Learning Similarity Functions for Topic Detection in Online Reputation Monitoring

Damiano Spina, Julio Gonzalo, Enrique Amigó
{damiano,julio,enrique}@lsi.uned.es
http://nlp.uned.es

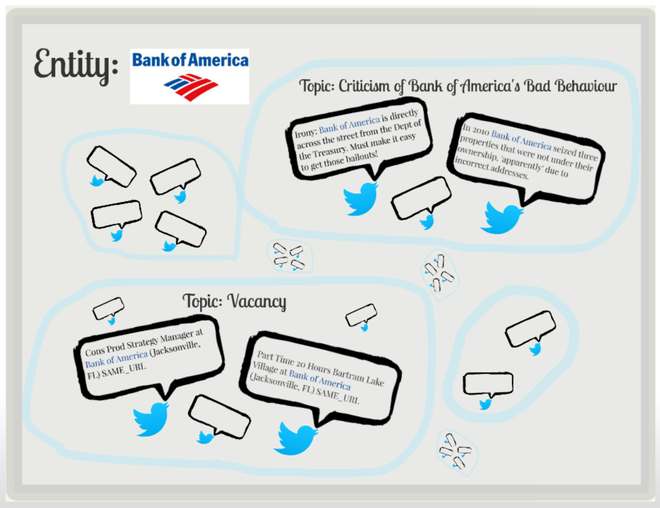
http://bit.ly/simFunctionsORM

SIGIR 2014 Workshop on Semantic Matching in Information Retrieval (SMIR)
Gold Coast, Australia. July 11th, 2014.

Topic Detection as a Clustering Task

What Are People Saying About a Given Entity Right Now?

- *Entity* = Brand, Company, Organization, Public Figure
- Early Detection of topics that may damage the reputation of the entity
- Twitter's Long Tail: Sparsity



RepLab@CLEF 2013

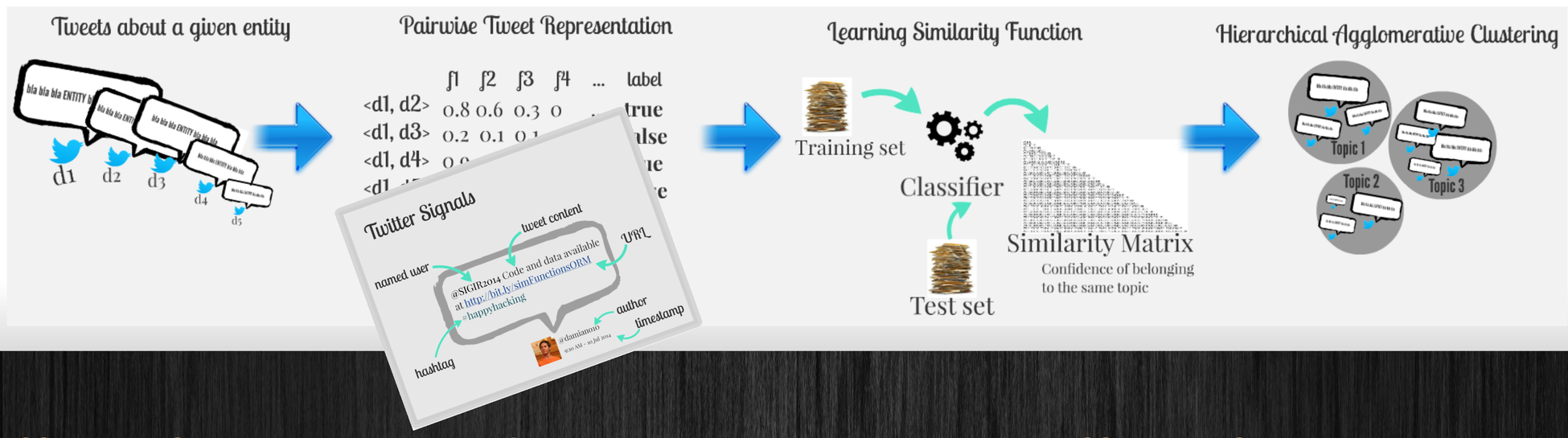
Evaluation of Online Reputation Monitoring Systems
http://nlp.uned.es/replab2013

Dataset for Topic Detection Task

61 entities
100,869 tweets annotated with 8,765 different topics

Training ≠ Test Topics, no overlap in time
544 (1,109) tweets and
57 (87) topics for training (testing) per entity

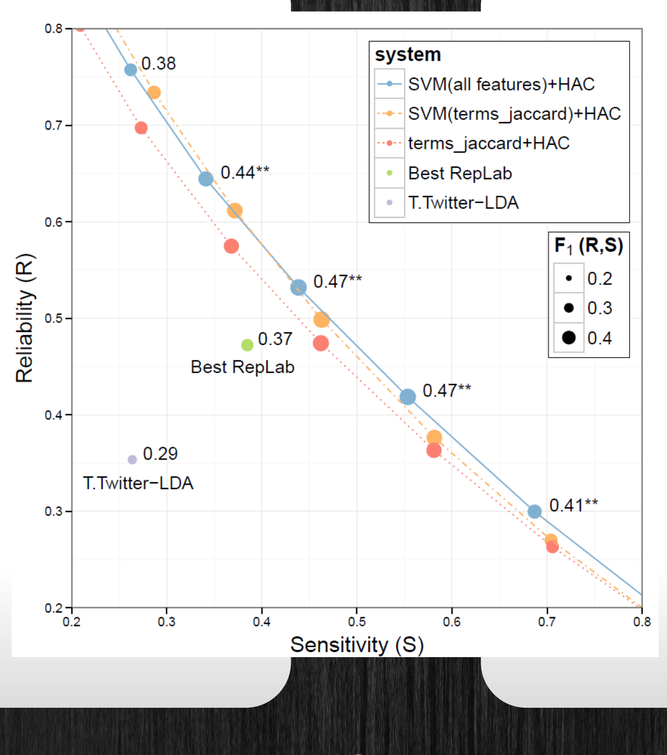
Learning Similarity Functions



Effect of Twitter Signals

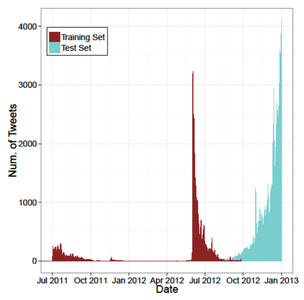
- Linear Support Vector Machine (SVM)
- High correlation with Maximal Pairwise Accuracy (theoretical upper bound)
- Results can be extrapolated to other Machine Learning algorithms
- Most signals are able to improve term-based features

System	AUC	MAR
Text only SVM(terms_jaccard)+HAC	0.40	0.59
All Twitter Signals SVM(all features)+HAC	0.41	0.61*



Effect of Training Process

- Supervised system consistently improves the performance of its unsupervised counterpart
- Regardless of how the similarity threshold is set
- Previous annotations can be used to learn better topic models, although differences are not large



System	AUC	MAR
Unsupervised terms_jaccard+HAC	0.38	0.57
Supervised SVM(terms_jaccard)+HAC	0.40	0.59*

Conclusions

- Twitter signals (authors, timestamps, hashtags, etc.) can indeed be used to improve topic detection w.r.t. textual content only
- It is possible to learn effectively from manually annotated topics (pairwise tweet similarity)
- Organizational ("Hate Opinions") vs. Event-like Topics ("Director of the Bank Accused of Evading Taxes")
- Learning Similarity Functions + Hierarchical Agglomerative Clustering (HAC) significantly outperforms the state-of-the-art (RepLab Topic Detection systems)
  - It gets close to the inter-annotator agreement rate: 0.48 F_1(R,S)
  - When data is sparse (as in the ORM scenario) it can be more effective than using probabilistic generative models such as Temporal Twitter-LDA

Acknowledgments. This research was partially supported by the European Community's FP7 Programme under grant agreement nr. 288024 (LiMoSINe), the Spanish Ministry of Education (FPU grant nr. AP2009-0507), the Spanish Ministry of Science and Innovation (Holopedia Project, TIN2010-21128-C02), the UNED (project nr. 2012V/PUNED/0004), the Regional Government of Madrid and the ESF under MA2VICMR (S2009/TIC-1542), a Google Faculty Research Award (Axiometrics Project) and a SIGIR Student Travel Grant.