universität innsbruck

DBIS DATABASES AND INFORMATION SYSTEMS

Overview of the Author Identification Task at PAN-2018:

Style Change Detection

Michael Tschuggnall, Günther Specht, Benno Stein, Martin Potthast

11.9.2018

PAN@CLEF 2018, Avignon, France

Content

1. Task description
2. Dataset
3. Submitted approaches
4. Results

Content

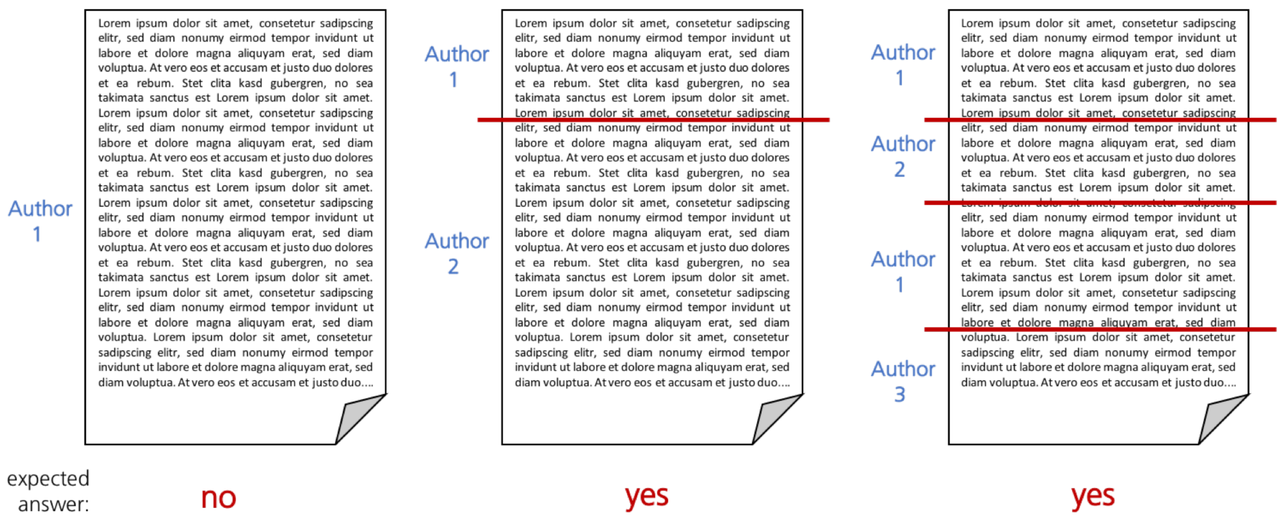
1. **Task description**
2. Dataset
3. Submitted approaches
4. Results

Task description

- Author Identification
  - Authorship Attribution (cross-domain)
  - **Style Change Detection**

- Substantially relaxed the previous **style *breach* detection** task:

  Given a previously unseen document, decide whether it was written by one or multiple authors, i.e., if it contains style changes or not.



Content

1. Task description
2. **Dataset**
3. Submitted approaches
4. Results

Dataset

- Goal: create a realistic, non-artificial and comprehensive dataset
- Optimal case
  - Find documents which are single-authored and ~~collaboratively written, indicated as such and freely available~~
- Alternative
  - Compile own dataset from freely available sources
  - Requirements
    - Find multiple authors that write about the same topic
    - Find texts that are freely available and of sufficient length
    - Multi-authored texts need to contain the same topic / subtopic
  - Q&A platform **StackExchange** fulfills these requirements

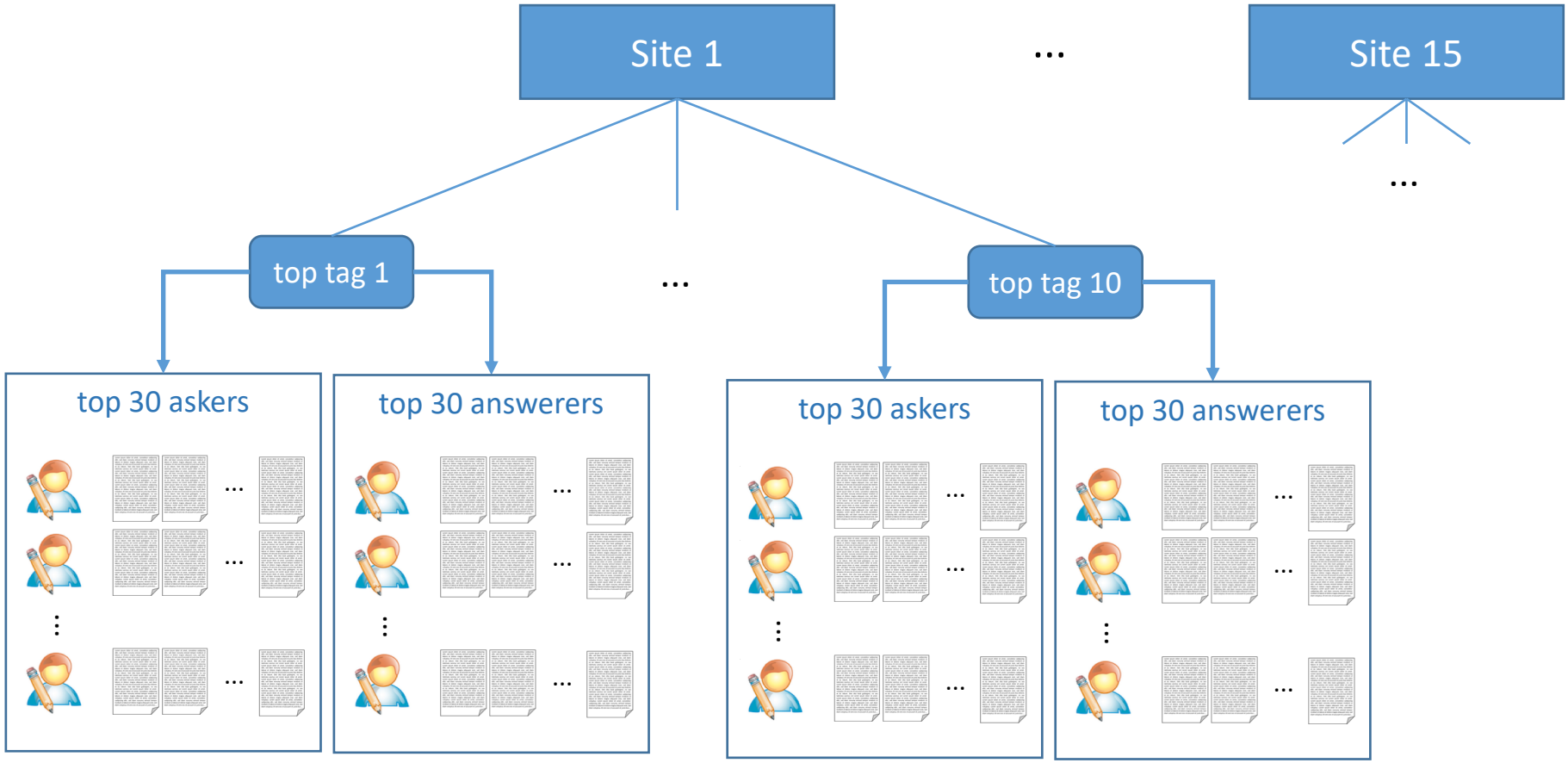
Dataset

- StackExchange consists of several **sites** (174 sites)

- Each question/answer of a site is associated with a set of **tags** (subtopics), example for the **Photography** site:
  - *lens, canon, nikon, lightroom, photoshop, ...*

Dataset

- Crawling process using the StackExchange API:



Dataset

- Using the raw texts, a **training** (50%), **validation** (25%) and **test** (25%) dataset has been created
  - Cleaning
    - Remove links
    - Remove images
    - Remove code snippets
    - Remove bullet lists
    - Remove block quotes
    - Remove very short questions/answers
    - Remove edited questions/answers
  - Each dataset contains 50% single-author documents and 50% multi-authored documents
  - Parameters

Parameter	Value/s
number of style changes	0–3
number of collaborating authors	1–3
document length	300–1000 tokens
change positions	at the end of / within paragraphs, mixed
segment length distribution	equalized / randomly
two-authors distributions	(A1-A2), (A1-A2-A1), (A1-A2-A1-A2)
three-authors distributions	(A1-A2-A3), (A1-A2-A1-A3), (A1-A2-A3-A1) (A2-A1-A3-A1)

Dataset

- Final datasets resulting from 15 selected sites

	Training				Validation				Test			
		Authors				Authors				Authors		
Site	**Problems**	**1**	**2**	**3**	**Problems**	**1**	**2**	**3**	**Problems**	**1**	**2**	**3**
bicycles	160	80	47	33	82	41	28	13	70	35	27	8
christianity	358	179	107	72	176	88	48	40	172	86	45	41
gaming	178	89	47	42	86	43	23	20	78	39	21	18
history	354	177	104	73	178	89	54	35	170	85	46	39
islam	166	83	49	34	86	43	31	12	72	36	20	16
linguistics	144	72	46	26	72	36	22	14	64	32	12	20
meta	196	98	56	42	94	47	30	17	90	45	30	15
parenting	178	89	54	35	92	46	32	14	78	39	27	12
philosophy	468	234	146	88	232	116	63	53	224	112	65	47
poker	100	50	35	15	48	24	14	10	42	21	13	8
politics	204	102	57	45	102	51	34	17	90	45	22	23
project man.	104	52	24	28	50	25	12	13	44	22	14	8
sports	102	51	34	17	54	27	20	7	40	20	12	8
stackoverflow	112	56	23	33	60	30	16	14	48	24	12	12
writers	156	78	43	35	80	40	25	15	70	35	18	17
	2980	**1490**	**872**	**618**	**1492**	**746**	**452**	**294**	**1352**	**676**	**384**	**292**

Content

1. Task description
2. Dataset
3. **Submitted approaches**
4. Results

Approaches

- 6 registrations, 5 submissions to TIRA

- **Hosseinia et al.**
  - two parallel recurrent neural networks
  - based on hierarchical structure of the language (parse tree) rather than n-grams
    - so called „Parse Tree Features“ traversing the parse tree
    - long running time due to parse tree computation
  - classification: reverse sentence order of the document, compute parse tree features and compare normal and reverse document using different (7) similarity functions
- **Khan**
  - Algorithmic approach
    - split documents by sentences
    - build groups of sentences, whereby sentences are shared within groups (at the segment position)
    - compute features for each group
      - most/least frequent stop words, most frequent words / word pairs, punctuation
  - style change calculation by computing a match score between two consecutive groups of sentences (sharing the middle one)

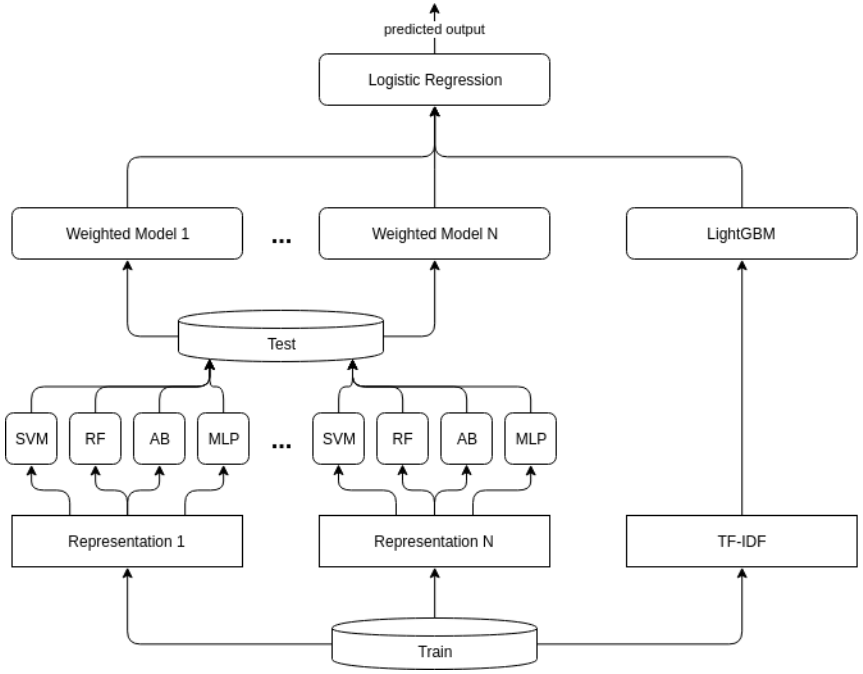
Approaches

- **Safin and Ogaltsov**
  - ensemble of 3 classifiers using 3 types of features:
    - text statistics (number of sentences, unique word frequencies, text length, punctuation, letter symbols frequencies) with Random Forest
    - 3000-dimensional vectors containing information about occurrences of char n-grams in the text
    - word n-grams for n=1-6 with logistic regression
  - final prediction: weighted sum of predictions of all 3 classifiers

- **Schaetti**
  - Character-based convolutional neural network (CNN)
    - Character-embedding-layer (dimension=50)
    - 3 different convolutional layers with 25 filters each: find effective 2-grams
    - 3 max-pooling layers
    - Final linear layer of size 2 for classification
  - Extended training corpus (25,000 documents)

Approaches

- **Zlatkova et al.**
  - preprosessing
    - replace URLS, long numbers, file paths, very long words with specific tokens
    - segment the document into 3 equal parts
  - various feature types
    - 1-5 word n-grams, grammar contractions, frequent words, lexical features, ...
  - classification using ensemble (stacking)
    - for each feature type: SVM, Random Forest, AdaBoost Trees, MLP
    - combiner: logistic regression using those models + tf-idf-based lightGBM



Content

1. Task description
2. Dataset
3. Submitted approaches
4. **Results**

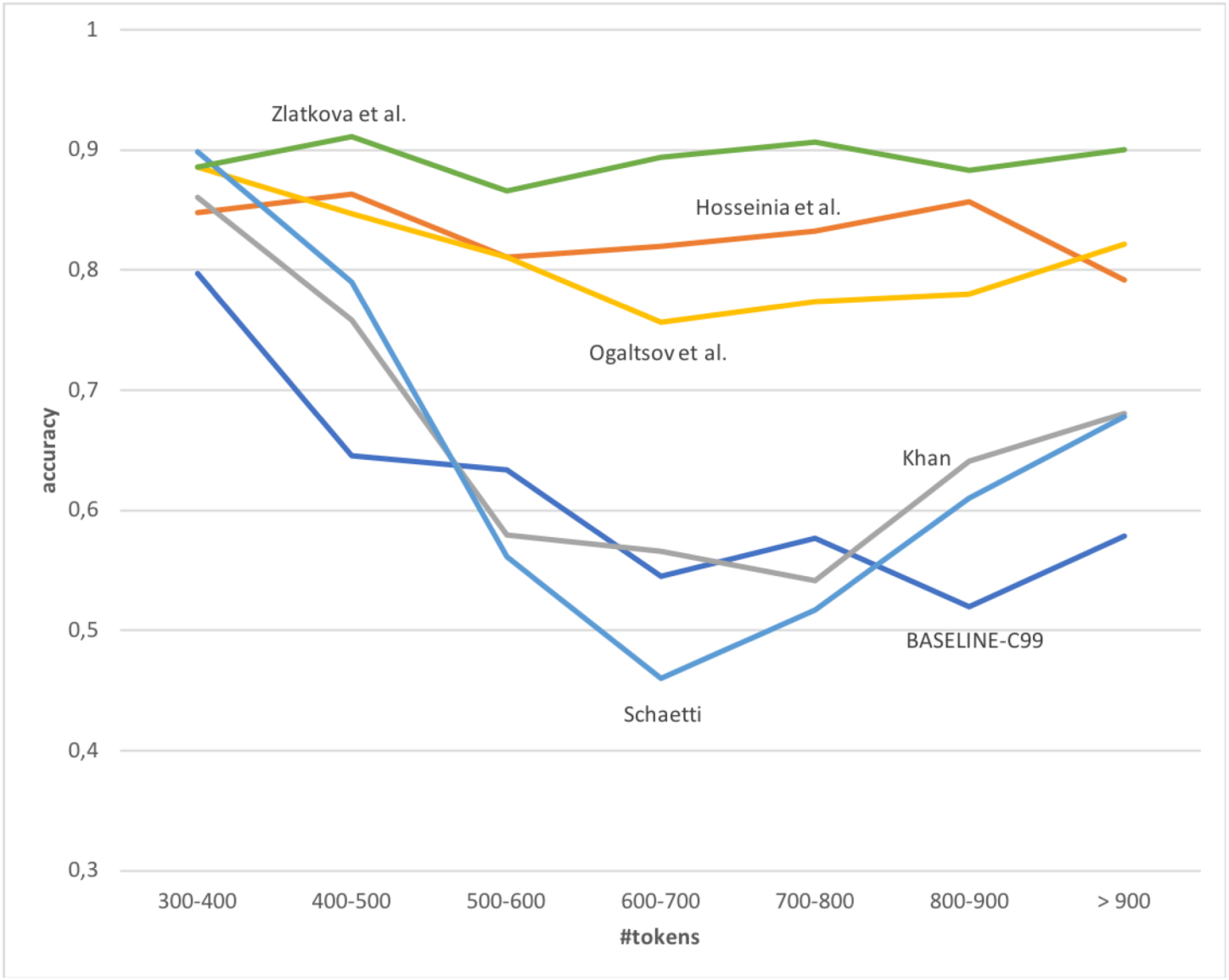
Baseline

- **BASELINE-rnd1**
  - Simple guessing (50%)
- **BASELINE-rnd2**
  - „advanced“ guessing using text length statistics
- **BASELINE-C99**
  - Utilize C99 text segmentation algorithm (Choi, 2000)
  - Let the algorithm determine the number of segments
  - If #segments = 1: predict no style changes, otherwise predict changes

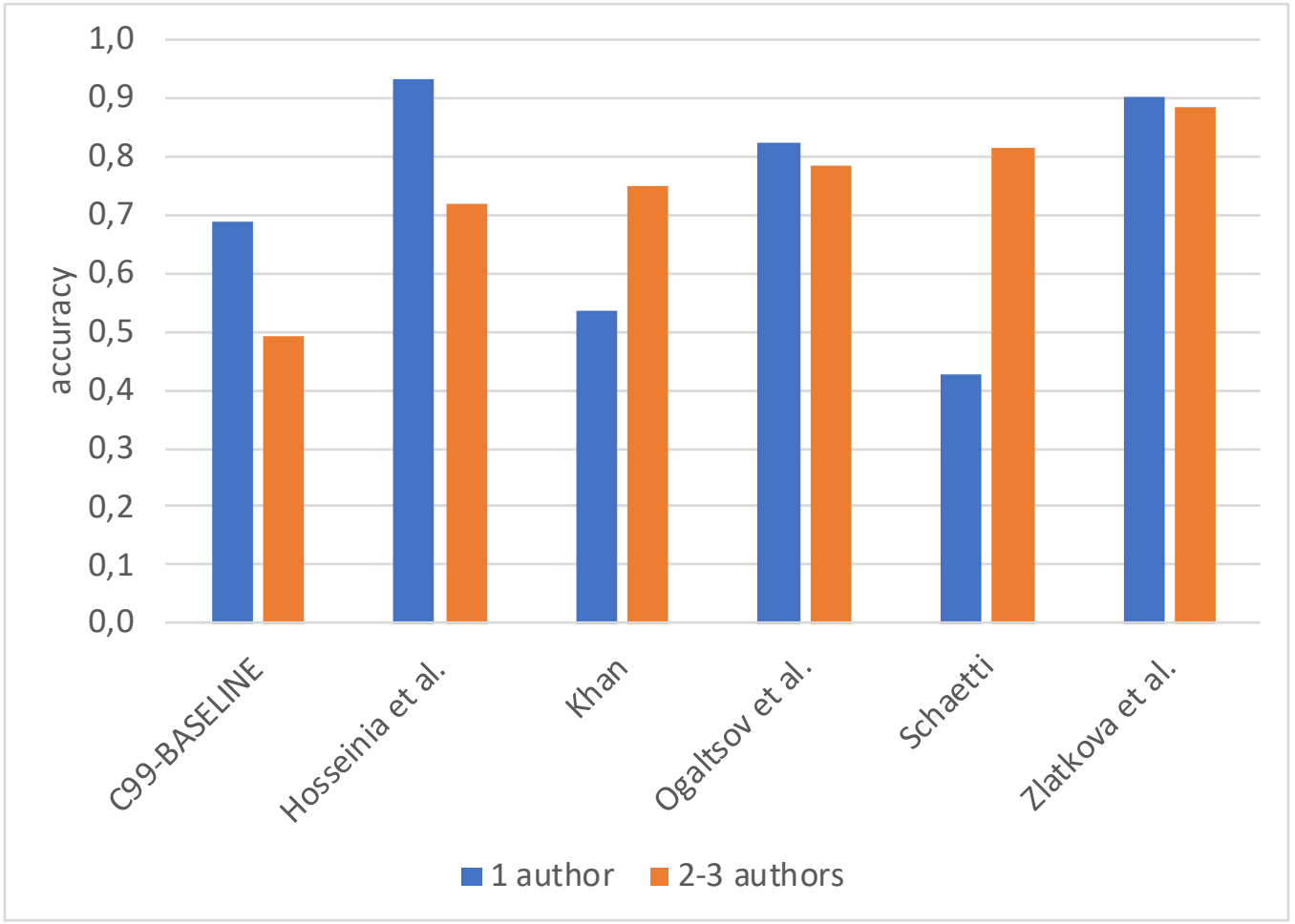
- Results

Submission	Accuracy	Runtime
Zlatkova et al.	**0.893**	01:35:25
Hosseinia & Mukherjee	0.825	10:12:28
Safin & Ogaltsov	0.803	00:05:15
Khan	0.643	00:01:10
Schaetti	0.621	00:03:36
BASELINE-C99	0.589	00:00:16
BASELINE-rnd2	0.560	-
BASELINE-rnd1	0.500	-

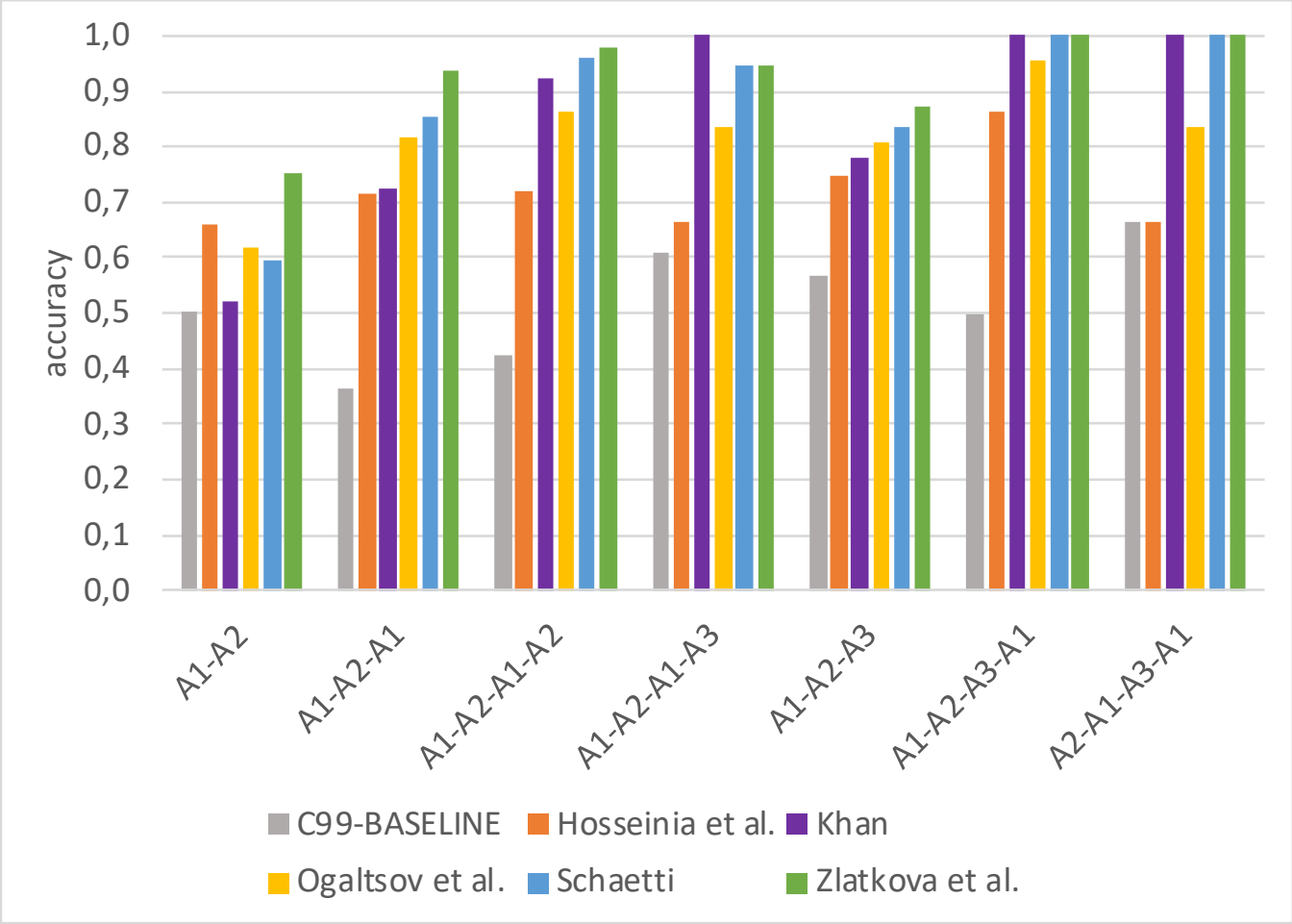
Results (#tokens)



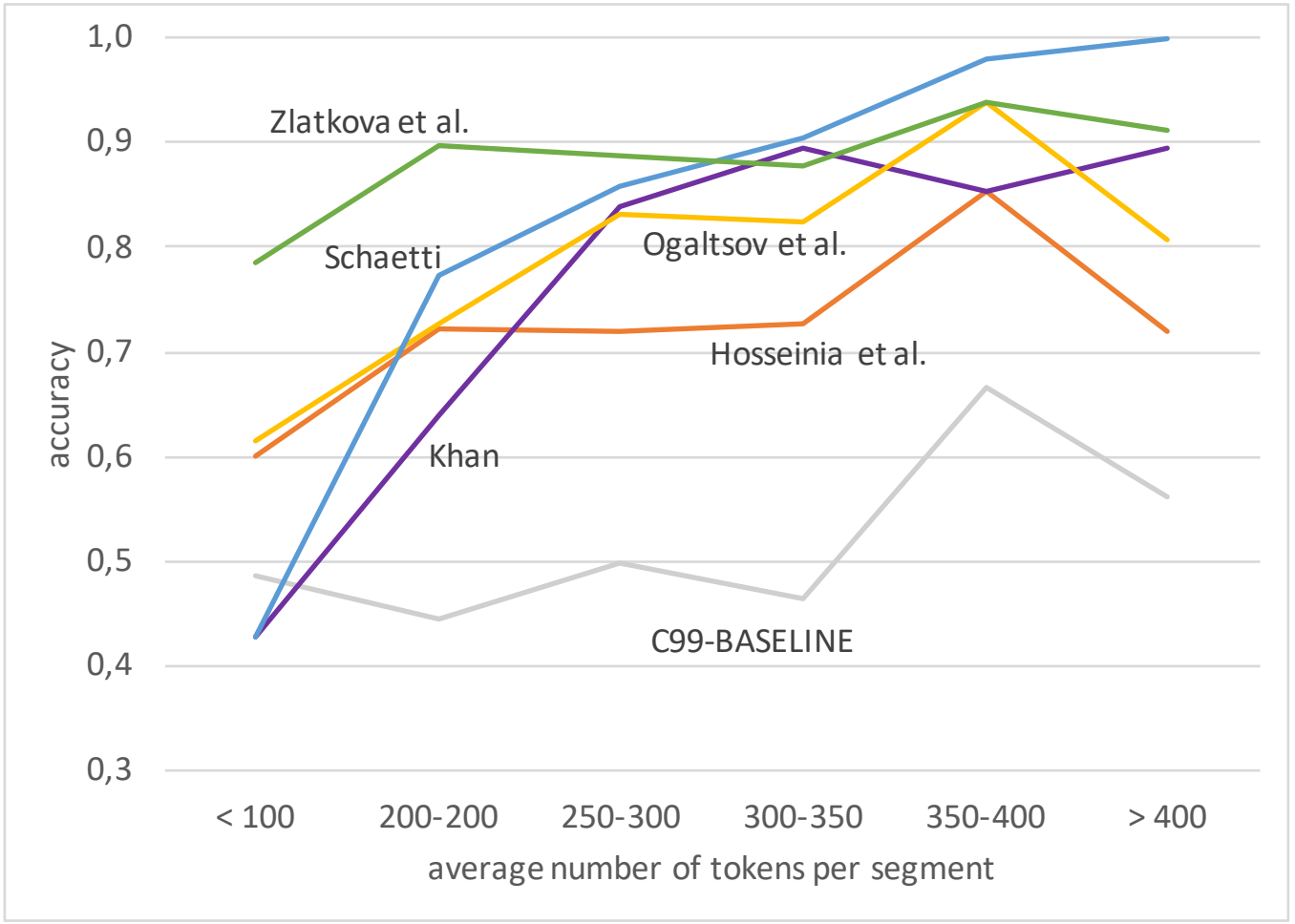
Results (#authors)



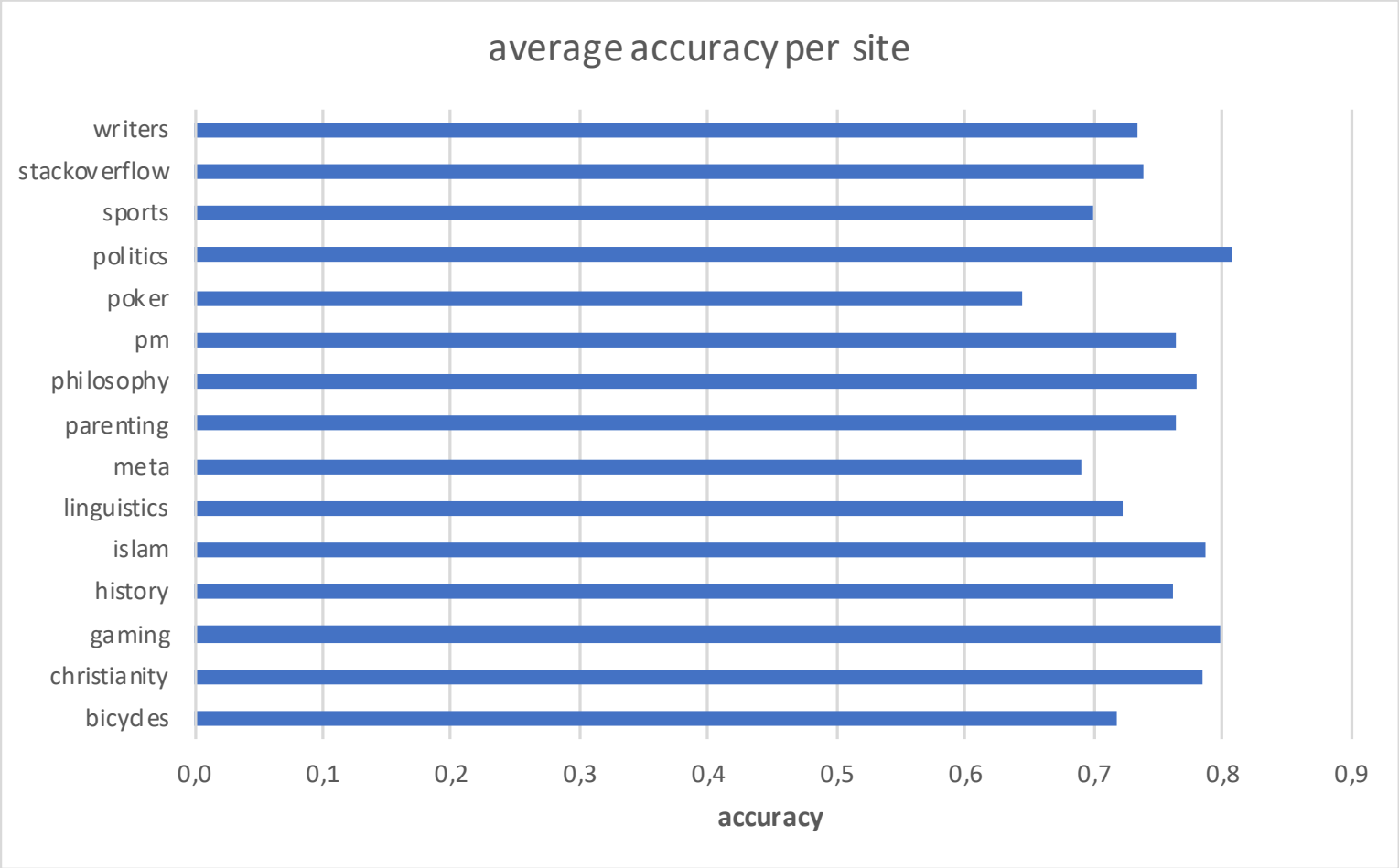
Results (author distribution)



Results (segment length)



Results (topic)



Results wrt subtopics

	subtopic (topic)	#docs	BASELINE-C99	Hosseinia et al.	Khan	Ogaltsov et al.	Schaetti	Zlatkova et al.	avg
best 10 subtopics	starcraft-2 (gaming)	12	0.67	0.83	0.92	**1.00**	**1.00**	0.92	0.93
	political-history (history)	14	0.43	**0.93**	0.86	**0.93**	**0.93**	**0.93**	0.91
	history (christianity)	10	0.80	0.90	0.80	**1.00**	0.80	0.90	0.88
	halal-haram (islam)	10	0.80	**1.00**	0.70	**1.00**	0.70	**1.00**	0.88
	war (history)	8	0.38	0.75	**1.00**	0.88	0.75	**1.00**	0.88
	economy (politics)	8	0.25	**1.00**	0.75	0.88	0.75	**1.00**	0.88
	exegesis (christianity)	22	0.73	0.91	0.82	0.91	0.73	**0.95**	0.86
	prophet-muhammad (islam)	8	0.63	**1.00**	0.88	0.88	0.50	**1.00**	0.85
	syntax (linguistics)	14	0.79	0.86	0.79	0.86	0.71	**1.00**	0.84
	election (politics)	10	0.60	0.80	**0.90**	**0.90**	0.70	**0.90**	0.84
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
worst 10 subtopics	feature-request (meta)	20	0.45	0.75	0.60	0.75	0.60	**0.80**	0.70
	discipline (parenting)	10	0.30	**0.90**	0.70	0.50	0.60	0.80	0.70
	scrum (pm)	14	**0.93**	**0.93**	0.43	0.79	0.57	0.79	0.70
	ancient-rome (history)	12	0.42	0.83	0.67	0.75	0.25	**0.92**	0.68
	lds (christianity)	14	0.43	0.71	0.57	**0.79**	0.50	0.71	0.66
	fiction (writers)	14	0.36	0.86	0.36	**0.93**	0.29	0.86	0.66
	nature-of-god (christianity)	12	0.75	0.75	0.50	0.67	0.42	**0.83**	0.63
	english (linguistics)	8	0.50	**0.75**	0.63	0.50	0.63	0.50	0.60
	world-war-two (history)	26	0.62	0.73	0.42	0.54	0.46	**0.81**	0.59
	support (meta)	10	0.40	**0.70**	0.50	0.60	0.40	**0.70**	0.58

universität innsbruck

DBIS
DATABASES AND INFORMATION SYSTEMS

Thank you.