Detecting Real-World Influence Through Twitter

Jean-Valère Cossu
NLP team @ University of Avignon, France

ENIC 2015

22 September 2015

Participants

- **J.-V. Cossu***, N. Dugue** and V. Labatut*
  * Université d'Avignon
  ** Université d'Orléans

Get started with RepLab
Online Reputation Monitoring on Twitter

What ?

- Assist brand managers in their daily work

How ?

- Learn experts behaviour about entities concerns
- Automatically propagate these assessments
- Justify the hypothesis

3 years of RepLab

2012

- Monitoring in unknown entity case
- New challenge, is it possible to compute the issue ?
- Features discovery

2013

- Machine Learning
- Messages selection/ranking
- Raises new issues

2014

- **User profiling**
- Age, gender, occupation, influence, traits and so on

RepLab: Author Profiling

Data

- Large profiles collection (7,000 profiles manually labelled)
- 2 economic domains Automotive, Banking
- Each profile comprises the last 600 posted tweets

Main objectives

- **Identify Opinion-Makers**
  70% are not influential - Search/IR
- Categorize profile according to their activity
- See PAN for gender and age identification

Piotr's profile

- Is Piotr an influent Twitter user (in his community-domain) ?
- What about Gender ? Age ? Traits ?

RepLab Author Profiling

- (Social) Network approaches
- Content-based approaches

You are what your network is

Features

- Public profile (description and personal data)
- (S)N features (followers, followees, etc.)
- Writing behaviour (retweets, hashtag, links)
- External data (Klout, Kred,Google)

Features issues

- Features names
- Features relevance to the problem
- Data crawling, impossible to collect the complete network

You are what you say

Profile representation

- A profile is a **Bag of Tweets** where each tweet is processed
  Tweets can be bag-of-words or set of features
- A profile is a vocabulary or a set of features (**User as Document**)

Related content-based features

- Bag of words binary representation (word presence)
- Tweet length, Special characters, Hashtags, Links

Extra-features

- Part-of-speech tagging, Named entities
- Tweet enrichment

You are what you say (2)

Proposed content-based features

- Bag of N-grams with TF, IDF
- Purity index (word distribution through classes) :

$$G(i) = \sum_{c \in \mathbb{C}} \mathbb{P}^2(i|c) = \sum_{c \in \mathbb{C}} \left(\frac{DF_i(c)}{DF_{\mathbb{T}}(i)} \right)^2 \quad (1)$$

Strategies

- Monolingual and domain specific model VS global model
- Majority vote over the profile

What about performances ?

Is tweet content sufficient ?

- Cosine distance between document and influents class' vocabulary
- No specific training, parameters tuning, or features selection

Author ranking performances

System	Automotive	Banking	Avg MAP
Cosine	**.803**	.626	**.714**
LIA_Participation	.764	**.652**	.708
REPLAB1	.721	.410	.565
Baseline	.370	.385	.378
Klout	.304	.275	.289

- LIA_Participation is a manually tuned KNN* (not replicable)
- REPLAB1 used hot topics information
- Baseline rank according to the Followers number
- Raw features (Tweets, followees etc.) are under Klout

What about performances ?

Influence detection classification

System	Automotive	Banking	Avg F
Cosine	**.833**	*.751*	**.792**
LIA_new	.702	.726	.714
Best_System	.696	.693	.694
Baseline	.500	.500	.500

- Same configuration ... same results

Wrap up !

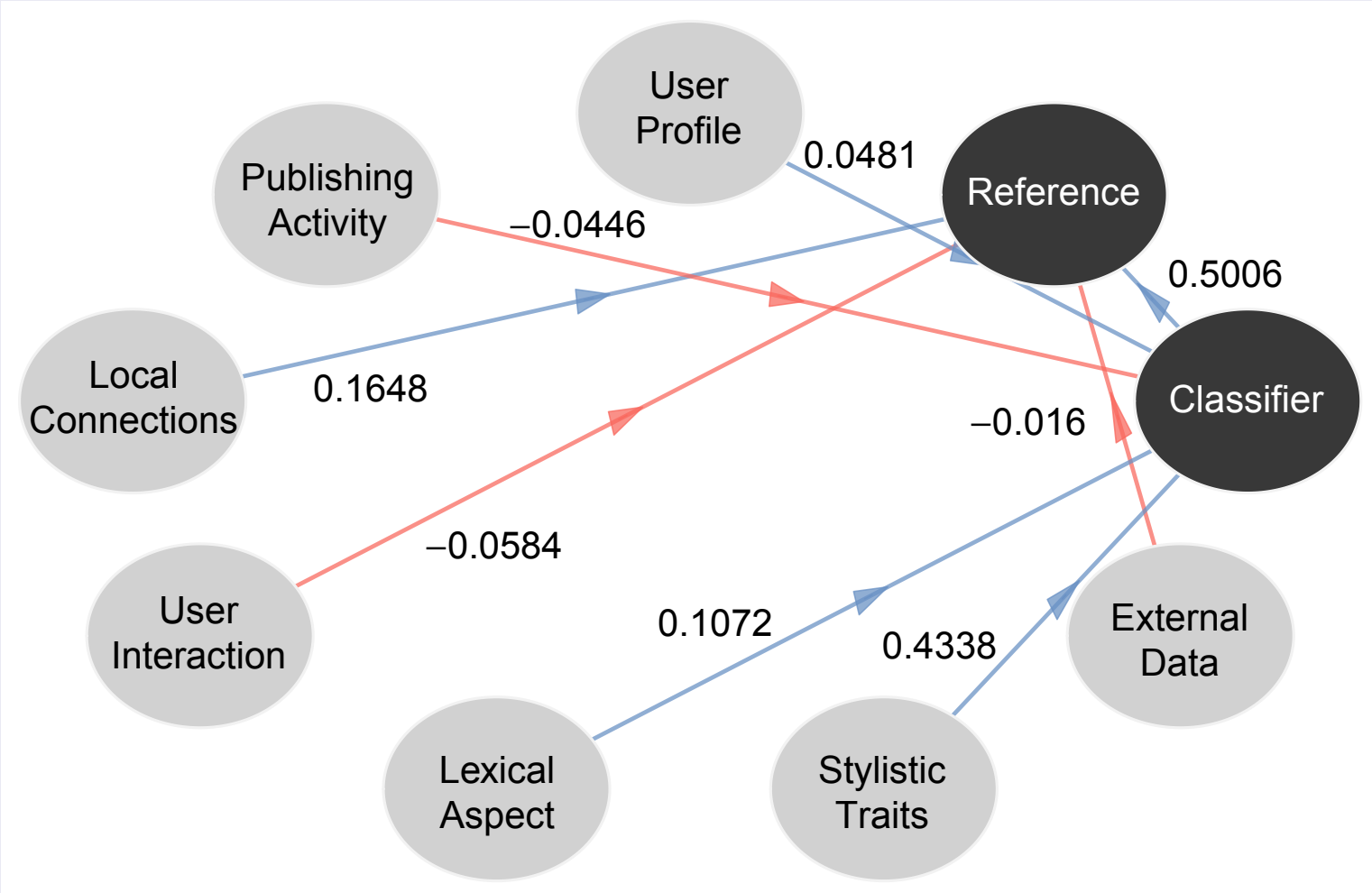
What we have done here ?

- **Distinguish influencers using their account characteristics**
- **Distinguish influencers using contents they produce**
- Ranking them
- Real influence is not node importance
  ** Results limited to this dataset and the annotation quality*

What's next ?

- Profile summarization in progress
  How much we need ?
  Does it work with only 50 tweets ?
- Visual modelling

Features behaviour regarding influence for banking



End

Thank you !

Contact:

- jvcossu@gmail.com
- www.jeanvalerecossu.fr

References

- **NLP-based classifiers to generalize experts assessments in E-Reputation monitoring**
  Cossu J-V., Ferreira E., Gaillard J., Janod K. and El-Bèze M. @CLEF 2015
- **Automatic Classification and PLS-PM Modeling for Profiling Reputation of Corporate Entities on Twitter**
  Cossu J-V., SanJuan E., Torres J-M. and El-Bèze M. @NLDB 2015
- **Overview of the 3rd Author Profiling task at PAN 2015.**
  Rangel F., Rosso P., Potthast M., Stein B., and Daelemans W., D. @CLEF 2015
- **Overview of RepLab 2014: Evaluating Online Reputation Monitoring Systems.**
  Amigó, E. and Carrillo de Albornoz, J. and Chugur, I. and Corujo, A. and Gonzalo, J. and Martín, T. and Meij, E. and de Rijke, M. and Spina, D. @CLEF 2014