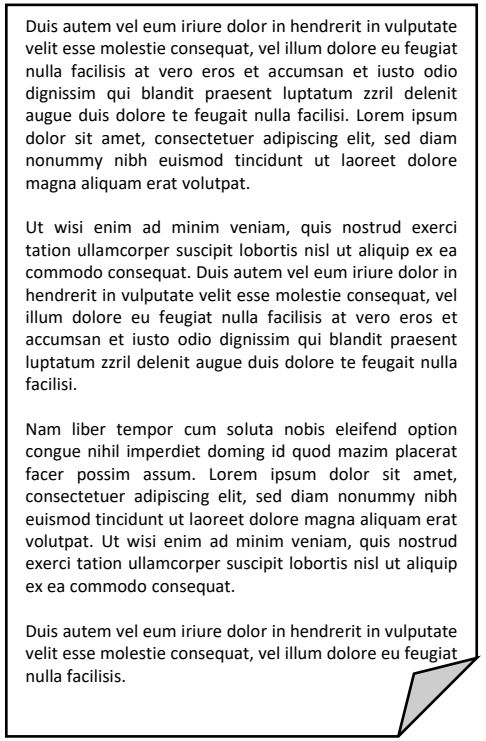
PAN@CLEF 2022
Style Change Detection Task

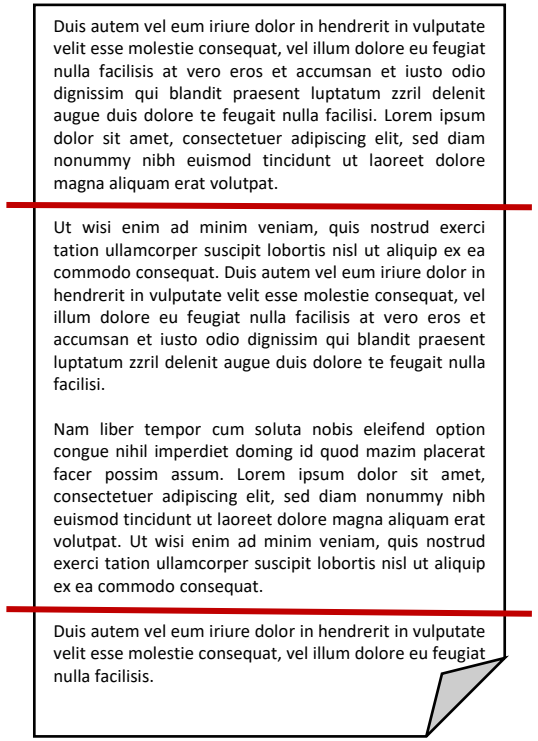
Eva Zangerle, Maximilian Mayerl, Martin Potthast, Benno Stein

Given a document,
find the **positions**
at which the **authorship changes**.

Style Change Detection Task



Style Change Detection Task



Looking back…

2017	Is the document multi-authored? If yes, find the positions of switches.
2018	Is the document written by a single or multiple authors?
2019	Is the document multi-authored? If yes, find the positions of switches. How many authors have collaborated on the document?
2020	Is the document multi-authored? For each pair of paragraphs, is there a style change between them?
2021	Is the document multi-authored? For each pair of paragraphs, is there a style change between them? Assign all paragraphs to an author id.

Task Description

[Task1] Style Change Basic: for a text written by two authors that contains a single style change only, find the position of this change (i.e., cut the text into the two authors' texts on the paragraph-level),

Task Description

[Task2] Style Change Advanced: for a text written by two or more authors, find all positions of writing style change (i.e., assign all paragraphs of the text uniquely to some author out of the number of authors assumed for the multi-author document)

Task Description

[Task3] Style Change Real-World: for a text written by two or more authors, find all positions of writing style change, where style changes now not only occur between paragraphs, but at the sentence level.

Task Description

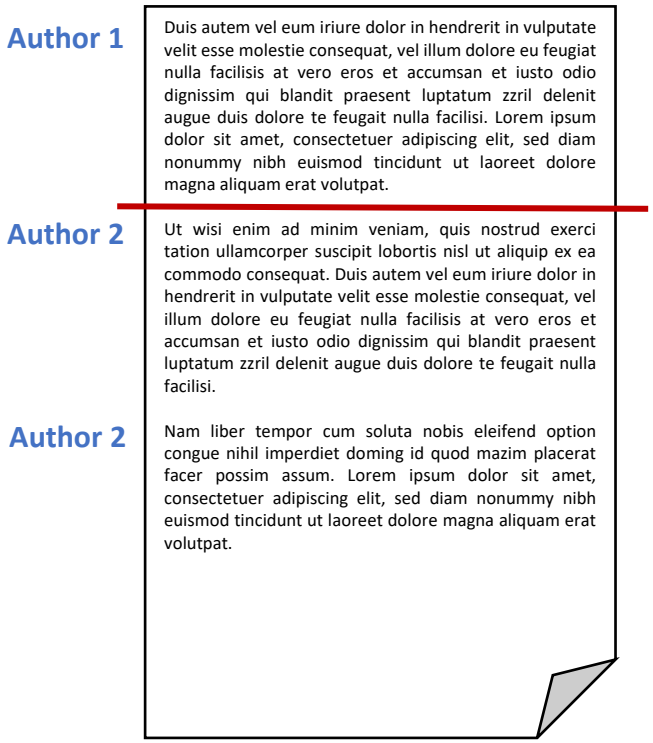
[Task1] Style Change Basic: for a text written by two authors that contains a single style change only, find the position of this change (i.e., cut the text into the two authors' texts on the paragraph-level),

[Task2] Style Change Advanced: for a text written by two or more authors, find all positions of writing style change (i.e., assign all paragraphs of the text uniquely to some author out of the number of authors assumed for the multi-author document)

[Task3] Style Change Real-World: for a text written by two or more authors, find all positions of writing style change, where style changes now not only occur between paragraphs, but at the sentence level.

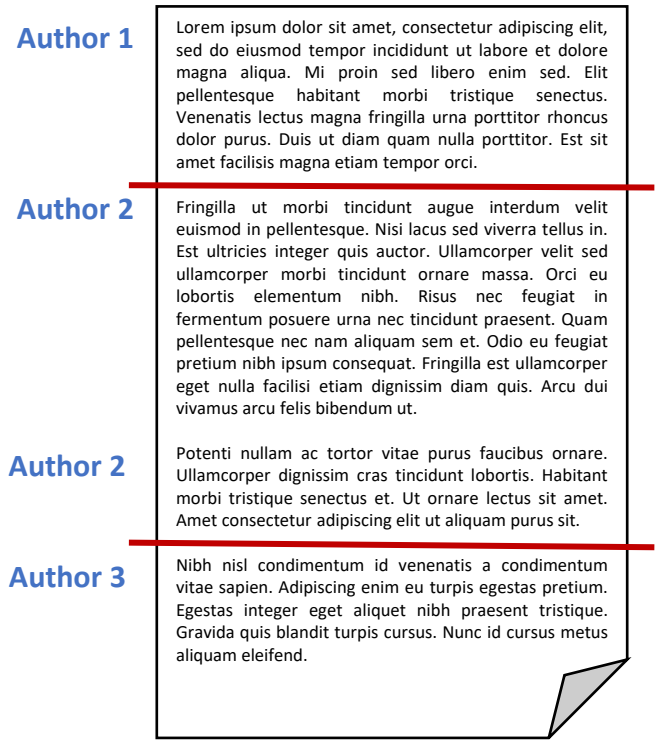
Task Description

Example Document A



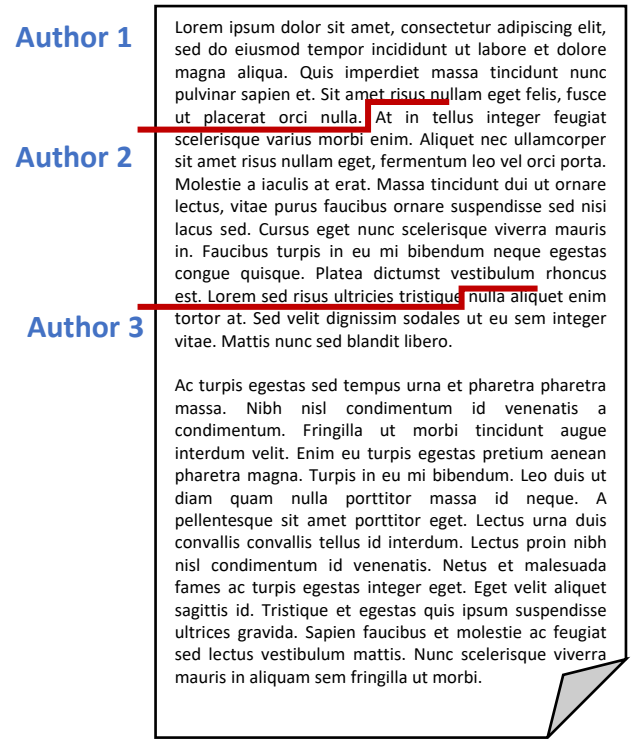
Task 1: [1,0]
Task 2: [1,2,2]

Example Document B



Task 2: [1,2,2,3]

Example Document C

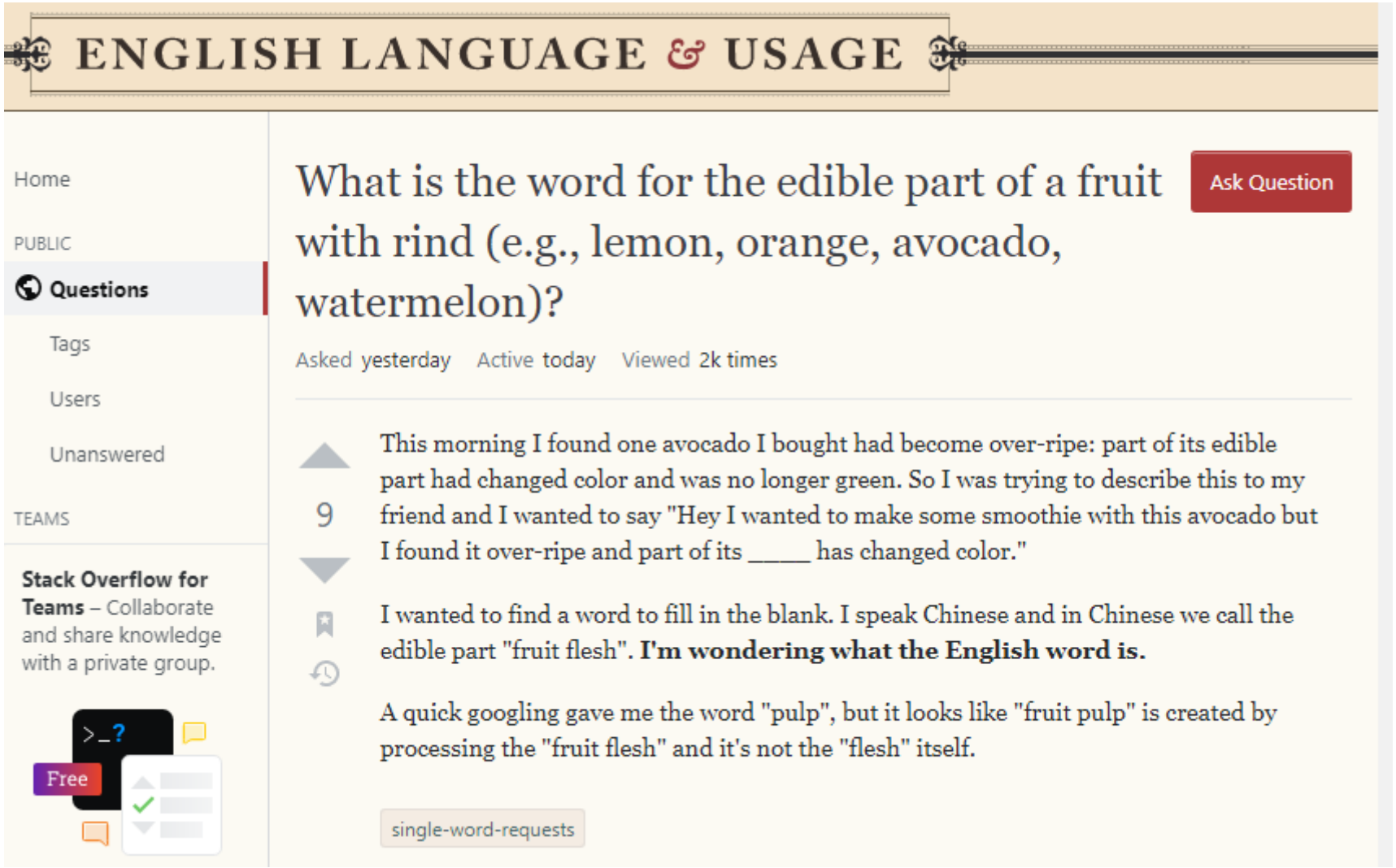


Task 3: [0,0,1,0,0,0,0,0,0,1,....]

Dataset

- Realistic, non-artificial and comprehensive dataset
- Requirements
    - Find multiple authors that write about the same topic
    - Multi-authored texts need to contain the same topic / subtopic
    - Find texts that are freely available and of sufficient length

- StackExchange consists of several sites (topics), data freely available
- Each question/answer associated with
    - Site
    - Set of tags (subtopics)
- Example: photography site – subtopics: lens, canon, nikon, lightroom, photoshop, …


ENGLISH LANGUAGE & USAGE
Home
PUBLIC
Questions
Tags
Users
Unanswered
TEAMS
Stack Overflow for Teams – Collaborate and share knowledge with a private group.
Free
What is the word for the edible part of a fruit with rind (e.g., lemon, orange, avocado, watermelon)?
Ask Question
Asked yesterday Active today Viewed 2k times
9
This morning I found one avocado I bought had become over-ripe: part of its edible part had changed color and was no longer green. So I was trying to describe this to my friend and I wanted to say "Hey I wanted to make some smoothie with this avocado but I found it over-ripe and part of its _____ has changed color."
I wanted to find a word to fill in the blank. I speak Chinese and in Chinese we call the edible part "fruit flesh". I'm wondering what the English word is.
A quick googling gave me the word "pulp", but it looks like "fruit pulp" is created by processing the "fruit flesh" and it's not the "flesh" itself.
single-word-requests

34

As the post suggests it can be called the

> **flesh**
>
> NOUN
>
> **1.2** The edible pulpy part of a fruit or vegetable.
> *halve the avocados and scrape out the flesh*

Calling it 'pulpy' doesn't mean it has been squashed. It means the soft part that isn't the skin/rind or pips/stone.

You don't have to call it 'fruit flesh', just 'flesh' will do when the fruit has already been mentioned.

> Hey I wanted to make some smoothie with this avocado but I found it over-ripe and part of its **flesh** has changed color.

From Lexico.

Dataset

- Cleaning
  - Remove links, images, code snippets, bullet lists, block quotes
  - Remove very short questions/answers
  - Remove edited questions/answers
- Randomly choose question/answer thread
- Randomly choose number of authors n in [1,5]
- Randomly select n authors
- Take their paragraphs in thread, shuffle, concatenate
- Check length of document (> 2 paragraphs, between 1,000 and 10,000 characters)

Dataset

Parameter	Setting
Number of authors	1-5
Document length	1,000-10,000 characters
Minimum paragraph length	100 characters
Minimum number of paragraphs	2
Change positions	Between paragraphs (tasks 1 and 2) Between sentences (task 3)
Document language	English

Task 1: 2,000 documents
Tasks 2 and 3: 10,000 documents
70:15:15 training/validation/test split

Evaluation

- Score each document by F_1 measure (macro-averaged)
- Three tasks evaluated independently

- Additionally for tasks 2 and 3:
  - Diarization Error Rate (DER): fraction of wrongly attributed paragraphs
  - Jaccard Error Rate (JER): author-specific error rate

$$JER(a) = 1.0 - \frac{|ref_a \cap pred_a|}{|ref_a \cup pred_a|}$$

Submissions

- Nine submissions (eight intrinsic, one extrinsic)

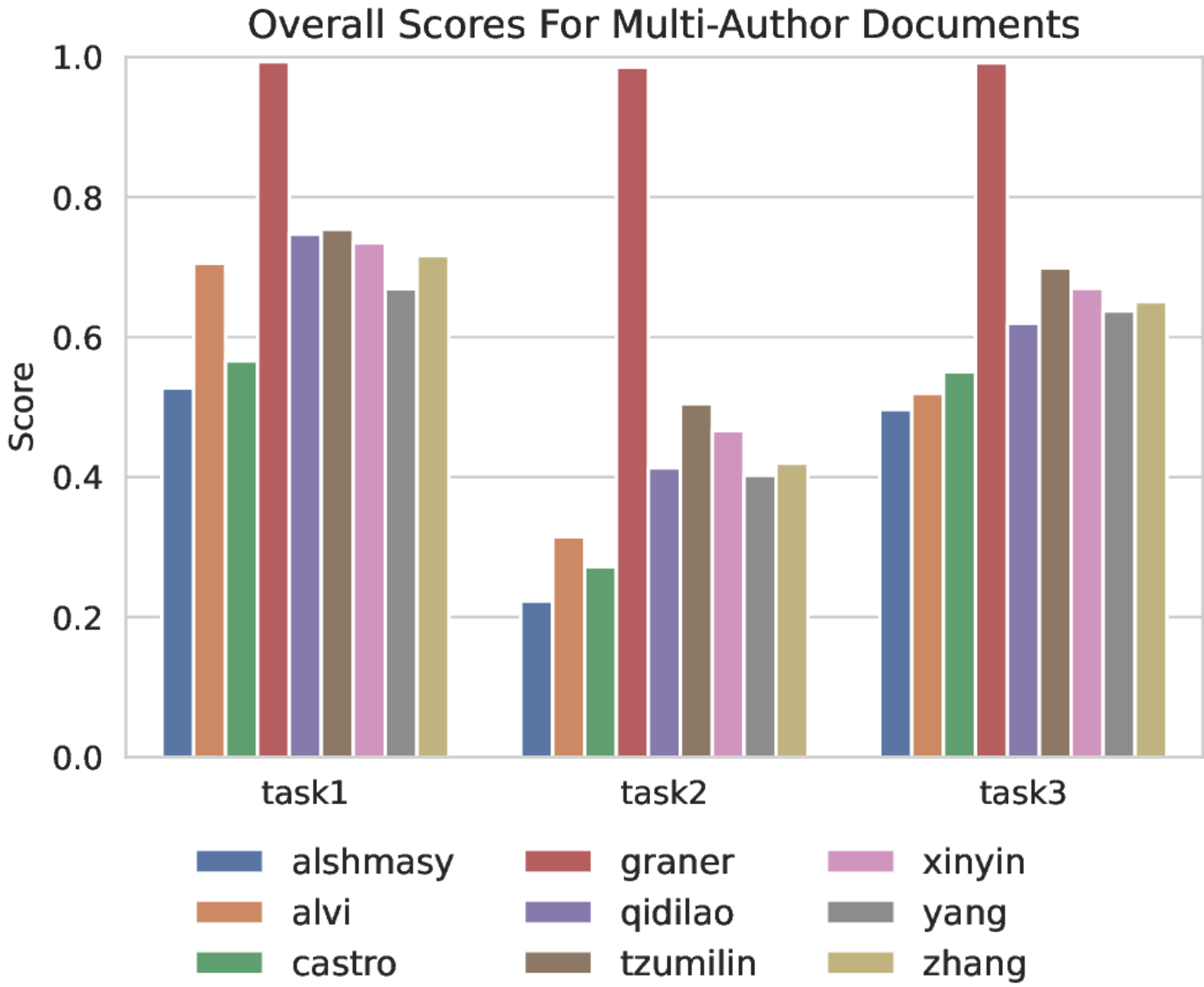
- Alshamasi and Menai (alshmasy): lexical and syntactic features, k-means clustering
- Alvi et al. (alvi): handcrafted discourse markers as features
- Jiang et al. (xinyin): Electra (masked LM), one model per task
- Lao et al. (qidilao): BERT, convolution layer, max pooling for binary class prediction
- Lin et al. (tzumilin): majority-vote ensemble of three LLMs (BERT, RoBERTa, ALBERT)
- Rodriguez-Losada and Castro (castro): transformer and lexical features, similarity for final prediction
- Zhang et al. (zhang): BERT (masked training) and prompt modeling for final prediction
- Zi et al. (yang): BERT (masked training), Bi-LSTM to capture context, fully connected layer for predictions
- Graner and Ranly (graner): extrinsic approach

Submissions

- Nine submissions (eight intrinsic, one extrinsic)

- Alshamasi and Menai (alshmasy): lexical and syntactic features, k-means clustering
- Alvi et al. (alvi): handcrafted discourse markers as features
- Jiang et al. (xinyin): Electra (masked LM), one model per task
- Lao et al. (qidilao): BERT, convolution layer, max pooling for binary class prediction
- Lin et al. (tzumilin): majority-vote ensemble of three LLMs (BERT, RoBERTa, ALBERT)
- Rodriguez-Losada and Castro (castro): transformer and lexical features, similarity for final prediction
- Zhang et al. (zhang): BERT (masked training) and prompt modeling for final prediction
- Zi et al. (yang): BERT (masked training), Bi-LSTM to capture context, fully connected layer for predictions
- Graner and Ranly (graner): extrinsic approach

Results

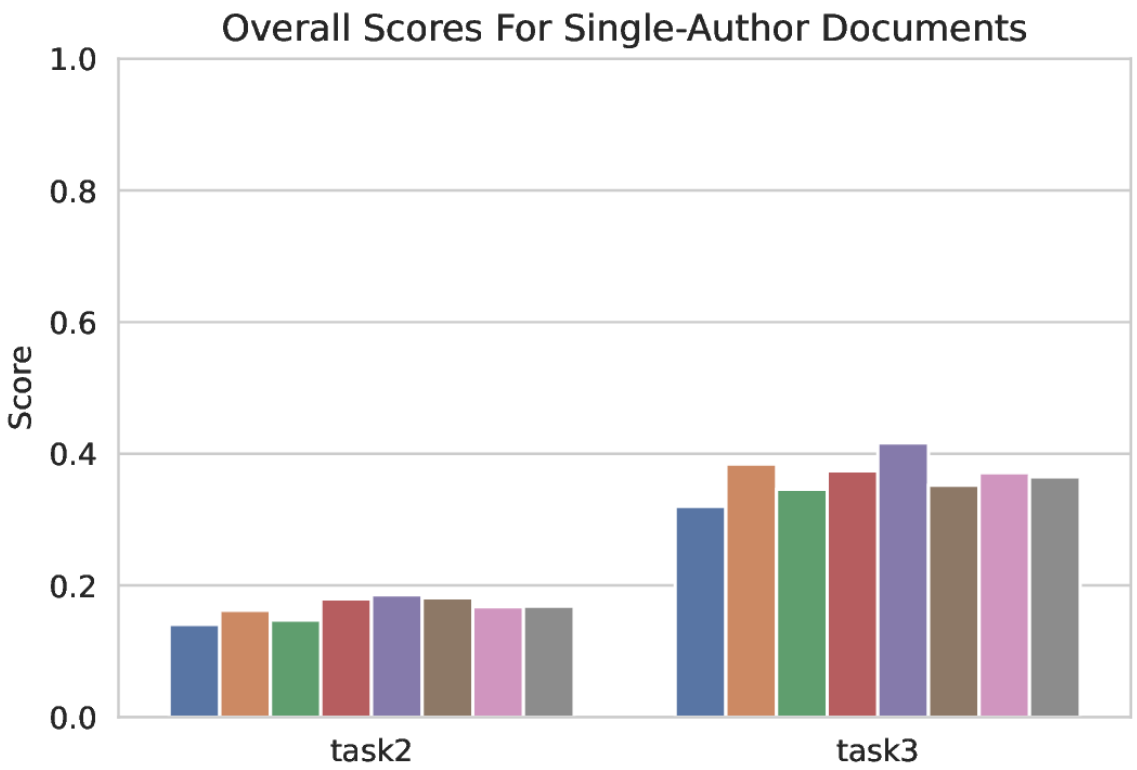

Overall Scores For Multi-Author Documents
Score
1.0
0.8
0.6
0.4
0.2
0.0
task1
task2
task3
alshmasy
alvi
castro
graner
qidilao
tzumilin
xinyin
yang
zhang

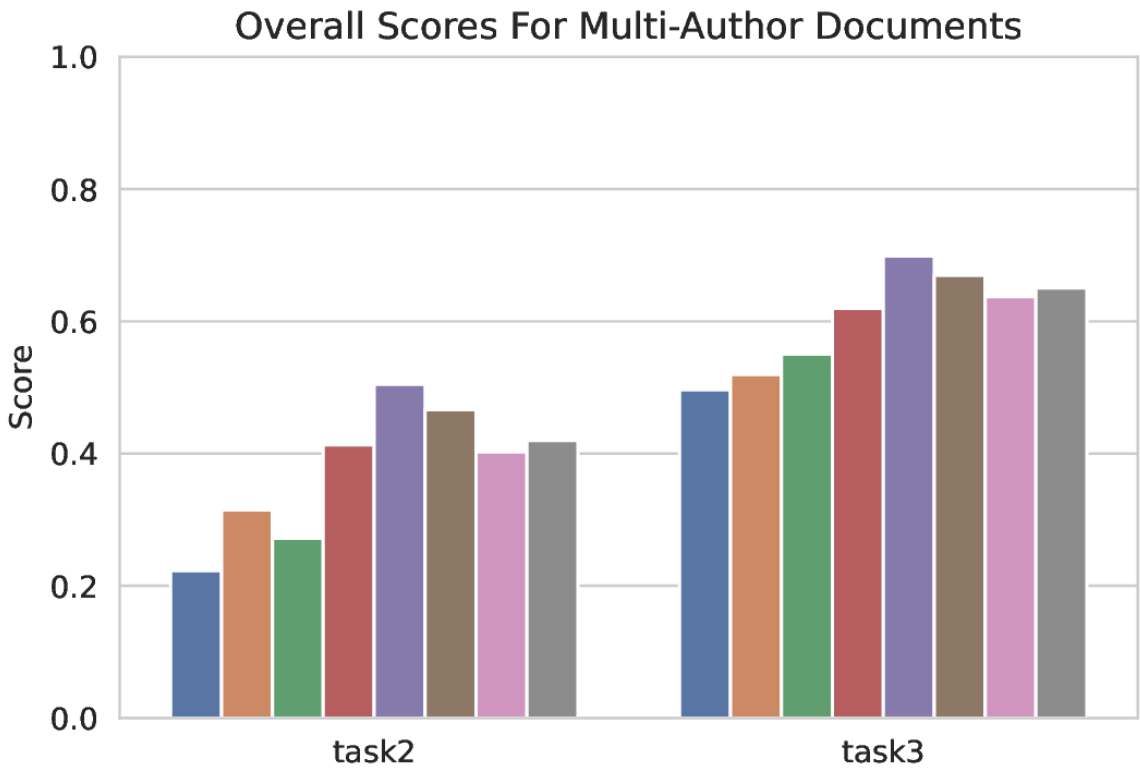
Overview

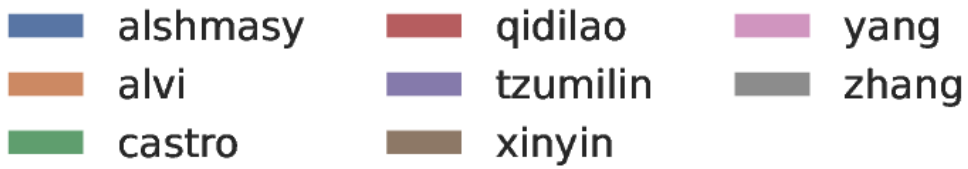
Participant	ST1 F_1	ST2 F_1	ST3 F_1	ST3 DER	ST3 JER
Intrinsic Approaches					
tzumilin22	**0.7540**	**0.5100**	**0.7156**	**0.1941**	**0.3095**
xinyin22	0.7346	0.4687	0.6720	0.2380	0.3138
qidilao22	0.7471	0.4170	0.6314	0.2636	0.3641
zhang22	0.7162	0.4174	0.6581	0.2886	0.3556
yang22	0.6690	0.4011	0.6483	0.2964	0.3677
alvi22	0.7052	0.3213	0.5636	0.3924	0.5218
castro22a	0.5661	0.2735	0.5565	0.4035	0.5771
alshmasy22	0.5272	0.2207	0.4995	0.4240	0.6444
Extrinsic Approaches					
graner22	0.9932	0.9855	0.9929	0.0040	0.0040
Baseline					
Random	0.3222	0.2651	0.4809	0.4568	0.5938

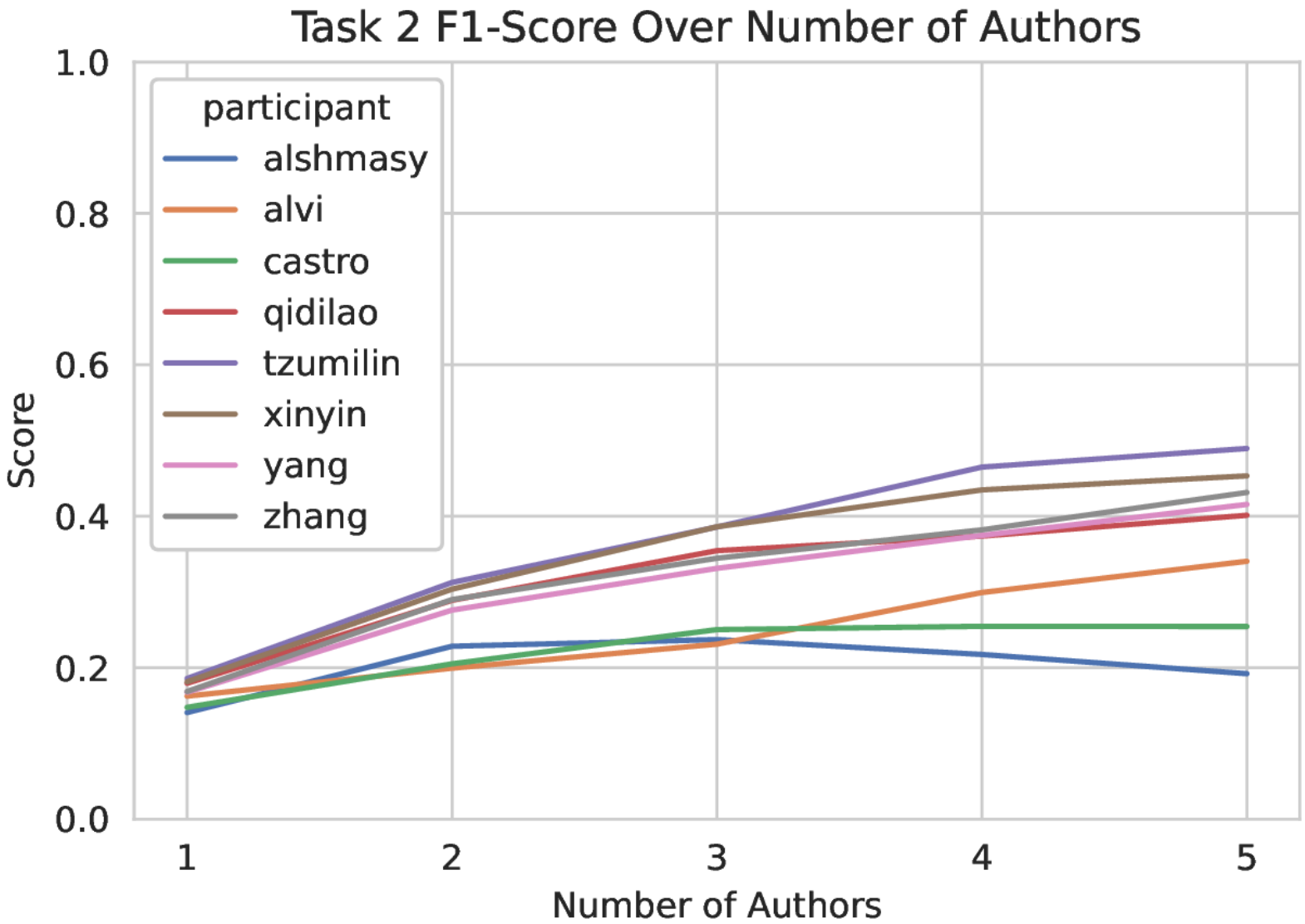
Overview

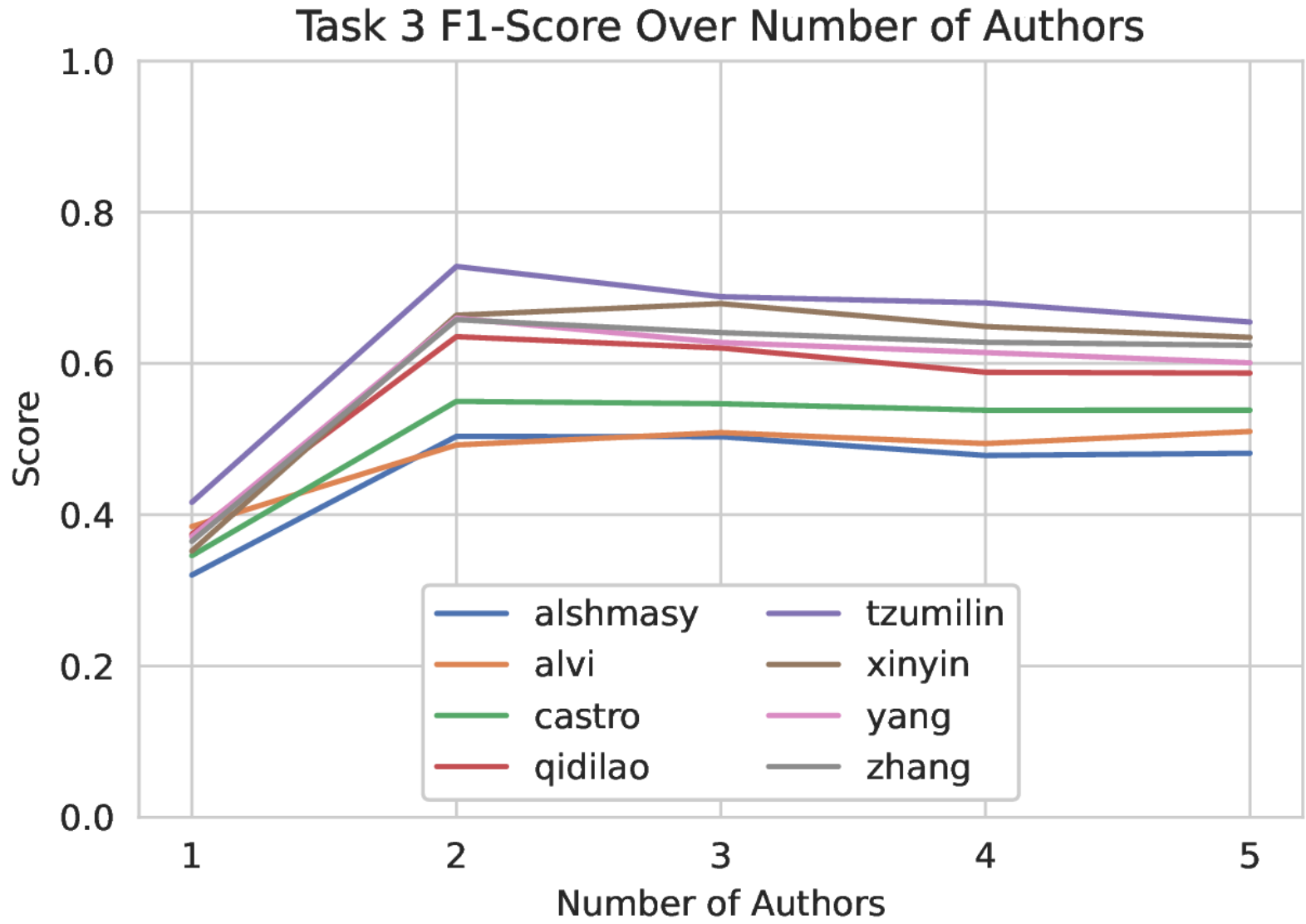
	Participant	ST1 F_1	ST2 F_1	ST3 F_1	ST3 DER	ST3 JER
	Intrinsic Approaches					
LLM	tzumilin22	**0.7540**	**0.5100**	**0.7156**	**0.1941**	**0.3095**
LLM	xinyin22	0.7346	0.4687	0.6720	0.2380	0.3138
LLM	qidilao22	0.7471	0.4170	0.6314	0.2636	0.3641
LLM	zhang22	0.7162	0.4174	0.6581	0.2886	0.3556
LLM	yang22	0.6690	0.4011	0.6483	0.2964	0.3677
	alvi22	0.7052	0.3213	0.5636	0.3924	0.5218
	castro22a	0.5661	0.2735	0.5565	0.4035	0.5771
	alshmasy22	0.5272	0.2207	0.4995	0.4240	0.6444
	Extrinsic Approaches					
	graner22	0.9932	0.9855	0.9929	0.0040	0.0040
	Baseline					
	Random	0.3222	0.2651	0.4809	0.4568	0.5938


Overall Scores For Single-Author Documents
Score
1.0
0.8
0.6
0.4
0.2
0.0
task2
task3


Overall Scores For Multi-Author Documents
Score
1.0
0.8
0.6
0.4
0.2
0.0
task2
task3


alshmasy
alvi
castro
qidilao
tzumilin
xinyin
yang
zhang


Task 2 F1-Score Over Number of Authors
participant
alshmasy
alvi
castro
qidilao
tzumilin
xinyin
yang
zhang
Score
1.0
0.8
0.6
0.4
0.2
0.0
1
2
3
4
5
Number of Authors


Task 3 F1-Score Over Number of Authors
Score
1.0
0.8
0.6
0.4
0.2
0.0
1
2
3
4
5
Number of Authors
alshmasy
alvi
castro
qidilao
tzumilin
xinyin
yang
zhang

Outlook PAN@CLEF 2023

- Style change detection often benefited from topic changes within documents
- Control for change of authorship and topic simultaneously

  → control for topic and style similarity in assembling documents

  → Uncover ability to capture/model style variations

- New dataset based on Reddit
- New "perspective" on task complexity: single task carried out on three datasets of increasing task complexity

Outlook PAN@CLEF 2023

Core task: Detect change of authorship on the paragraph level

1. Easy dataset: paragraphs cover a variety of topics
2. Medium dataset: topical variety is limited, focus on style information
3. Hard dataset: all paragraphs cover the same topic, focus only on style information

Outlook PAN@CLEF 2023

Core task: Detect change of authorship on the paragraph level

1. Easy dataset: paragraphs cover a variety of topics
2. Medium dataset: topical variety is limited, focus on style information
3. Hard dataset: all paragraphs cover the same topic, focus only on style information

Outlook PAN@CLEF 2023

Core task: Detect change of authorship on the paragraph level

1. Easy dataset: paragraphs cover a variety of topics
2. Medium dataset: topical variety is limited, focus on style information
3. Hard dataset: all paragraphs cover the same topic, focus only on style information

Any feedback welcome ☺

See you at