LIA RepLab systems
Author Profiling & Reputation Dimensions

J.V. Cossu
University of Avignon, France

CLEF - RepLab 2014

17 September 2014

Participants

- **J.-V. Cossu**, K. Janod, E. Ferreira, J. Gaillard and M. El-Bèze

Plan

- Introduction
- Profiling
  - Categorization sub-task
  - Small overview of Ranking sub-task
- Dimensions task
- Conclusion

Introduction : LIA participation to RepLab 2014

Idea

- Adaptation of LIA systems from speech recognition
- Light classifier adaptation to Ranking and Profiling
- Monolingual VS multilingual approaches
- Global VS domain-specific approaches
- Homogeneity rules

Participation to each sub-tasks

- Author Ranking (3 monolingual systems - 1 run)
- Author Categorization (4 systems - 3 runs)
- Reputation Dimensions (5 systems - 5 runs)
- Combination using majority weighted vote

Author Categorization sub-task

LIA participation in Author Categorization sub-task

Main idea

- Consider Author Categorization as tweet categorization
- A profile is equivalent to a bag of tweets where each one is tagged
- Majority vote over the profile
- Monolingual and domain specific model VS global model

Features

- Bag of words representation (TF, IDF, purity index)
- Unigrams, bi-grams, skip-grams (distance = 1) , tri-grams
- Assumption English words in Spanish tweets keep the English meaning
- Metadata: Language, Domain, Date, Author, ...
- Long lower-cased words (length>3), no punctuation
- No PoS or extra NLP resources

Systems description

- LIA_AC_1 : HMM and Poisson combination for English and Spanish
  - Poisson is used for fast match component in speech recognition
  - Fits the sparse distribution of relevant features for small classes

  Cosine was added for Spanish since there are less Spanish tweets
  Each domain has been processed separately
- LIA_AC_2 : HMM and Cosine with global models
  - 2 pass classification
  - "Non influencer" + (undecidable and professional) VS the rest
  - Then 2 classifiers (undecidable VS professional) and standard categorization
- LIA_AC_3 : Majority vote, giving more importance to small classes

Label distribution

Label	Train	Test
Public Institution	40	90
NGO	102	233
Stockholder	0	7
Investor	3	0
Sportsmen	57	208
Journalist	466	991
Employee	4	14
Undecidable	1028	1412
Celebrity	61	208
Professional	594	1546
Company	145	222

Local optimization for each run

- Wrt to the distribution if the best hypothesis is an over populated class ...
- The second hypothesis is an under populated class
- We permute

Imply sacrifice : losses in terms of accuracy but improvements with small classes
Similar to Gambit on chess

Official results (Average Accuracy)

#Run-ID	Automotive	Banking	Misc	Average
LIA_AC_1	0,445	**0,502**	0,461	**0,473**
Baseline-SVM	0,426	0,494	-	0,460
MF-Basline	**0,450**	0,420	**0,51**	0,435
LIA_AC_2	0,356	0,397	0,376	0,377
LIA_AC_3	0,292	0,308	0,369	0,300

Author Ranking sub-task

LIA participation in Author Ranking sub-task

System description

- LIA_AR_1 : HMM and Poisson combination for English and Spanish
  Each domain has been processed separately
  HMM with global model added in parity cases

Same process but ...

- Binary classification problem for each author
- Ranking according to the "influencer" probability over the bag of tweets
- Offset and threshold for the permutation
- Global model system showed no improvements

Official results (Average MAP)

#Run-ID	Automotive	Banking	Average MAP
Best	**0,721**	0,410	**0,565**
LIA_AR_1	0,502	**0,450**	0,476
Baseline	0,370	0,385	0,378

Reputation Dimension task

LIA participation in Reputation dimension task

Main idea

- Try different approaches from speech recognition

Systems description

- LIA_DIM_1 : Conditional random field
  Tagging unigram (with 5 neighbors context) and bigram then voting
- LIA_DIM_2 : Monolingual Multilayer Perceptron
  3 layers : 1 input, 1 hidden, 1 output
  Using unigram (with 5 neighbors context) and bigram
- LIA_DIM_3 : Naive use of continuous Word2Vec
  Using Brown corpus and background tweets
  We build a vector representing each class label
  Distance from each word to each label and majority vote
- LIA_DIM_4 : HMM and Cosine with global models
  HMM 80 % and Cosine 20 %
- LIA_DIM_5 : Majority vote, giving more importance to small classes

LIA participation in Reputation dimension task

Official results (F-Score and Accuracy)

#Run-ID	F-Score	Accuracy
Best	**0,473**	**0,732**
SVM Baseline	0,380	0,622
LIA_DIM_2	0,258	0,612
LIA_DIM_1	0,258	0,607
LIA_DIM_5	0,238	0,595
LIA_DIM_3	0,160	0,549
Naive Baseline	0,152	-
LIA_DIM_3	0,121	0,356

Bad results ...

According to classes distribution we predicted too much P&S

LIA participation in Reputation dimension task

New proposal

- Global cosine classifier (2 versions)
  No homogeneity rule
  With(out) features selection
  - Selecting the best weighted values of TF, IDF and Purity index

New results

#Run-ID	F-Score	Accuracy
Best	0,473	**0,732**
Cosine-FS	**0,521**	0,716
Cosine-Basic	**0,482**	0,661
SVM Baseline	0,380	0,622
LIA_DIM_2	0,258	0,612

Results near the best run without contextualization
Improvements with tweet expansion or active learning ?

Conclusion

Analyses

- Lack of time for the Profiling task due to the amount of data
- Merging strategy failed
- Monolingual combined performed better
- Homogeneity rule still need to be improved
- Better features selection increased the results

Perspectives

- Try profile summarization
- Use tweet expansion
- Exploit the unlabeled tweets

Any questions or suggestions ?