentiment analysis on **French** texts

2. Multilingual experiment (solution 1):

Basic **French** sentiment analysis on **French** texts

3. Cross-lingual experiment (solution 2):

English sentiment analysis on **French** machine translated texts



Results

- Machine translation approach (solution 2) outperforms *
- MT system choice has low impact! (Google, Bing, NRC)
- Combining approaches is even better



Sweet, extra time!

Naive Bayes
Decision Trees
Random Forests

US Patent 7,860,872
CDN Patent 2,635,420

MediaMiser's system is **trainable, ensemble-based** and has **language-specific** routines.

Tokenization
Sentence splitting
Negation handling

Bag-of-words
Sentiment lexicons
Ontological data
Meta data