Towards the improvement of topic priority assignment using various topic detection methods for e-reputation monitoring on Twitter (short)

J.V. Cossu
University of Avignon, France

NLDB'2014

June 19 2014

Participants

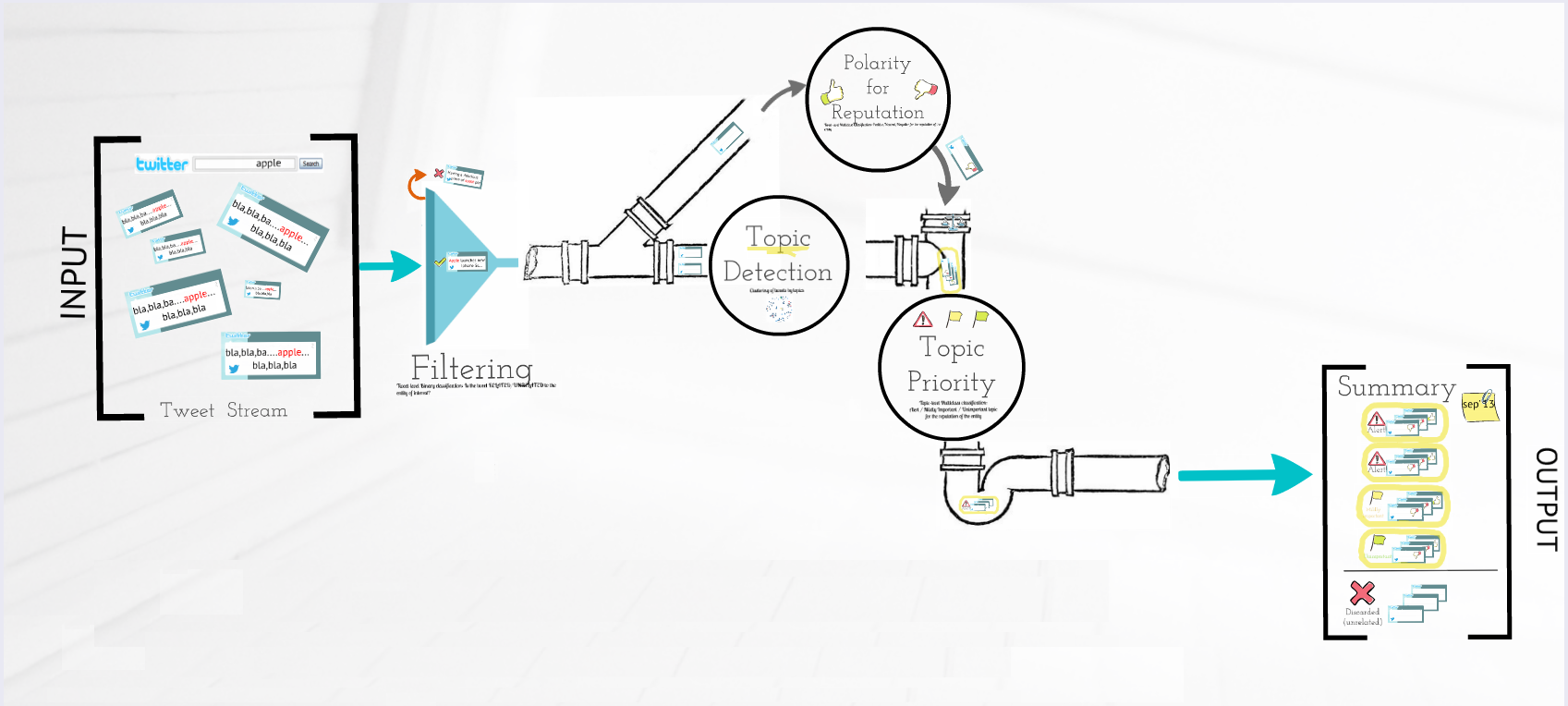
J.-V. Cossu*, B. Bigot, L. Bonnefoy, G. Senay

Opinion database

E-reputation monitoring issue

- Group tweets depending on trends
- Rank them to find those needing a reaction

Replab 2013 challenge



- RepLab 2013 Overview from Julio Gonzalo (UNED) (September 2013)

Opinion database

Data

- Over 142,000 tweets (80% English, 20% Spanish)
- 61 entities spread in 4 domains

Filtering

- Distinguishing whether "Stanford" refers to the University or the place
- Binary classification

"Barclays" plans additional job cuts in the next two years

Polarity for Reputation

- Does this information have implication on "Barclays" reputation ?
- 3 reputation classes : POSITIVE, NEUTRAL and NEGATIVE

Opinion database

Topic Detection (clustering)

- Grouping tweets referring to the same subject/event/conversation
- Around 8,300 topics (3,400 in train and 5,300 for the test set)
- Hard classification issue with 500 common topics between train and test
- Hard clustering problem due to the huge graphs 2,200 nodes per entity

Topic Priority (ranking)

- Predict the priority level of above clusters
- Negative tweets may impact the priority rank
- User's influence may receive an higher priority
- Relationship : IMPORTANT > MILDLY > NOT IMPORTANT

Our work

Approach

- Focus on Topic Detection and Priority
- Multilingual approach with global model
- Considering both clustering and ranking problems as a classification issue
- Relationships : Topic -> Priority

Metrics

- Bag of relationship ($<$, $>$, $=$) between documents
- R&S can be seen as precision and recall of relationships

Our work

Tweets Clustering

- Identification of headwords for each topic in the training-set

Tweets Ranking

- Work on tweets content
- Tokens representing user identity and entity background

Combination

- State of the art merging strategies failed
- Topic Detection > Topic Ranking
- Work on the topic "bag of tweets"
- Majority vote strategy

Our work

Topic detection Systems

- Hierarchical clustering using Jaccard similarity
- K-Means clustering using Jaccard similarity
- MAP classifier

Topic priority Systems

- KNN Jaccard distance over discriminant bow representation
- KBA'2012 system

Baseline

- Memory test, match the closest tweet (Jaccard word similarity) in the training set

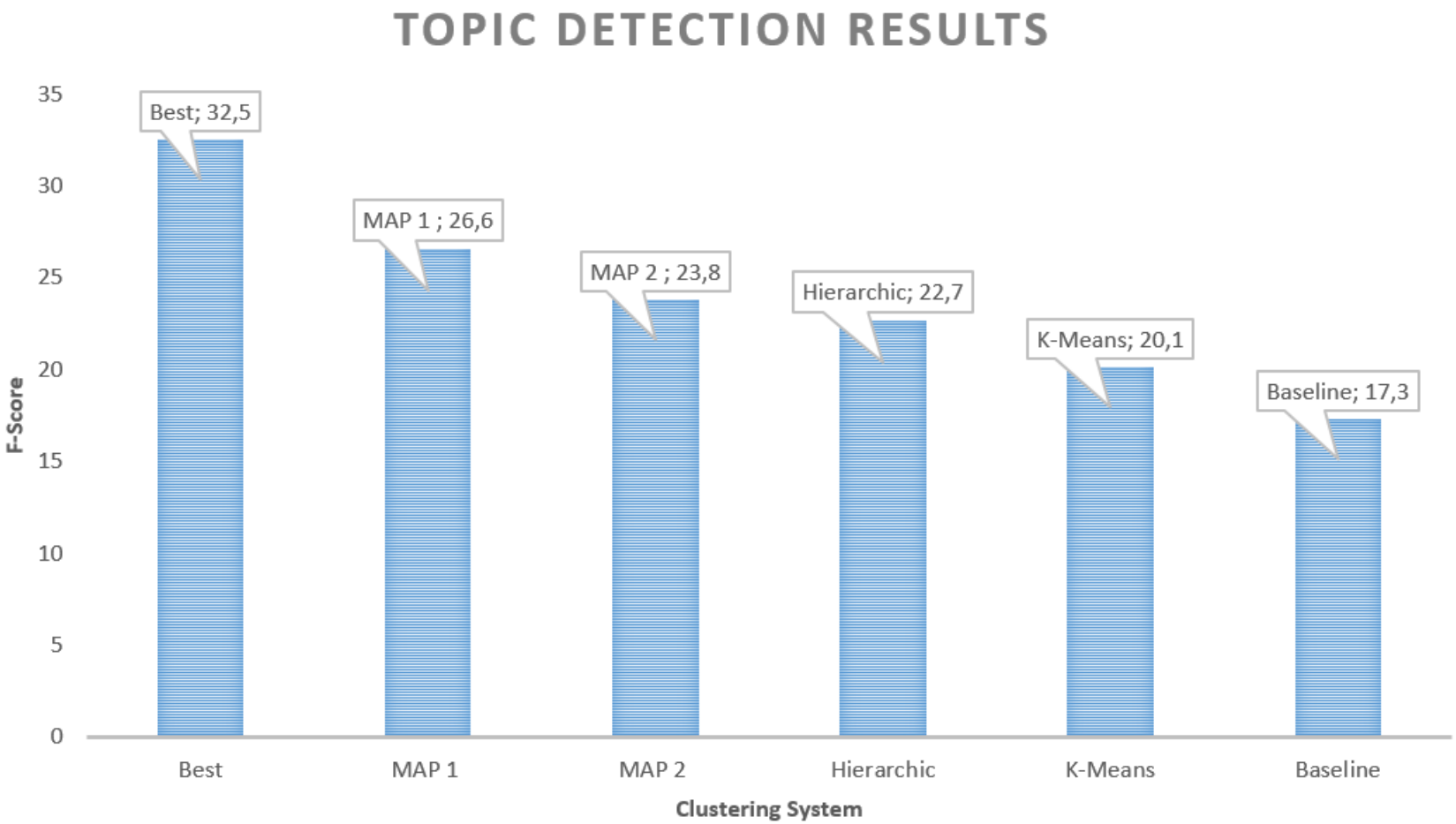
Maximum a Posteriori [Hazen, 2011] Uses features purity

Tweet 281764404670889984 AlvinWongCH RL2013D04E145

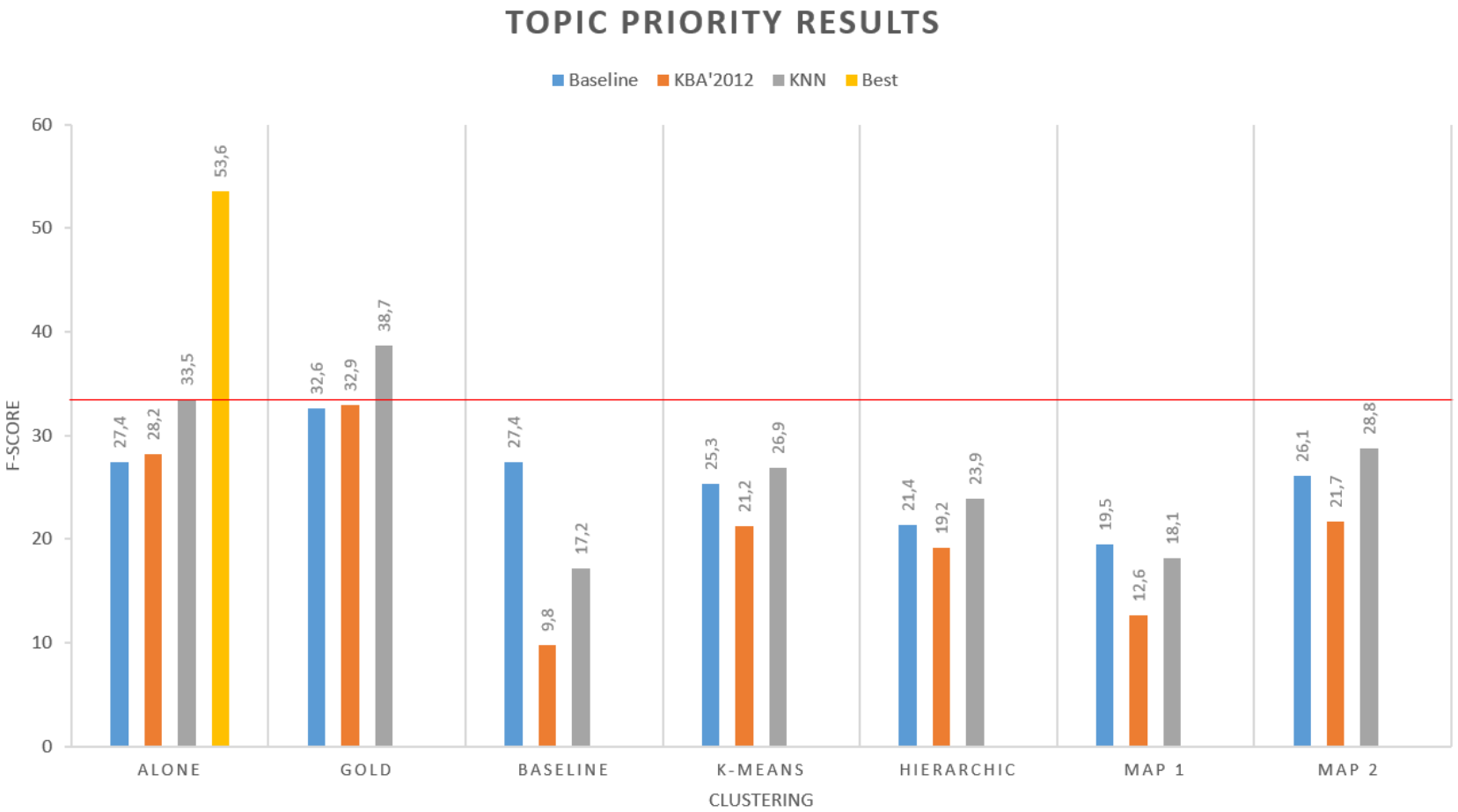
#nowplaying *adele- skyfall*
TOPIC -> Now playing/listening (Ranked MILDLY*IMPORTANT*)

now playing		listening, playing alicia's songs		now playing/listening	
jovi	0.022	alicia	0.027	whitney	0.021
bon	0.022	keys	0.026	houston	0.020
np	0.009	np	0.013	np	0.013
nowplaying	0.007	**nowplaying**	0.008	adele	0.012
listening	0.006	ft	0.007	**nowplaying**	0.009
bad	0.004	usher	0.007	you	0.006
always	0.004	by	0.006	love	0.006
bed	0.004	listening	0.005	by	0.007
by	0.004	song	0.005	listening	0.005

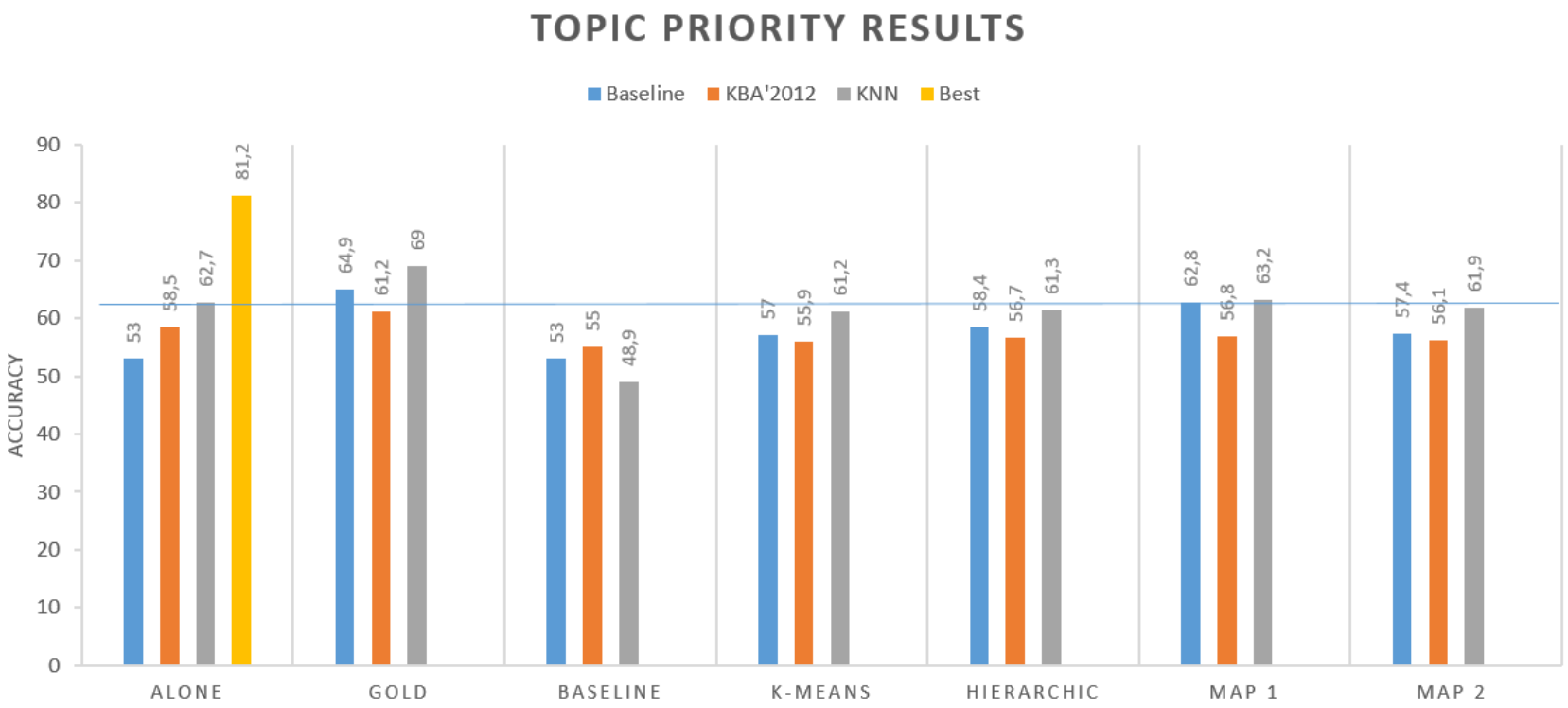
Results - Topic Detection



Results - Topic Priority



Results - Topic Priority



Results

Topic Detection

System	RELIABILITY	SENSITIVITY	F-MEASURE
Best	**.462**	.324	**.325**
MAP 1	.193	**.497**	.266
MAP 2	.381	.172	.238
Hierarchic	.261	.220	.227
K-means	.308	.157	.201
Baseline	.152	.217	.173

Topic Priority as Tweet Priority

System	RELIABILITY	SENSITIVITY	F-MEASURE
KNN	.387	**.315**	**.335**
KBA	.315	.276	.282
Baseline	**.403**	.248	.274
Gold+MAP2	.756	.518	.602
Best answers	.789	.481	.536
KNN+Gold	.549	.345	.387

Results

Topic Priority with MAP 1

System	RELIABILITY	SENSITIVITY	F-MEASURE
Baseline	.383	**.151**	**.195**
KBA	**.551**	.098	.126
KNN	.413	.136	.181

Topic Priority with MAP 2

System	RELIABILITY	SENSITIVITY	F-MEASURE
Baseline	**.406**	.214	.261
KBA	.361	.171	.217
KNN	.405	**.249**	**.288**

Conclusion

Classification issues

- Classification over the tweet or cluster bag of words
- Multilingual classification
- Similar topics : Pictures on social networks, Pictures posted, Pictures posted in social networks, Pictures posted on social networks

Combination issues

- Selection strategies are not adapted to the ranking issue
- Additional information provided by our Topic Detection is not efficient

End

Any questions ?