universität
innsbruck

PAN@CLEF 2020
Style Change Detection Task

Eva Zangerle, Maximilian Mayerl, Günther Specht, Martin Potthast, Benno Stein

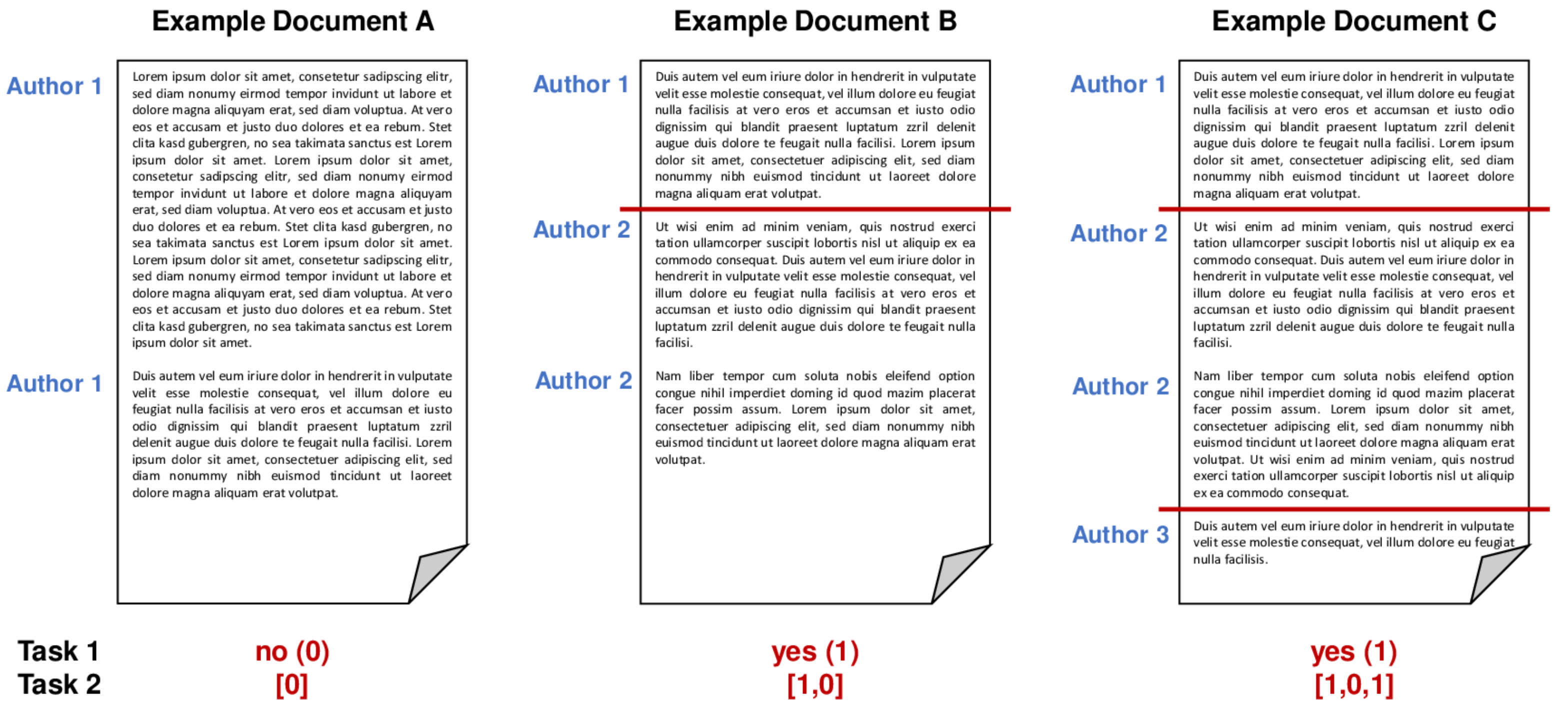
Task Description

Given a document, participants should answer the following questions:

(a) Is the document written by one or more authors, i.e., do style changes exist or not?

(b) Between which consecutive paragraphs in the document do style changes occur?

Task Description



Example Document A

Author 1

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

Author 1

Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi. Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat.

Example Document B

Author 1

Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi. Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat.

Author 2

Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat. Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi.

Author 2

Nam liber tempor cum soluta nobis eleifend option congue nihil imperdiet doming id quod mazim placerat facer possim assum. Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat.

Example Document C

Author 1

Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi. Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat.

Author 2

Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat. Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi.

Author 2

Nam liber tempor cum soluta nobis eleifend option congue nihil imperdiet doming id quod mazim placerat facer possim assum. Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat. Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat.

Author 3

Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis.

	Example Document A	Example Document B	Example Document C
Task 1	no (0)	yes (1)	yes (1)
Task 2	[0]	[1,0]	[1,0,1]

Dataset

- Realistic, non-artificial and comprehensive dataset
- 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

- Q&A platform **StackExchange** fulfills these requirements

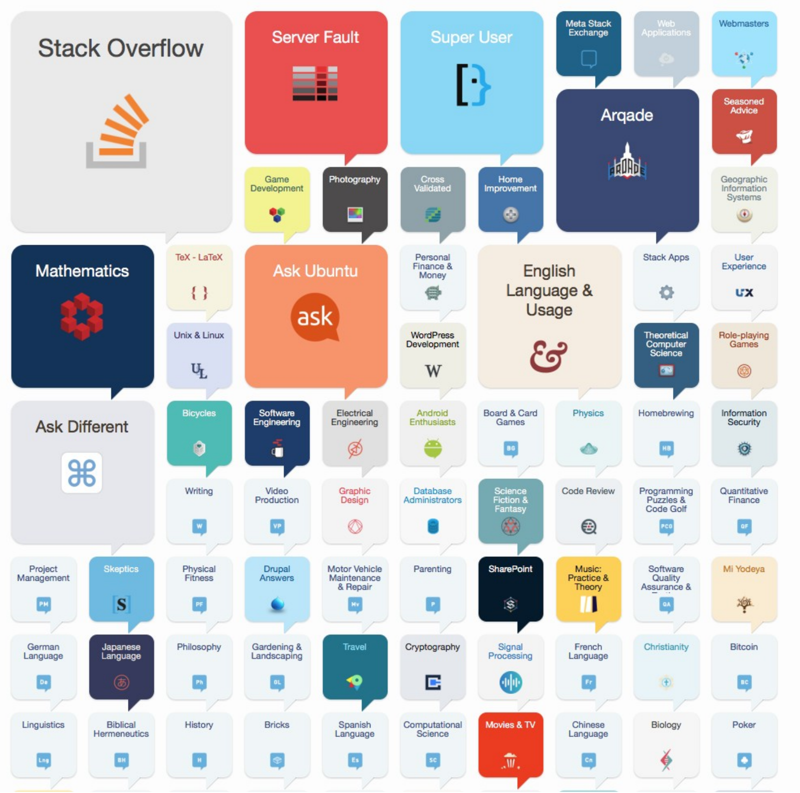
Dataset

StackExchange consists of several sites (176 sites), data freely available

Each question/answer is associated with a site, giving it a broad topic.

Example sites:

- data science
- economics
- literature
- philosophy



Dataset

- Cleaning
  - Remove links
  - Remove images
  - Remove code snippets
  - Remove bullet lists
  - Remove block quotes
  - Remove very short questions/answers
  - Remove edited questions/answers
  - Remove questions/answers not written in English
- Using the raw texts, a **training** (50%), **validation** (25%) and **test** (25%) dataset has been created
- Each dataset contains 50% single-author documents and 50% multi-authored documents

Parameters

Parameter	Configuration Options
Number of style changes	0-10
Number of collaborating authors	1-3
Document length	1,000-3,000 tokens
Change positions	between paragraphs
Document language	English

Dataset

Two datasets for the task, differing in how broad the range of topics included in them is:

- `dataset-narrow`: questions/answers from 12 sites, covering topics related to computing technology
- `dataset-wide`: questions/answers from 25 sites, covering a wide range of topics, including astronomy, economics, history, linguistics, mathematics, etc.

Evaluation

- F1 score
- Score for a subtask: average of scores for both dataset
- Overall score: average of the scores for the subtasks

Approaches

3 submissions to TIRA, 2 submitted working notes papers:

Mixed Style Feature Representation and B-maximal Clustering (Castro-Castro et al.)

- 185 stylometric features: character-based/lexical/syntactic features, explicitly excluding features which capture the semantics of the text
- Similarity between paragraphs = number of similar features in both paragraphs
- Cluster paragraphs into authors using B0-maximal clustering

Style Change Detection Using BERT (Iyer and Vosoughi)

- Use BERT as a feature extractor to describe paragraphs and documents
- Random Forest classifiers

Baseline

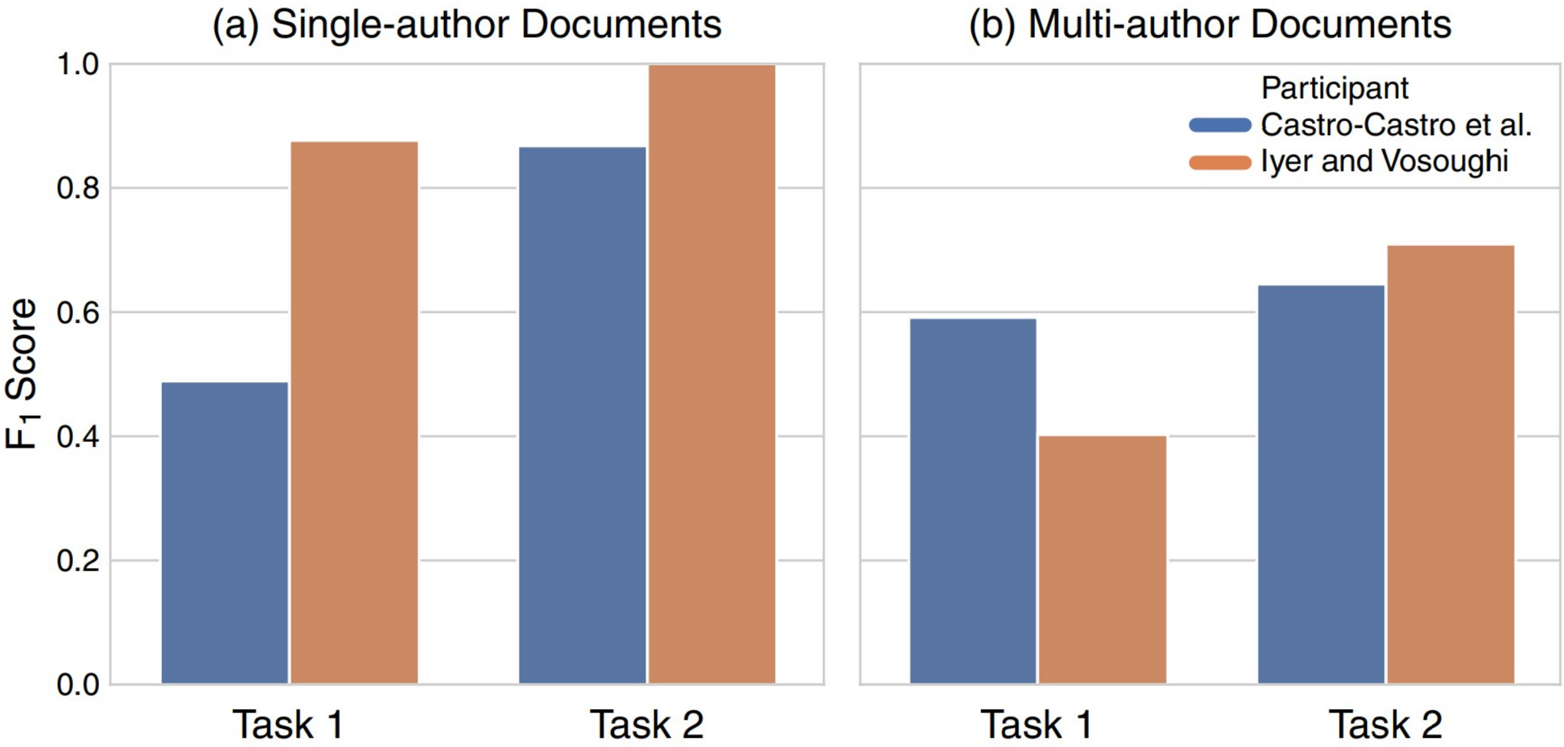
We also evaluated a simple random baseline:

- Task 1: randomly predict the document to be single- or multi-authored (equal chance)
- Task 2: randomly predict there to be a style change between any pair of consecutive paragraphs (equal chance)

Results

Participant	Task 1 (F1)	Task 2 (F1)	Average (F1)
Iyer and Vosoughi	0.6401	0.8567	0.7484
Castro-Castro et al.	0.5399	0.7579	0.6489
Nath	0.5204	0.7526	0.6365
Baseline (random)	0.5007	0.5001	0.5004

Single- vs Multi-author Documents



Impact of Topical Breadth

Participant	Task 1 Narrow	Task 1 Wide	Task 2 Narrow	Task 2 Wide
Iyer and Vosoughi	0.7042	0.5760	0.8823	0.8310
Castro-Castro et al.	0.5379	0.5419	0.8242	0.6915

Conclusion

- Style change detection task
- Two subtasks were tackled
- Unfortunately only two submissions
- For next year: Repeat the same type of task with a dataset that has stronger topical coherence within its documents.
  - We are looking forward to your participation!