FIRE

Nicholas Mamo

nicholas.mamo.14@um.edu.mt

Joel Azzopardi

joel.azzopardi@um.edu.mt

FIRE: Finding Important News REports

- Too much information, too little time
- Solution to find emerging news and related reports
- Crowd-sourced using Twitter

Twitter

- Popular social network
- Open
- Rich in information
- Developer tools

Objectives

- Examine tweets' effect on clustering
- Establish clusters' potential to represent topics
- Identify the news categories that can be detected
- Ascertain the relationship between spam and tweet features

Workflow


t-1
t
t+1
Sampling
Processing
Sampling
Processing
Sampling
Processing

Sampling

- English tweets are collected from the public stream
- Content and author features are stored alongside the messages
- Around 0.2% of all tweets are collected in any given time-window

Sampling

Preprocessing

Nutrition

Clustering

Postprocessing

Emerging Topic Detection

Preprocessing

- A conservative approach is adopted to remove spam and noise
- Tweets are converted into documents in the vector space using the bag-of-words model
- Content and author features are saved as document attributes

Sampling

Preprocessing

Nutrition

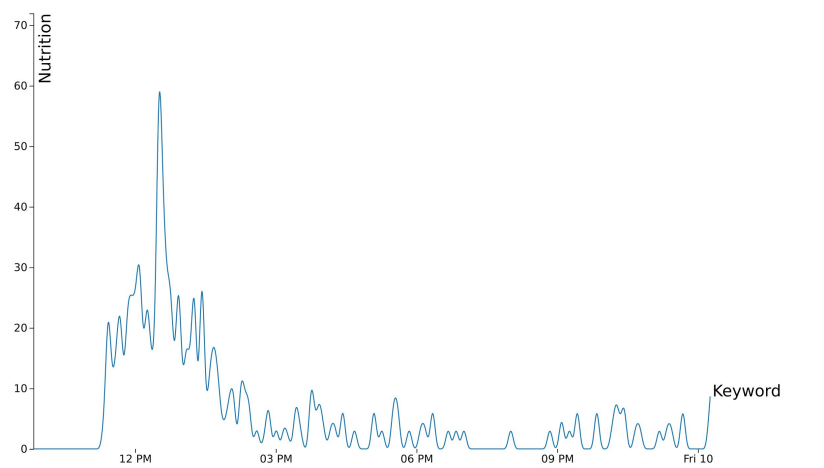
Clustering

Postprocessing

Emerging Topic Detection

Nutrition

- A measure of a term's popularity, based only on keyword usage



Sampling

Preprocessing

Nutrition

Clustering

Postprocessing

Emerging Topic Detection

Clustering

- Alike documents are grouped together to form clusters
- An adaptation of the No-K-Means algorithm is used, allowing big clusters to grow bigger
- Clusters are analogous to stories

Sampling

Preprocessing

Nutrition

Clustering

Postprocessing

Emerging Topic Detection

Postprocessing

- Noisy and spam clusters are removed
- Named entities are extracted from cluster labels
- Nutrition is used to calculate burstiness, representing the increase in interest in keywords

Sampling

Preprocessing

Nutrition

Clustering

Postprocessing

Emerging Topic Detection

Emerging Topic Detection

- A neural network delivers a judgement on the newsworthiness of stories
- Topic tracking is performed with approved topics, and relevant news reports are fetched

Evaluation

- Corpus from McMinn et al. (2013) was used to evaluate the clustering component
- The topic detection algorithm’s results were compared with Twitter’s trends and news portals
- A neural network was trained and evaluated on real events across a week

Results

- Brevity does not affect clustering
- Content and author features useful to detect and remove spam
- News stories can be detected much earlier than Twitter's trends mechanism

		US	UK
Time (m)	FIRE	-25.1	**-38.7**
	Cataldi et al. (2014)	25.6	17.4
F1	FIRE	**0.35**	0.27
	Cataldi et al. (2014)	0.18	0.16

Conclusion

- Twitter necessitates efficient noise removal
- However, it is a promising tool for emerging topic detection

FIRE

Nicholas Mamo

nicholas.mamo.14@um.edu.mt

Joel Azzopardi

joel.azzopardi@um.edu.mt