Constraint-based Reputation Modeling on microblogs
ML CIMI Toulouse WorkShop 2

Jean Valère Cossu, Eric SanJuan, Juan Manuel Torres Moreno, Marc El Bèze

Universit d'Avignon

eric.sanjuan@univ-avignon.fr

November 10, 2015

Summary

Problematic

Given an entity (public figure), a set of pre-defined attributes, a stream of public Micro-Blogs we seek to model their impact on two objective measures:

1. External public Opinion expressed about the entity,
2. Internal Stimuli Priority,

Challenge

Dimensions to explain reputation and recommend actions rely on anonymous analyst background. They are biased, vague, approximative, inconsistent BUT expensive.

Approach

Problem

Investigate the issue of modeling corporate entities' online reputation.

Method

Optimal Bayesian Network among latent variables defined by analyst.

Tool

Partial Least Square Path Modeling combining Multivariate Factor Analysis with Simulated annealing.

Solution

1. Start from a small pool of Micro-Blogs annotated by experts for Priority or current blog readers for Opinion,
2. Apply a Learn-To-Rank procedure to score Dimensions, Opinion and Priority on all available Micro-Blogs content.
3. Build an optimal interaction model between Dimensions and objective measures.

Experimental setup

Input: small set of annotations

Dimensions to analyze e-Reputation are set up by analyst based on a small set of manual annotations.

Enrichment: automatic annotated MicroBlogs

Simple but diverse Machine Learning (ML) Natural Language Processing (NLP) approaches are used to generalize these annotations.

Evaluation

Domains

1. Automotive,
2. Banking,
3. Music
4. University

Feature

Ranking of dimensions by decreasing impact over opinion or priority

Gold Standard

CLEF RepLab 2012-2014 dataset

Replab Dataset

Annotated Micro Blogs

- 61 entities (DbPedia entries: automakers, banks, artist and universities)
- 33 952 twitts queried based on these keywords
- Dimensions (7 tags with no overlap)
- Polarity (3 levels: Negative, Neutral, Positive)
- Priority (3 levels: Alert, Important, Non important)

Constrained Latent variables

7 Dimensions

- Performance: *long term business success and financial soundness of the company. Goldman Profit Rises but Revenue Falls: Goldman Sachs reported a second-quarter profit of $1.05 billion*
- Products& Services: *BMW To Launch M3 and M5 In Matte Colors: Red, Blue, White but no black...*
- Leadership:*Goldman Sachs estimates the gross margin on ACI software to be 95% ...*
- Citizenship, responsibility, including ethical aspects of business: integrity, transparency and accountability: *Find out more about Santander Universities scholarships, grants, awards and SME Internship Programme*

Constrained Latent variables

7 Dimensions

- Governance Related to the relationship between the company and the public authorities: *Judge orders Barclays to reveal names of 208 staff linked to Libor probe via Telegraph*
- Workplace Related to the working environment and the companys ability to attract, form and keep talented and highly qualified people.
- Innovation, novel ideas incorporated into products *Eddy Merckx Cycles announced a partnership with Lexus to develop their ETT Hme trial bike.*

Scoring Protocol

n+1 binary classifiers

- Cosine distance $\times$ multiple term weighting functions,
- Jaccard index $\times$ multiple term weighting functions,
- Linear SVM (keep it simple)
- k-NN $\times$ multiple values of k
- Word2Vect $\times$ different distances

Multidimensional numerical sphere

Dimensioon:

$$\#\text{Dimensions} \times \#\text{Classifiers}$$

$$+\#\text{Opinion or Priority} \times \#\text{Classifiers}$$

Normalization

Each score provided by kNN, Cosinus and Jaccard has been normalized using the sum of all scores related to a micro blog.

External Model

Latent Variables entailment

Given a trend Ω of Micro-Blogs texts, the analyst provides a family $\mathcal{S} = S_1, ..., S_D$ of subsets of Ω, each representing a Dimension, an Opinion or an *"Alert"* that we shall model as a latent variable.

Training

S_i are used to train a sequence of K independent classifiers over each set that will score the possibility for a new Micro-Blog text to be included by the analyst in some S_i.

Normalization

Given D sets of K scoring normalized functions such as:

1. $f_{i,k} : \omega \in \Omega \mapsto f(\omega) \in [0, 1]$
2. $\omega \in S_i \Rightarrow (\forall 1 \leq k \leq D) f_{i,k}(\omega) = 1$

Each normalized scoring function $f_{i,k}$ defines a discrete smoothed probability function $P_{i,k}$ over Ω defined by:

$$P_{i,k}(\{\omega\}) = (1 - \lambda) f_{i,k}(\omega) + \lambda E(f_{i,k})$$

where λ is a smoothing parameter in $[0, 1]$ and $E(f_{i,k})$ is the expectation of $f_{i,k}$ over a trend of T Micro-Blogs:

$$\frac{1}{|T|} \sum_{\omega \in T} f_{i,k}(\omega)$$

External Latent Model

Bayesian Network

For each latent variable S_i there exists: $\alpha_i = (\alpha_{i,1}, ..., \alpha_{i,k})$, $\alpha_{i,k} > 0$ such that the probability P_i for a Micro-Blogs ω to be associated by the analyst with S_i verifies:

$$P_i(\{\omega\}) = \prod_{k=1}^{K} P_{i,k}(\{\omega\})^{\alpha_{i,k}}$$

Internal Latent Model

Bayesian Network

For each case we look for a partition D_1, D_2 and two vectors $\beta(x) = (\beta_1, ..., \beta_{n_x})$ of dimensions maximizing:

$$P_X(\{\omega\}) = \prod_{j=1}^{x_n} P_j(\{\omega\})^{\beta_j}$$

Latent variables

Dimensions

Products and services, Citizenship, Governance, Innovation, Leadership, Performance, Workplace

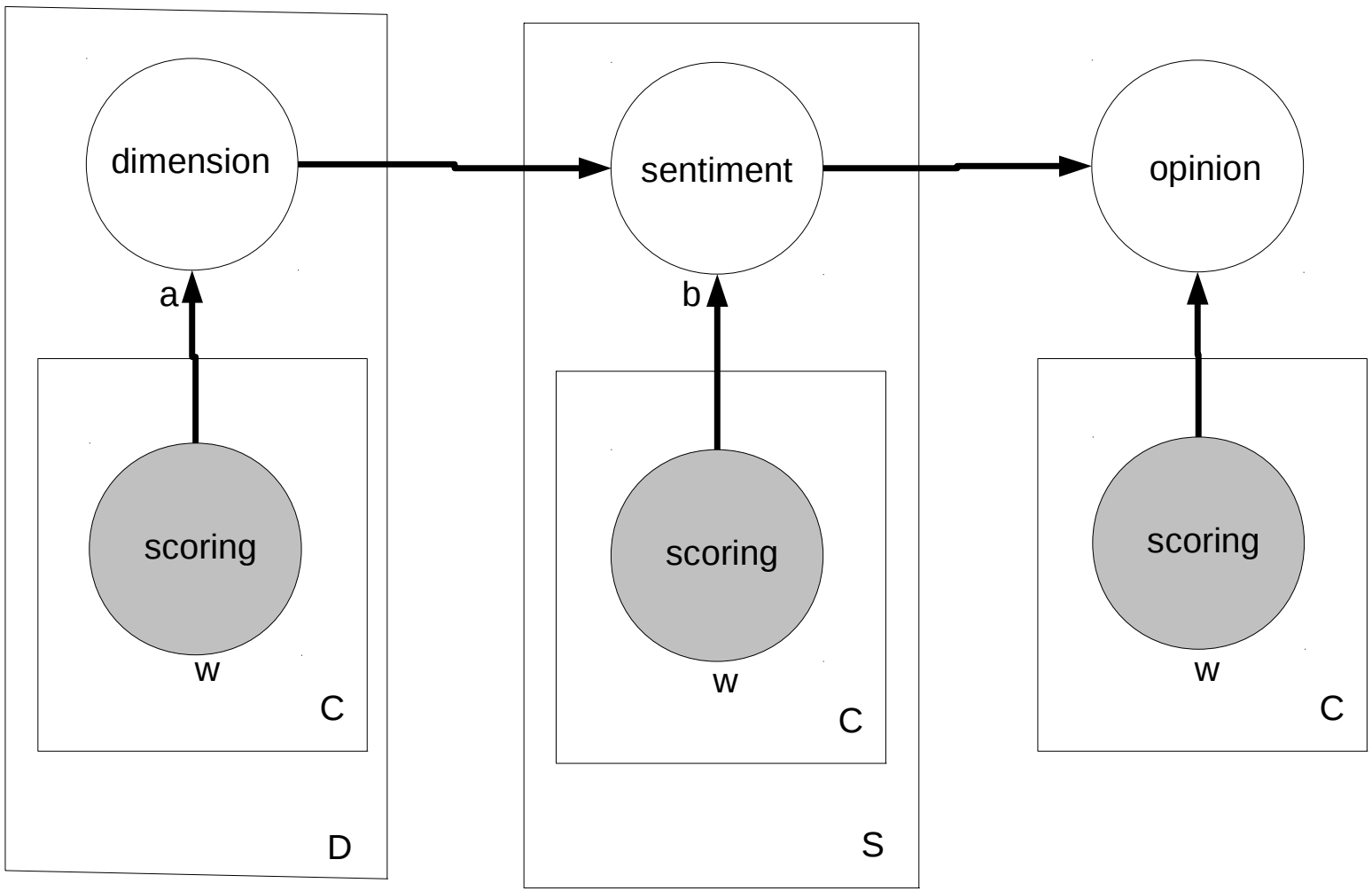
Target: Opinion or Priority

- Positive, Neutral, Negative
- Alert, Important, Non-Important

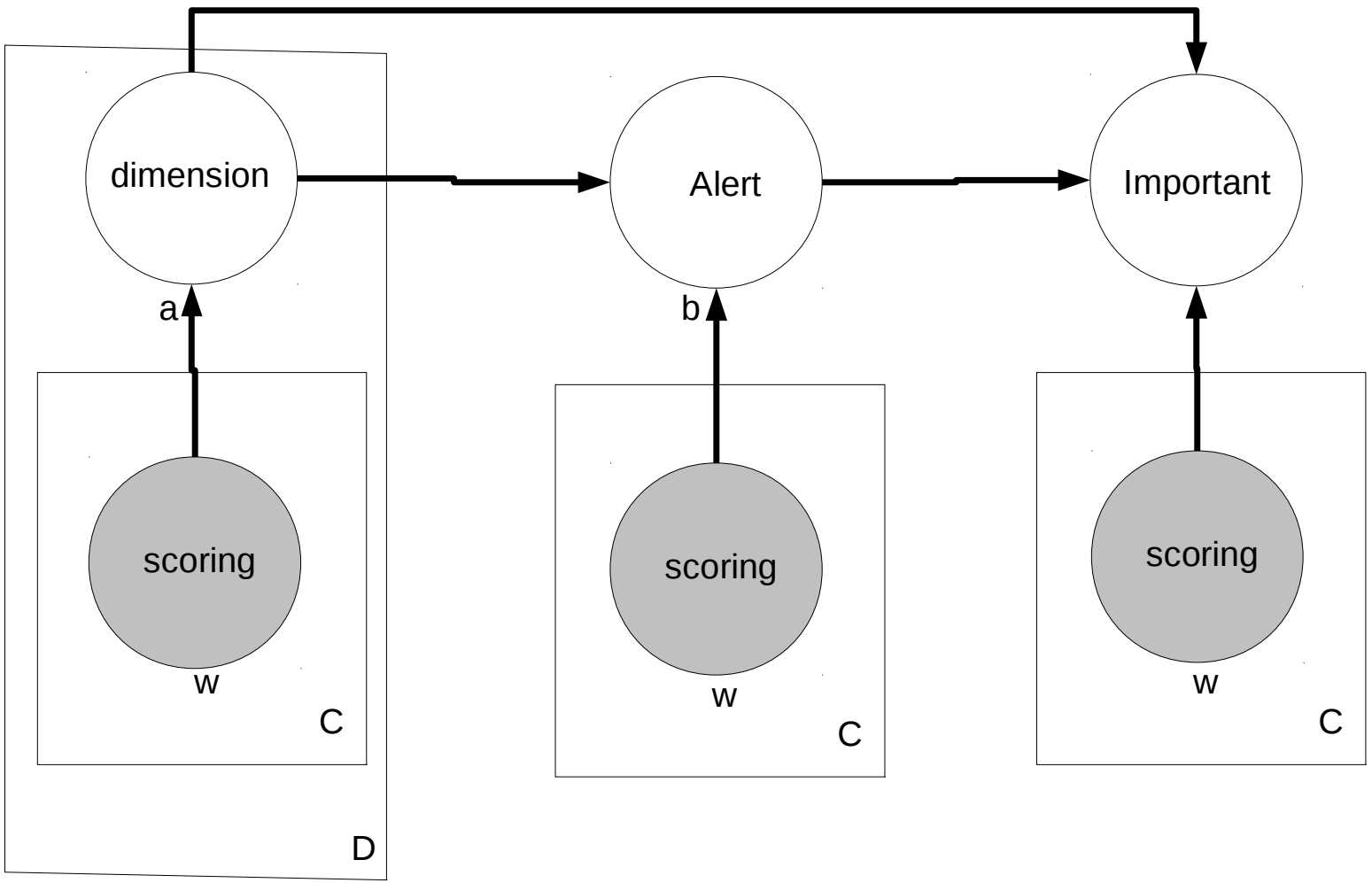
Smoothed probabilities based on scoring functions

$$P_i(\{\omega\}) = \prod_{k=1}^{K} P_{i,k}(\{\omega\})^{\alpha_{i,k}}; P_X(\{\omega\}) = \prod_{j=1}^{x_n} P_j(\{\omega\})^{\beta_j}$$

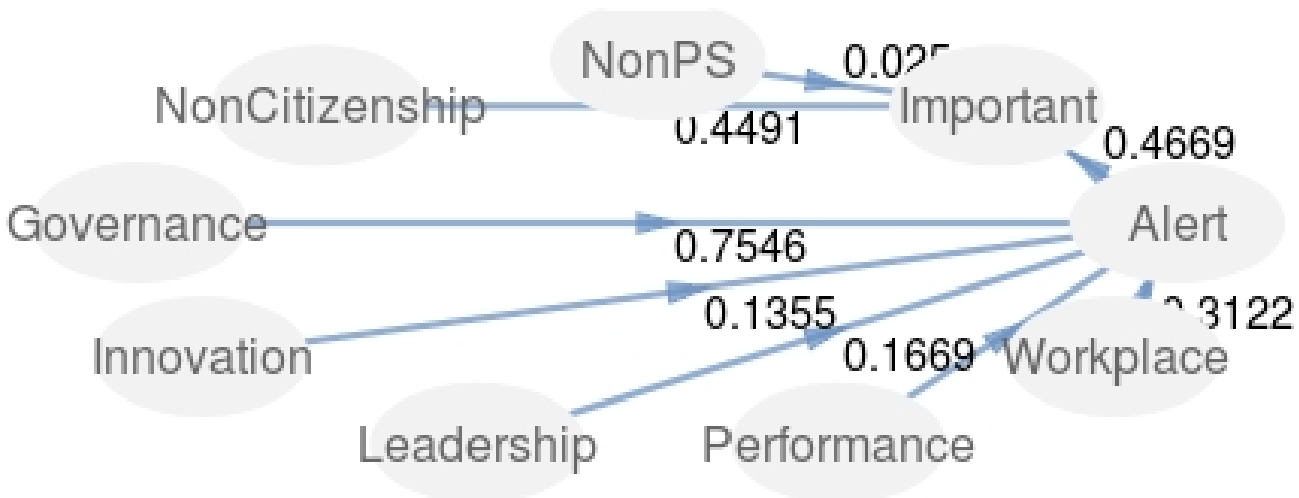
Path Opinion model



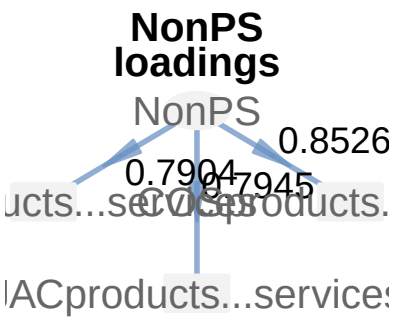
Path Alert model

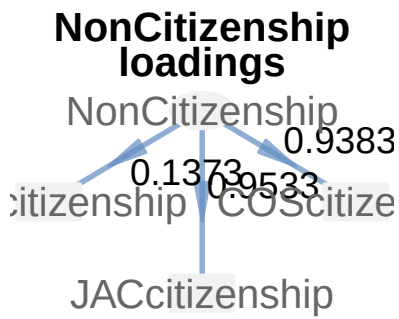


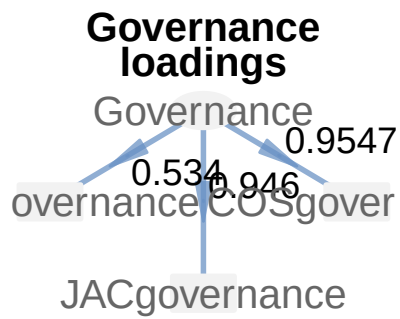
Results on Alert

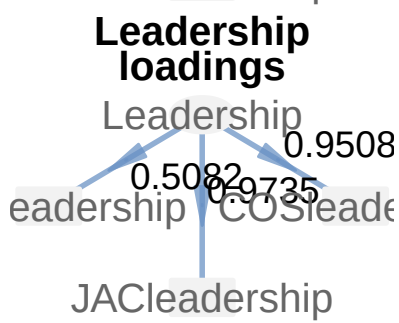


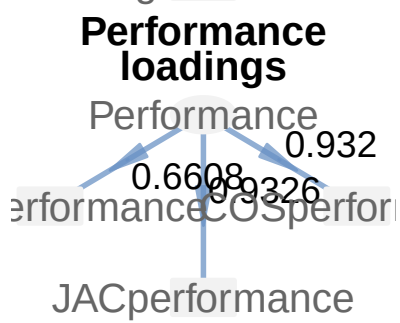
Alert Internal Model Results

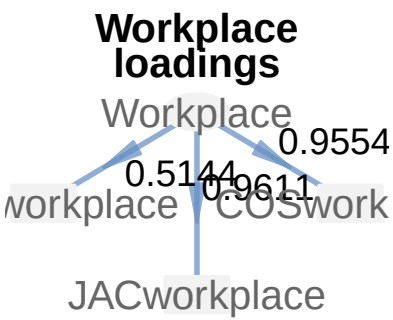


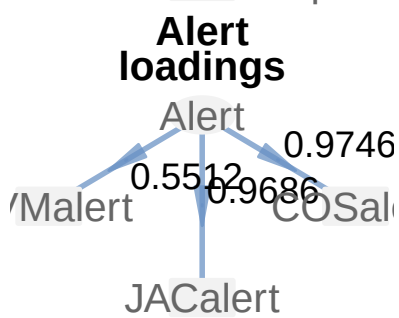


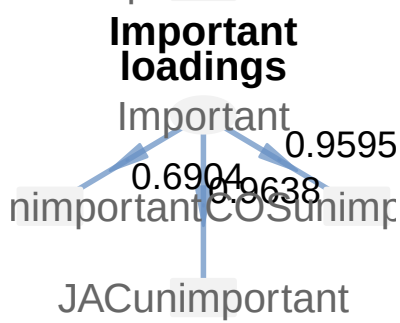












Evaluation

Methodology

Ability of our probabilistic path model to rank Dimensions by decreasing impact over Opinion or Priority.

- Pearson's product-moment correlation $= 0.9759958$, p-value $= 0.0008574$
- Kendall's rank correlation $\tau = 1$, p-value $= 0.002778$

Conclusion

- Supervised path modeling algorithm to explains Opinion or Priority scores based on selected Concepts.
- Robustness of the resulting model has been evaluated over the multilingual CLEF RepLab dataset.

References

- Multi-dimensional Reputation Modeling Using Micro-blog Contents. ISMIS 2015
- Automatic Classification and PLS-PM Modeling for Profiling Reputation of Corporate Entities on Twitter. NLDB 2015
- Bilingual and Cross Domain Politics Analysis. Research in Computing Science 85: 9-19 (2014)

Thank you