Big Data Research as Interaction between Topic Models and Expert Data: A New Approach to Capturing National Online Debates

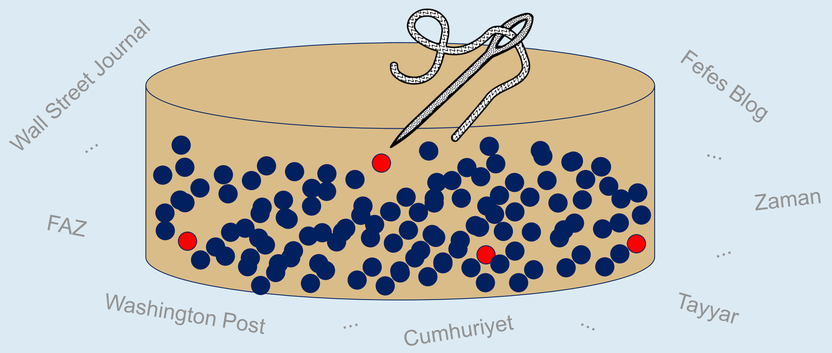
Cäcilia Zirn, Eike Mark Rinke, Charlotte Löb, & Hartmut Wessler

TASK

Dataset

- **1** year of scraping
- **101** sources
- Leading national blogs & news websites
- **6** countries
  - Australia, Germany, Lebanon, Turkey, Switzerland, USA
- **4** languages
  - English, German, Arabic, Turkish
- **1.6 million articles**

Find the needle in the haystack!



General goal: More insightful topic identification

- Using expert domain knowledge for better automated identification of documents from prespecified broad topic domain ("meta-topic")
  - For classification of documents by meta-topic
  - For cross-media and cross-national comparison of topic structures within a specific meta-topic

Specific application: Expert knowledge for improved classification of documents by topic

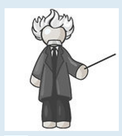
- Finding articles that belong to at least one topic that are part of meta-topic "role of religion in public life"

TOPIC MODELING APPROACH

TOPIC MODEL

Training	Inference
• input: document collection, # of expected topics (≈ 500) • output: topics defined by weighted words • algorithm: Latent Dirichlet Allocation	• input: documents • output: topic proportions of each document

Expert



Seed Terms

crucifix
class room
bavaria

burqa
teacher
…

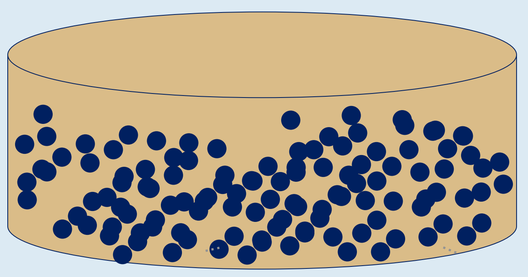
T1 T2 T3 T4 T5 T6 T7
T8 T9 T10 T11 T12 T13 …

T5	T134	T205
paris	...	...
authorities	...	...
terrorists	...	...
effort	...	...
senate	...	...

Identification of relevant topics

60% T1
20% T2
20% T3-T500

Select documents that contain relevant topics
(at least to user-specified minimum)



WHY DID YOU NOT USE ...

Machine Learning	Information Retrieval	Clustering	Seed-based LDA	Crowd Sourcing
• **no training data** available • irrelevant classes unknown • classes (relevant vs. irrelevant) **unbalanced**	• expert-generated list of relevant items probably **not complete** • collecting all possible search terms not feasible (**synonyms, paraphrases**)	• high **cognitive effort** required to comprehend clusters • **words** can be assigned to **a single cluster** only (though they might be relevant for several topics) • only **single meaning** per word can be used	• **presumptions** about expected issues **bias topic creation** and might **limitate results** • requires **comprehensive** issue (seed) list	• coders must **follow complex instructions** and decide carefully • **difficult to scale** to millions of documents • **requires preexisting domain knowledge** • **difficult to verify** coding quality with extremely few positive codes

mzes mannheimer zentrum für europäische sozialforschung

UNIVERSITÄT MANNHEIM