LIA@RepLab 2013

J.-V. Cossu[1], B. Bigot[1], L. Bonnefoy[1,2], M. Morchid[1], X. Bost[1], G. Senay[1], R. Dufour[1],
V. Bouvier[2,3], J.-M. Torres-Moreno[1], M. El-Bèze[1]

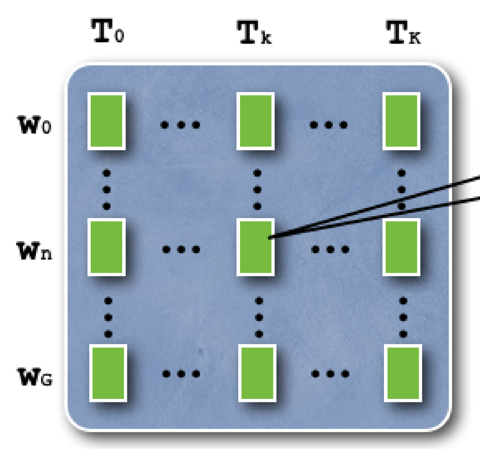
[1] - LIA, University of Avignon, France [2] - iSmart, Aix-en-Provence, France [3] - LSIS, University of Aix-Marseille, France

Abstract

This paper presents the participation of the Computer Science Laboratory of Avignon (LIA) to RepLab 2013. We first present our most efficient systems, then a view of all our contributions to the challenging task of online Reputation analysis based on the extraction of the information conveyed in tweets (i.e. content and metadata). We aimed at participating as a laboratory and we have applied several methods derived from different domains (IR, spoken document retrieval). As the methods presented below rely on very different approaches, we have also checked how combining system outputs by the use of merging algorithms could improve the performances according to different metrics.

Most efficient methods

- Maximum a Posteriori Feature Selection (Topic Detection)
  - Used previously for topic detection in conversations
  - 3 steps:
    1. Detection of the headwords (HW) in the Tweets related to one specific topic
    2. Selection using a Maximum A Posteriori Probability (MAP) estimator
    3. Ranking considering a purity criterion



$$p(T_k|w_n) = \frac{c(w_n, T_k) + Kp(T_n)}{\sum_{j=1}^{K} c(w_n, T_j) + K}$$

$$? \quad \hat{T} = \arg\max_k \left[H^W(T_k|Y_d) \right]$$

With w_n the headword n (word, bigram or distance bigram) and T_k the topic k.

- k-NN with discriminant features (Filtering, Polarity and Topic-Priority)
  - A very improved version of the baseline
  - Best results for Filtering, Polarity and Topic-Priority tasks
  - Similarity using the Jaccard measure on a discriminant representation of bag-of-words
  - Use of metadata

Other methods

- Adaptation of the LIA@KBA2012 (Filtering and Topic-Priority)
  - Captures intrinsic characteristics of highly relevant documents
  - Document centric features, entity's profile features, and time features
  - One classifier for each category of entities
- Continuous Context Models (Filtering, Polarity and Topic-Priority)
  - Previously used for spoken name recognition in speech
  - Probabilistic model of the positional and lexical dependencies between a word and its context
  - Context vectors are built around anchor words (i.e. hashtags, "@'s usernames")
  - One model for each entity
- TF-IDF-Gini approach with a SVM classifier (Polarity and Topic-Priority)
- Boosting classification approach (Polarity and Topic-Priority)
- Cosine with TF-IDF and Gini purity criteria (Filtering, Polarity and Topic-Priority)
- Ultrastemming + n-grams (Filtering)

Merging algorithms

For Filtering, Polarity and Priority, our systems have been merged using the 2 following methods.

- ELECTRE
  - Ranks the entity labels with regards to how a label dominates another one
- PROMETHEE
  - Compares several alternative of actions taken by pair
  - Measures the capacity of an entity label to dominate or being dominated

Results

# Method	Accuracy	F-Measure
k-NN	.8720	**.3819**
KBA 2012	.8764	.3412
Baseline	*.8714*	*.3255*
Linear combination	**.8827**	.3127
ELECTRE	.8792	.3024
PROMETHEE	.8745	.2962
Cosine	.8351	.2720
Median	*.8260*	*.2655*
Ultra-stemming	.8067	.1870
CCM	.8000	.1265

Submitted systems to Filtering Task

#Method	Accuracy	F-Measure
k-NN	.6275	**.3351**
Baseline	*.6007*	*.2965*
KBA 2012	.5858	.2820
Boosting	.6405	.2680
Cosine	.6167	.2657
ELECTRE	.6514	.2530
PROMETHEE	.6470	.2513
Linear combination	**.6527**	.2510
Median	*.5734*	*.2496*
SVM	.5758	.1457
CCM	.5424	.1367

Submitted systems to Priority Task

# Method	Reliability	Sensitivity	F-Measure
MAP Feature Selection	.2187	**.3468**	**.2463**
Median	**.3659**	*.2180*	*.1954*
Baseline	*.1525*	*.2173*	*.1735*

Best submitted runs to Topic Detection Task compared to Median and Baseline

Conclusion

- A large variety of approaches and performances
- Several systems combined in order to benefit from this diversity
- **Perspectives:**
  - Integration of the metrics in the merging algorithms
  - Possible improvement by considering a subset of systems selected before the merging step