Universität Bremen

Hochschule Karlsruhe Technik und Wirtschaft
UNIVERSITY OF APPLIED SCIENCES

Automated Intrinsic Text Classification for Component Content Management Applications in Technical Communication

Jan Oevermann
jan.oevermann@hs-karlsruhe.de

Wolfgang Ziegler
wolfgang.ziegler@hs-karlsruhe.de

DocEng 2016, Vienna, 15.09.2016

Motivation

- Semantic access to information through classification
- Demand for automation in industry use cases

- Adapt existing ML methods for
  Component Content Management
- Little research on Technical Communication topics

Technical Communication

- Writing documentation (and more)
- Complex information management
- Legal obligations and international standards

- Component Content Management
  - Modularized content for reuse and translation
  - XML-based information models
  - Metadata and classification models
  - Single Source Publishing

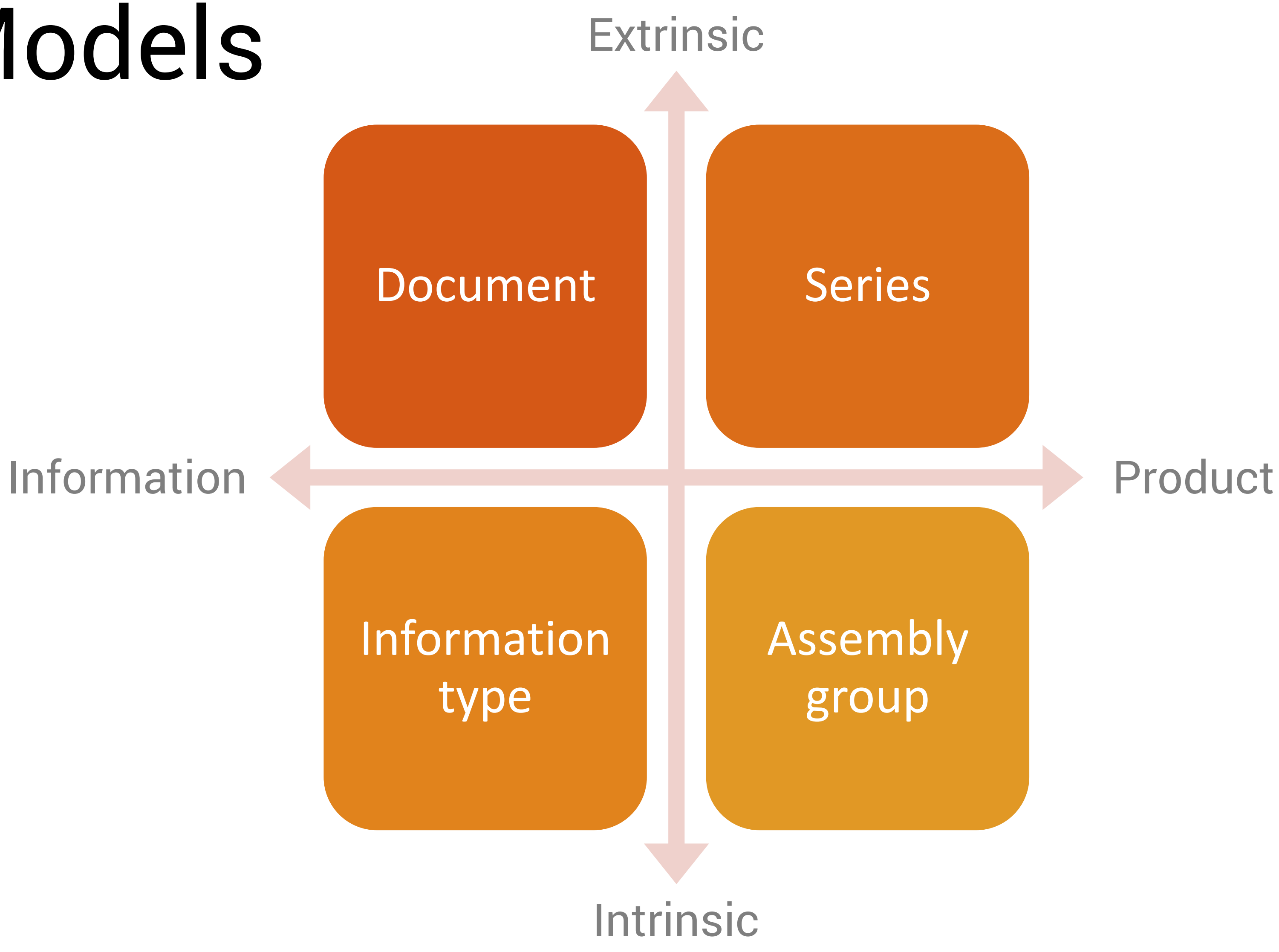
Methodology

1. Characterize relevant properties of CCM
2. Derive implications for classification
3. Verify with real-world data sets
   (Vector space classification)

Classification Models

- PI classification model (Ziegler 2011)
- Organized in taxonomies

- Focus on intrinsic information classification



Use Cases

- Content delivery portals
- Automated publishing
- Dynamic linking

	Series	Model	Project
Safety advice	C-123		
Product description		C-321	
Operation Main Engine	C-159		C-158
Maintenance			C-123

Characteristics

- Standardized patterns
- Specific terminology
- Size of content
- Training and validation data
- Quality requirements

3.3.2 Starting the engine

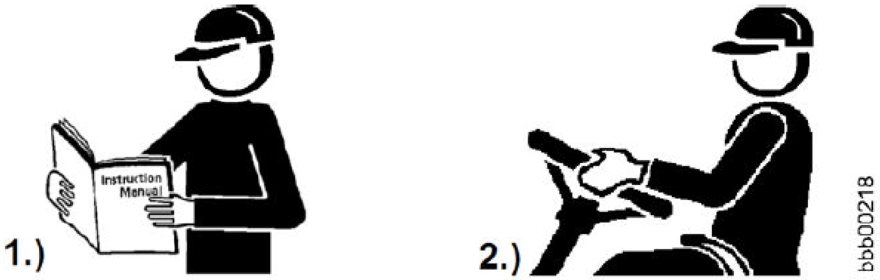


Fig. 288: Operating manual

1.) Make sure you have read and understood the operator's manual

2.) Then you are rea
machine

Only operate the machine after you have read and understood th
manual!

Note
The machine is equipped with a hydrostatic travel drive.

▶ You cannot start the engine by bump-starting it or towing it.

Data sets

Set	Sector	Units	Words/Unit	Classes
A	Construction equipment	570	173	11
B	Medical lab equipment	278	41	10
C	Security printing presses	3947	97	22

- XML-based content components
- Manually classified
- German language

Implications

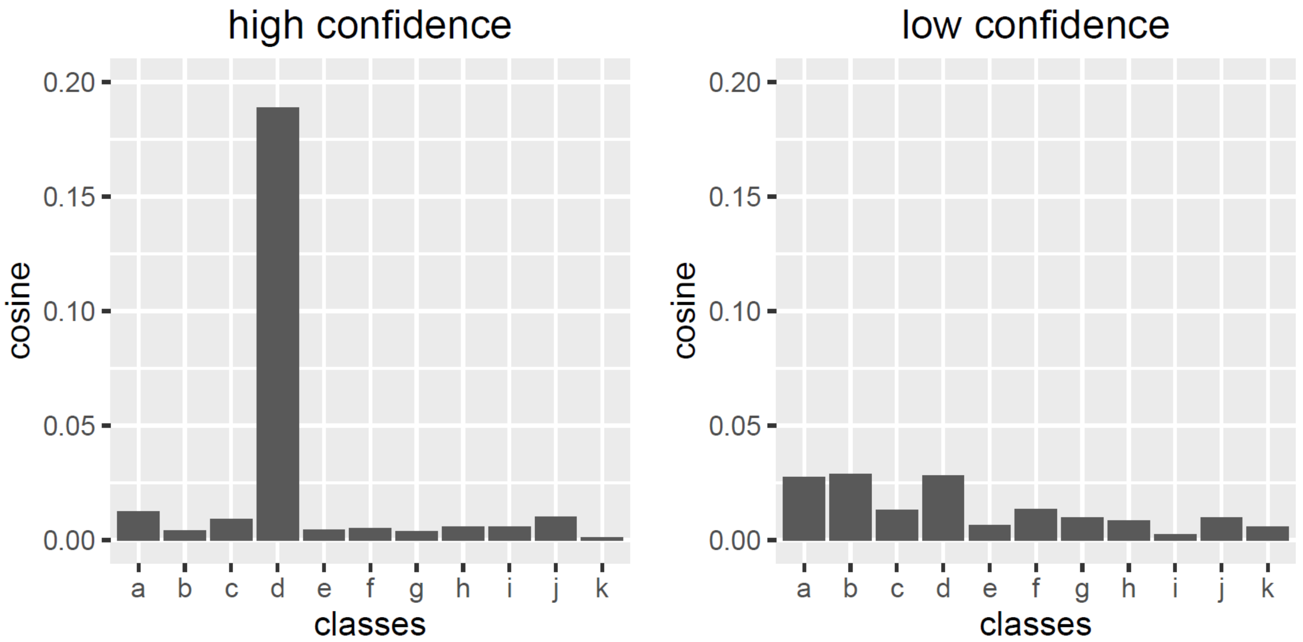
- Semantic quantifiers

$$tf_{iq} = tf_i * q \quad \text{for } q > 0$$

- Confidence scoring

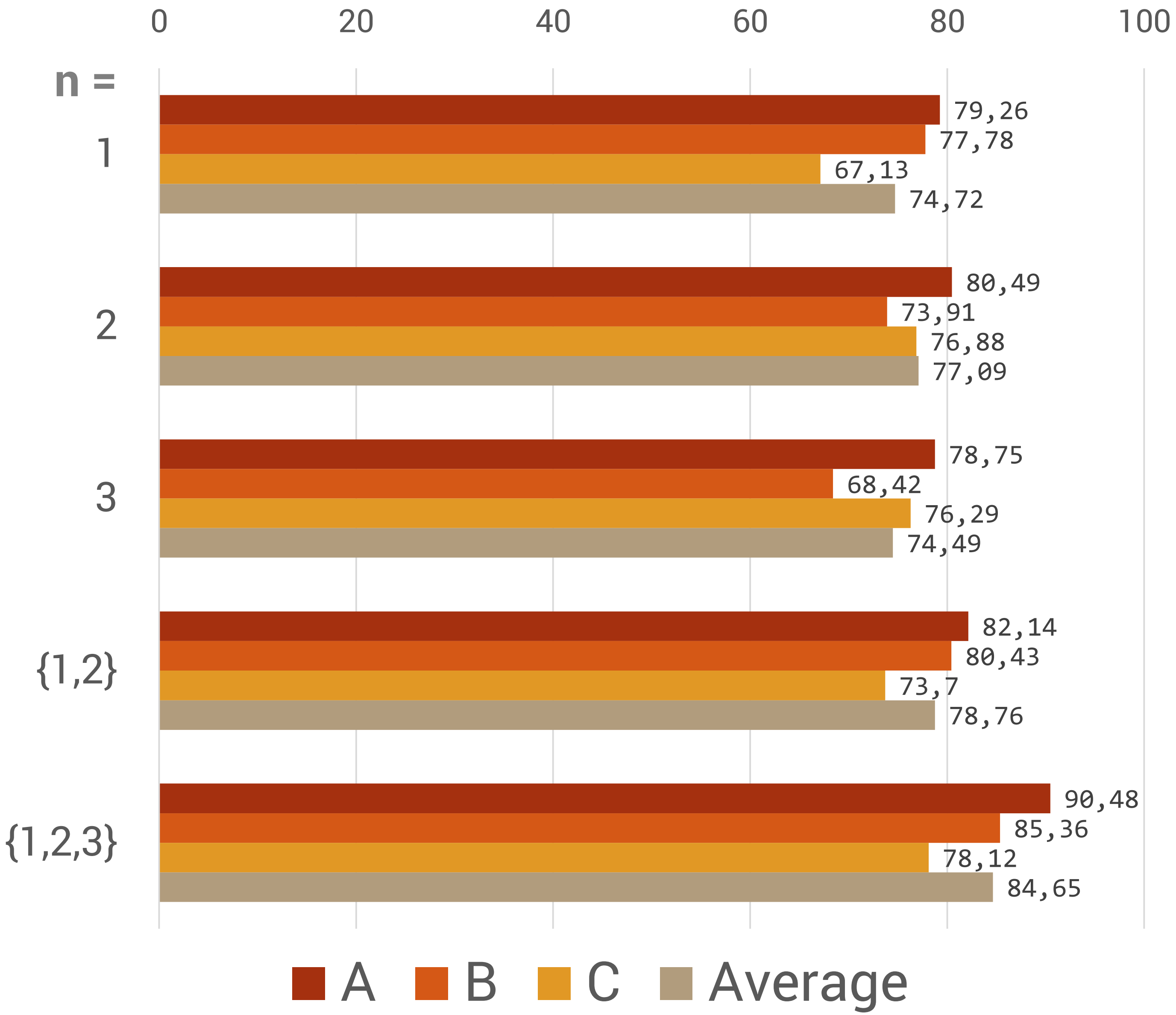
$$p = \frac{s_1 - s_2}{s_1 - s_n}$$

Instead of softmax or standard deviation



Feature selection

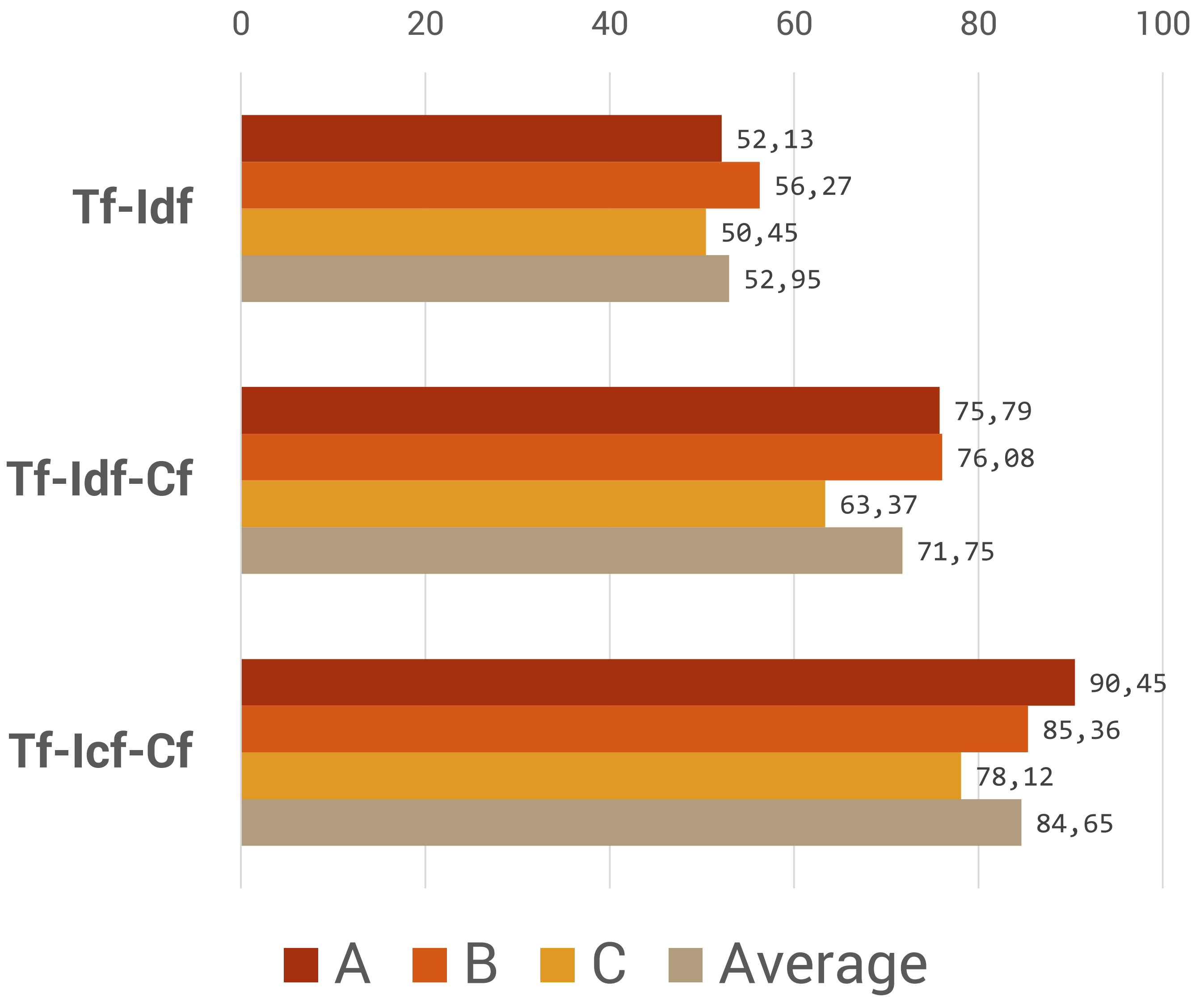
- Smaller total number of features
  - Standardization of wording and patterns
  - Size of content components
- Single words and patterns important
  - Combination of words and patterns



Token weighting

- Tf-Idf
  - Good for documents
- Tf-Idf-Cf (Liu/Yang 2012)
  - In-class characteristics
- Tf-Icf-Cf:

$$w_{ij} = \log(1 + tf_i) * \log(1 + \frac{|C|}{tf_i}) * \frac{tf_{ij}}{C_j}$$



Applications

- Data migration
- Key figures (QA)

- Authoring assistance
- Content delivery portals (API, Import hook)

Results & Observations

- CCM has different requirements than document classification
- Technical content is well suited for automated classification

- Set of adjustments for content components to improve results
- Working prototype:
  REST API for classification of content components

Related work & Outlook

- Soto et al. (2015): Similarity-Based Support for Text Reuse in Technical Writing
- Oevermann (2016): Reconstructing Semantic Structures in Technical Documentation with Vector Space Classification

- Apply results to unstructured technical content
- Use more advanced machine learning or deep learning technologies

Contact

Jan Oevermann

jan.oevermann@hs-karlsruhe.de

www.janoevermann.de