TOUCHÉ 2023

CLEF 2023 Thessaloniki

Touché @ CLEF 2023

Argument and Causal Retrieval

Alexander Bondarenko
Maik Fröbe
Johannes Kiesel
Ferdinand Schlatt
Valentin Barriere
Brian Ravenet
Léo Hemamou
Simon Luck
Jan Heinrich Reimer
Benno Stein
Martin Potthast
Matthias Hagen

[touche.webis.de]

Touché: Argument and Causal Retrieval

Shared Tasks

Task 1: Argument Retrieval for Controversial Questions

- ❑ Retrieve relevant and high-quality argumentative documents, detect stance

Touché: Argument and Causal Retrieval

Shared Tasks

Task 1: Argument Retrieval for Controversial Questions

- ❑ Retrieve relevant and high-quality argumentative documents, detect stance

Task 2: Evidence Retrieval for Causal Questions

- ❑ Retrieve and rank causality-related documents and detect causal stance

Touché: Argument and Causal Retrieval

Shared Tasks

TOUCHÉ
2023

Task 1: Argument Retrieval for Controversial Questions

- ❑ Retrieve relevant and high-quality argumentative documents, detect stance

Task 2: Evidence Retrieval for Causal Questions

- ❑ Retrieve and rank causality-related documents and detect causal stance

Task 3: Image Retrieval for Arguments

- ❑ Retrieve images for each stance (pro / con) that support that stance

Touché: Argument and Causal Retrieval

Shared Tasks

Task 1: Argument Retrieval for Controversial Questions

- ❑ Retrieve relevant and high-quality argumentative documents, detect stance

Task 2: Evidence Retrieval for Causal Questions

- ❑ Retrieve and rank causality-related documents and detect causal stance

Task 3: Image Retrieval for Arguments

- ❑ Retrieve images for each stance (pro / con) that support that stance

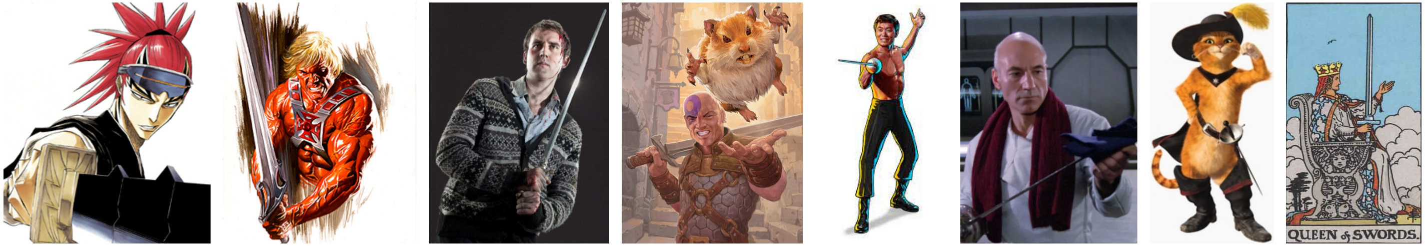
Task 4: Multilingual and Multi-target Stance Classification

- ❑ Detect the stance of a comment on a proposal

Touché: Argument and Causal Retrieval

Lab Statistics

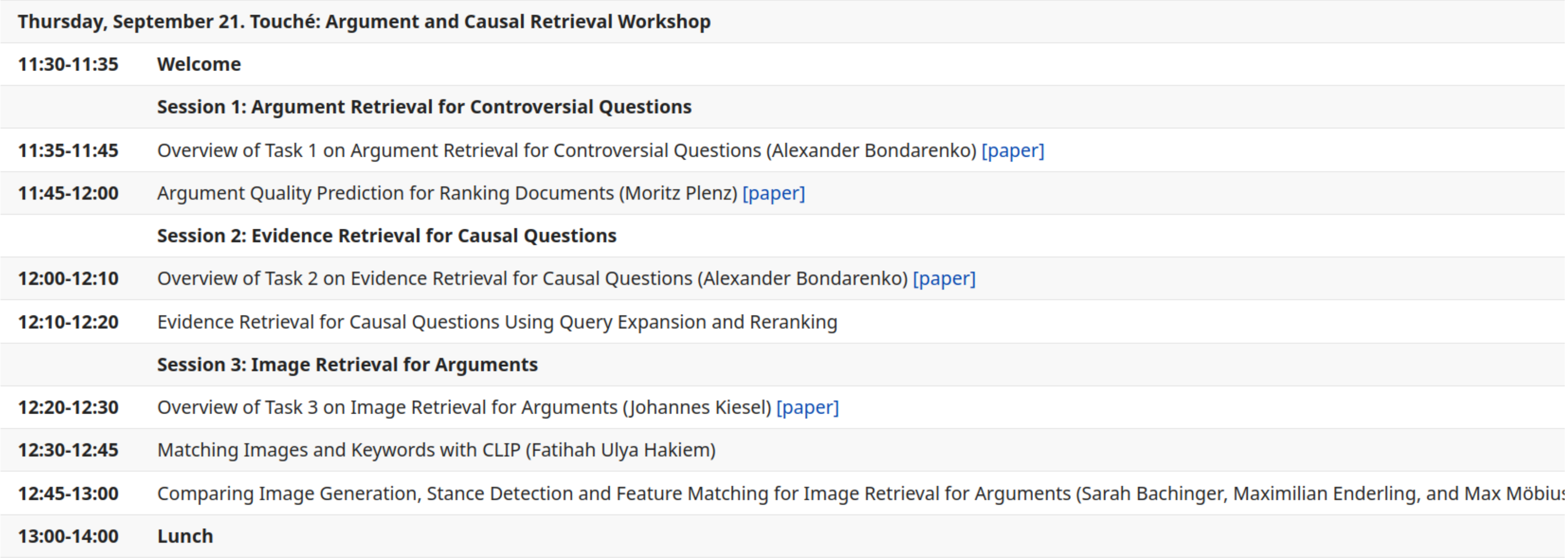
- Registrations: 41 teams (vs. 58 teams last year)
- Nicknames: Real or fictional fencers / swordfighters (e.g., Zorro)
- Submissions: 7 participating teams (vs. 23 last year)
- Approaches: 30 valid runs were evaluated (vs. 84 last year)
- Judgments: 1 500 web documents, 700 images, 25 000 comments



Touché: Argument and Causal Retrieval

Workshop Program

[touche.webis.de]

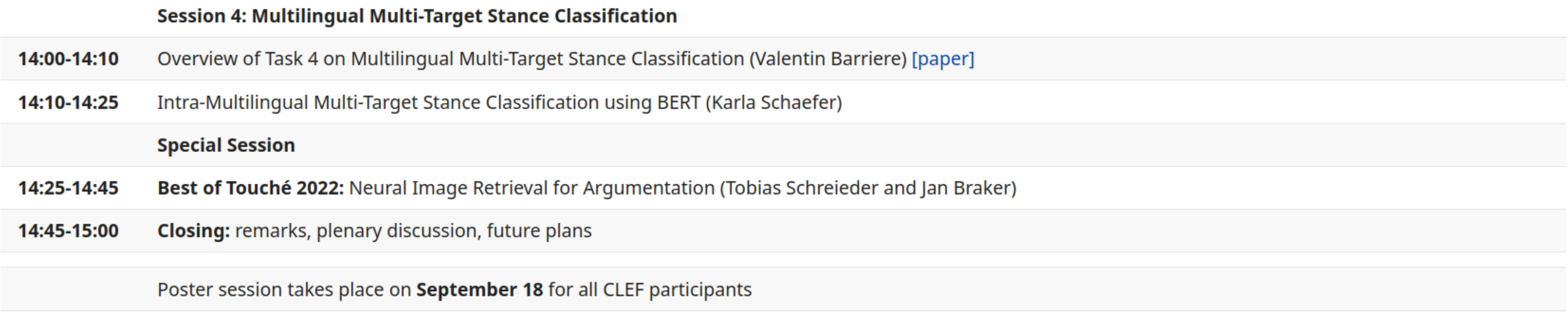


Thursday, September 21. Touché: Argument and Causal Retrieval Workshop	
11:30-11:35	**Welcome**
	Session 1: Argument Retrieval for Controversial Questions
11:35-11:45	Overview of Task 1 on Argument Retrieval for Controversial Questions (Alexander Bondarenko) [paper]
11:45-12:00	Argument Quality Prediction for Ranking Documents (Moritz Plenz) [paper]
	Session 2: Evidence Retrieval for Causal Questions
12:00-12:10	Overview of Task 2 on Evidence Retrieval for Causal Questions (Alexander Bondarenko) [paper]
12:10-12:20	Evidence Retrieval for Causal Questions Using Query Expansion and Reranking
	Session 3: Image Retrieval for Arguments
12:20-12:30	Overview of Task 3 on Image Retrieval for Arguments (Johannes Kiesel) [paper]
12:30-12:45	Matching Images and Keywords with CLIP (Fatihah Ulya Hakiem)
12:45-13:00	Comparing Image Generation, Stance Detection and Feature Matching for Image Retrieval for Arguments (Sarah Bachinger, Maximilian Enderling, and Max Möbius
13:00-14:00	**Lunch**

Touché: Argument and Causal Retrieval

Workshop Program

[touche.webis.de]

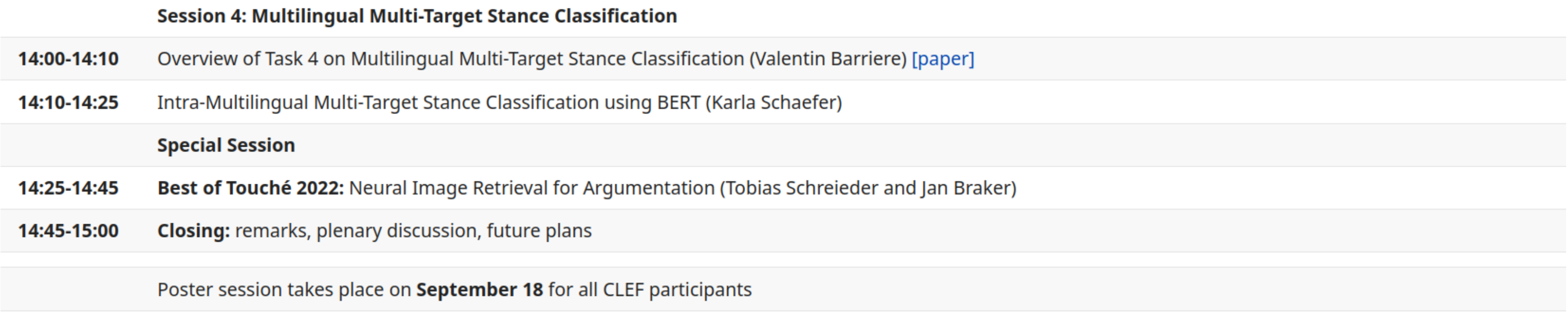


	Session 4: Multilingual Multi-Target Stance Classification
14:00-14:10	Overview of Task 4 on Multilingual Multi-Target Stance Classification (Valentin Barriere) [paper]
14:10-14:25	Intra-Multilingual Multi-Target Stance Classification using BERT (Karla Schaefer)
	Special Session
14:25-14:45	**Best of Touché 2022:** Neural Image Retrieval for Argumentation (Tobias Schreieder and Jan Braker)
14:45-15:00	**Closing:** remarks, plenary discussion, future plans
	Poster session takes place on **September 18** for all CLEF participants

Touché: Argument and Causal Retrieval

Workshop Program

[touche.webis.de]



	Session 4: Multilingual Multi-Target Stance Classification
14:00-14:10	Overview of Task 4 on Multilingual Multi-Target Stance Classification (Valentin Barriere) [paper]
14:10-14:25	Intra-Multilingual Multi-Target Stance Classification using BERT (Karla Schaefer)
	Special Session
14:25-14:45	**Best of Touché 2022:** Neural Image Retrieval for Argumentation (Tobias Schreieder and Jan Braker)
14:45-15:00	**Closing:** remarks, plenary discussion, future plans
	Poster session takes place on **September 18** for all CLEF participants

Spoiler: Touché will run again at CLEF 2024 (but with new tasks)

Submit your extended working notes to ECIR 2024

Session 1: Argument Retrieval for Controversial Questions

Moderator: Alexander Bondarenko

Argument and Argumentation

Argument:

- A conclusion (claim) supported by premises (reasons) [Walton et al. 2008]
- Conveys a stance on a controversial topic [Freeley and Steinberg, 2009]

Conclusion *Argumentation will be a key element of conversational agents.*

Premise 1 *Superficial conversation (“gossip”) is not enough.*

Premise 2 *Users want to know the “Why” to make informed decisions.*

Argumentation:

- Usage of arguments to achieve persuasion, agreement, . . .
- Decision making and opinion formation processes

Touché: Argument and Causal Retrieval

Task

Task 1: Argument Retrieval for Controversial Questions

- ❑ Scenario: Users search for arguments on controversial topics
- ❑ Task: Retrieve and rank relevant and high-quality arg. documents identify the document stance
- ❑ Data: ClueWeb22-B (200 million documents); also available via [ChatNoir]

- ❑ Run submissions similar to “classical” TREC tracks
- ❑ Software submissions in TIRA [tira.io]

Touché: Argument and Causal Retrieval

Topics

Example topic for Task 1:

Title	*Should teachers get tenure?*
Description	*A user has heard that some countries do give teachers tenure and others don't. Interested in the reasoning for or against tenure, the user searches for positive and negative arguments [...]*
Narrative	*Highly relevant arguments make a clear statement about tenure for teachers in schools or universities. Relevant arguments consider tenure more generally, not specifically for teachers, or, instead of talking about tenure, consider the situation of teachers' financial independence.*

Touché: Argument and Causal Retrieval

Evaluation

Document relevance (nDCG@10):

🙌	Highly relevant to the topic
👍	(Partially) relevant to the topic
👎	Everything else

Rhetorical argument quality (nDCG@10):

🙌	Proper language, good structure, good grammar, easy to follow
👍	Proper language but broken logic / hard to follow, or vice versa
👎	Profanity, hard to follow, grammar issues / no arguments at all

Document stance (macro-avg. F1):

🤺	Pro, con, neutral, no stance

Touché: Argument and Causal Retrieval

Session 1: Participant Paper Presentation

- 1 team (Renji Abarai) submitted 7 runs
- Baseline (Puss in Boots): BM25F-based ChatNoir; Flan-T5 for stance
- 747 documents manually judged (relevance, argument quality, and stance)

Session 2: Evidence Retrieval for Causal Questions

Moderator: Alexander Bondarenko (on behalf of Ferdinand Schlatt)

Touché: Argument and Causal Retrieval

Causality

Cause–Effect relationships:

- An integral part of human reasoning; an association of two ideas because of experiencing their regular conjunction [Khoo, 2002]
- A cause is an insufficient but necessary part of unnecessary but sufficient conditions for an effect (INUS) [Mackie, 1980]

Fuel-soaked Rag	➜ House Fire
Sufficient condition	{Fuel-soaked rag, spark, wooden house, . . . }
Unnecessary condition	Other possible conditions exist
Necessary part	Without the rag, no fire would happen
Insufficient part	Only the rag would not cause the fire

Touché: Argument and Causal Retrieval

Task

Task 2: Retrieving and analyzing evidence for causal claims

- ❑ Scenario: Users want to know if two events are causally related
- ❑ Goal: Help to find evidence for or against a causal claim
- ❑ Task: Retrieve and rank documents containing evidence identify the document stance
- ❑ Data: ClueWeb22-B (200 million documents); also available via [ChatNoir]

- ❑ Run submissions similar to “classical” TREC tracks
- ❑ Software submissions in TIRA [tira.io]

Touché: Argument and Causal Retrieval

Topics

Example topic for Task 2:

Title	*Could sun exposure cause hair loss?*
Cause	sun exposure
Effect	hair loss
Description	*A user is wondering how to protect against hair loss and specifically, if an increased exposure to sunlight can cause hair loss.*
Narrative	*Highly relevant documents will provide information on a potential causal connection between exposure to sunlight and hair loss (medically: alopecia). This includes documents stating or giving evidence that the first is (or is not) a cause of the other. Documents stating that there is not enough evidence to decide either way are also highly relevant. [. . .]*

Touché: Argument and Causal Retrieval

Evaluation

TOUCHÉ 2023

Document relevance (nDCG@5):

🙌	Highly relevant to the topic
👍	(Partially) relevant to the topic
👎	Everything else

Document stance (macro-avg. F1):

✅	Supporting Evidence
❌	Refuting Evidence
❓	Neutral Evidence

©touche.webis.de 2023

Touché: Argument and Causal Retrieval

Session 2: Participant Paper Presentation

- 1 team (He-Man) submitted 3 runs
- Baseline (Puss in Boots): BM25F-based ChatNoir; Flan-T5 for stance
- 718 documents manually judged (relevance and stance)

Touché: Argument and Causal Retrieval

TOUCHÉ 2023

Evidence Retrieval for Causal Questions Using Query Expansion and Reranking

Aron Gaden, Niklas Rausch, Bruno Reinhold, and Lukas Zeit-Altpeter

Friedrich-Schiller-Universität Jena

Touché: Argument and Causal Retrieval

Approach

- Query: Could *sun exposure* cause *hair loss*?
- First-stage retrieval with ChatNoir: (1) original and (2) expanded query
- Dependency tree parsing to extract cause, effect, and causal phrase
- Query expansion with synonyms from CauseNet
- Query expansion with terms generated by ChatGPT

Touché: Argument and Causal Retrieval

Approach

- ❑ Query: Could *sun exposure* cause *hair loss*?
- ❑ First-stage retrieval with ChatNoir: (1) original and (2) expanded query
- ❑ Dependency tree parsing to extract cause, effect, and causal phrase
- ❑ Query expansion with synonyms from CauseNet
- ❑ Query expansion with terms generated by ChatGPT

- ❑ Re-ranking using a position bias
- ❑ Dependency tree parsing: cause, effect, and causal phrase (in documents)
- ❑ Documents containing the causal relationship from the original query earlier in the document are ranked higher

Touché: Argument and Causal Retrieval

Results

Team	Run Tag	nDCG@5 Relevance	F1 macro Stance
He-Man	no_expansion_rerank	0.657†	–
Puss in Boots	**ChatNoir**	**0.585**	**0.256**
He-Man	gpt_expansion_rerank	0.374	–
He-Man	causenet_expansion_rerank	0.268	–

Touché: Argument and Causal Retrieval

Results

Team	Run Tag	nDCG@5 Relevance	F1 macro Stance
He-Man	no_expansion_rerank	0.657[†]	–
Puss in Boots	**ChatNoir**	**0.585**	**0.256**
He-Man	gpt_expansion_rerank	0.374	–
He-Man	causenet_expansion_rerank	0.268	–

- Simple yet effective approach
- A high-precision but low-recall solution
- Error: (drinking wine, blood urine) → (eating food, diarrhea)
- Room for future research

©touche.webis.de 2023

Touché: Argument and Causal Retrieval

Session 3: Image Retrieval for Arguments

Moderator: Johannes Kiesel

©touche.webis.de 2023

Touché: Argument and Causal Retrieval

Shared Task

Task 3: Image retrieval for arguments

- Scenario: Users search for images to corroborate their argumentation
- Task: Retrieve and rank images to support or attack a given stance
- Data: 56 000 web images with respective web documents and Google Cloud Vision data

- Run submissions similar to “classical” TREC tracks
- Software submissions in TIRA [tira.io]

Touché: Argument and Causal Retrieval

Statistics

- Submissions: 3 participating teams (+ baseline)

- Approaches: 12 valid runs were evaluated (+ baseline)
- Baseline: Re-implementation of Aramis approach
- Evaluation: 7 000 images-topic pairs judged manually

- Matthew Lewis as Neville Longbottom in “Harry Potter”
- George Takei as Hikaru Sulu in “Star Trek”
- Patrick Stewart as Jean-Luc Picard in “Star Trek”
- Minsc (and Boo) by u/Kazuliski (on Reddit)

Touché: Argument and Causal Retrieval

Results

TOUCHÉ 2023

		Precision@10		
Team	**Run**	**On-topic**	**Arg.**	**Stance**
Neville Longbottom	clip_chatgpt_args.raw	0.785	0.338	0.222
Hikaru Sulu	Keywords	0.664	0.350	0.185
Jean-Luc Picard	No stance detection	0.523	0.292	0.162
Minsc	Baseline (Aramis)	0.376	0.194	0.102
Boromir	On 2022 data	0.878	0.768	0.425

©touche.webis.de 2023

Touché: Argument and Causal Retrieval

Results

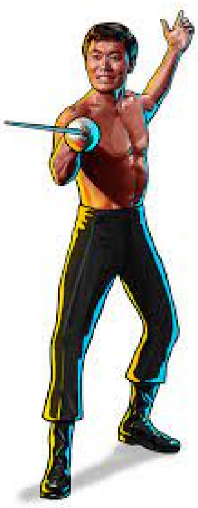
		Precision@10		
Team	**Run**	**On-topic**	**Arg.**	**Stance**
Neville Longbottom	clip_chatgpt_args.raw	0.785	0.338	0.222
Hikaru Sulu	Keywords	0.664	0.350	0.185
Jean-Luc Picard	No stance detection	0.523	0.292	0.162
Minsc	Baseline (Aramis)	0.376	0.194	0.102
Boromir	On 2022 data	0.878	0.768	0.425

Neville Longbottom

- ❑ ChatGPT for generating arguments for topic + stance
- ❑ ChatGPT for generating image descriptions for arguments
- ❑ CLIP for ranking images by similarity to descriptions
- ❑ Experimented with re-ranking using description for other stance or IBM’s debater pro-con score

Touché: Argument and Causal Retrieval

Session 3: Participants' paper presentations



TOUCHÉ
2023

Session 4: Multilingual and Multi-target Stance Classification

Moderator: Valentin Barriere

Task

Task 4: Multilingual and Multi-target Stance Classification

- Scenario: Stakeholders want to get an overview about citizens' opinions on an important societal topic
- Task: Detect the stance of a comment towards a proposal
- Data: 4 200 proposals and 20 000 comments focused on various topics written in 26 different languages

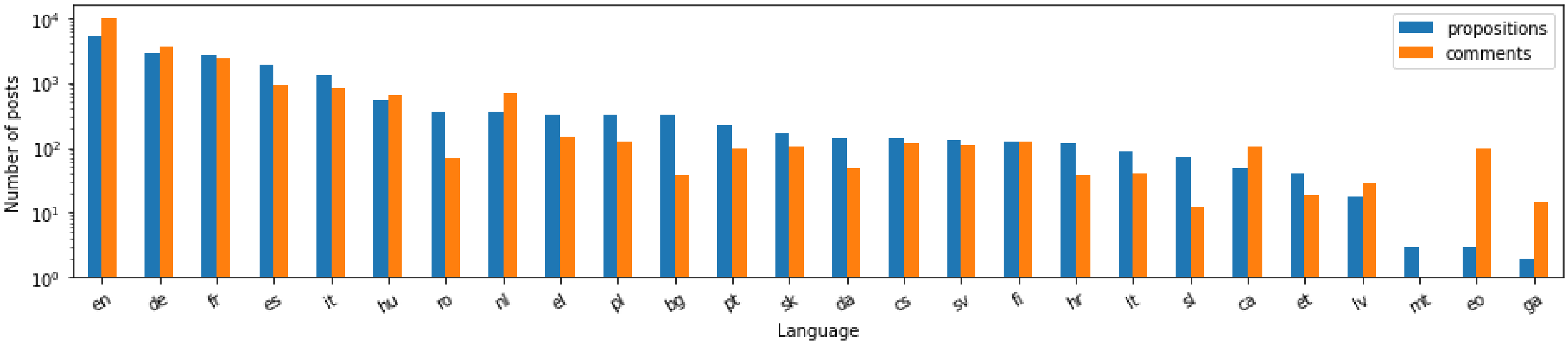
Touché: Argument and Causal Retrieval

Task

TOUCHÉ 2023

Task 4: Multilingual and Multi-target Stance Classification

- Scenario: Stakeholders want to get an overview about citizens' opinions on an important societal topic
- Task: Detect the stance of a comment towards a proposal
- Data: 4 200 proposals and 20 000 comments focused on various topics written in 26 different languages



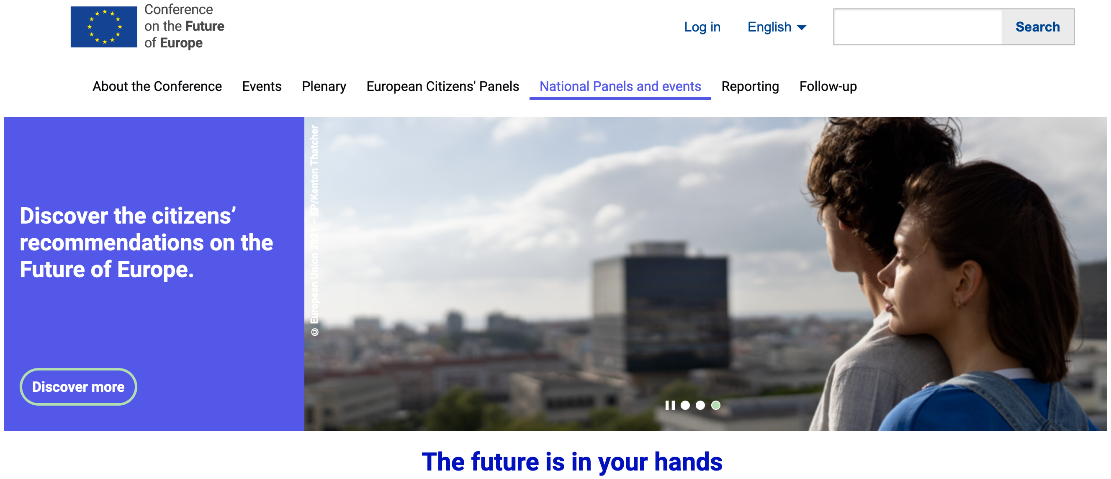
Touché: Argument and Causal Retrieval

Data

TOUCHÉ 2023

Example data instance for Task 4:

Title	Topic	Proposal	Comment	Stance
Focus on anti-aging and longevity research	Health	The EU has presented their green paper on aging, and correctly named the aging . . .	The idea of prevention being better than a cure is nothing new or revolutionary. Rejuvenation . . .	In favor
Encourage people eat less meat	Climate change	I think it would be great that everyone gets a meat card. You take the card to the store . . .	La valeur nutritionnelle de la viande reste un argument très fort en faveur de la consommation . . .	Against



Discover the citizens' recommendations on the Future of Europe.

The future is in your hands

Touché: Argument and Causal Retrieval

Evaluation

- Subtask 1: Cross-debate classification
- Subtask 2: All-data-available classification
- Baselines: a) always predict the majority class ‘in favor’ (Cavalier Simple)
  b) multilingual masked language model XLM-R (Cavalier)
- Participants: 2 teams, 8 runs

Barriere, Valentin, and Alexandra Balahur. “Multilingual Multi-Target Stance Recognition in Online Public Consultations.” Mathematics 11, no. 9 (2023): 2161.

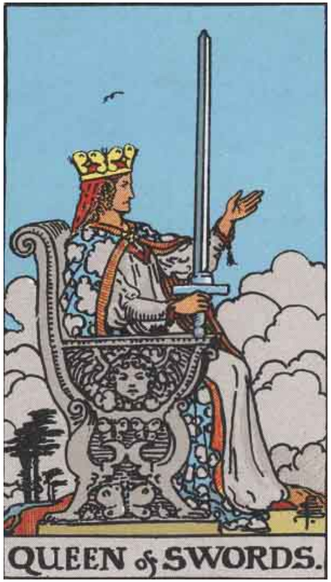
Touché: Argument and Causal Retrieval

Evaluation

- Subtask 1: Cross-debate classification
- Subtask 2: All-data-available classification
- Baselines: a) always predict the majority class ‘in favor’ (Cavalier Simple)
  b) multilingual masked language model XLM-R (Cavalier)
- Participants: 2 teams, 8 runs

	F1 macro							**Acc.**
Team	**en**	**fr**	**de**	**it**	**hu**	**el**	**All**	
Subtask 1: Cross-debate classification								
Cavalier	**59.4**	**54.9**	**54.6**	**54.9**	**52.8**	**54.2**	**57.7**	**63.0**
Queen of Swords	44.8	41.3	34.5	37.7	40.5	38.9	41.7	60.5
Cavalier Simple	**24.4**	**24.2**	**20.3**	**25.1**	**29.3**	**17.1**	**23.7**	**55.2**
Subtask 2: All-data-available classification								
Cavalier	**57.2**	**54.6**	**58.8**	**68.5**	**50.9**	**56.6**	**59.3**	**67.3**
Silver Surfer	36.7	33.9	30.2	37.8	38.0	33.3	35.0	55.1
. . .								
Queen of Swords	35.1	31.5	26.2	40.9	43.0	35.7	32.4	61.6
. . .								

Session 4: Participant’s paper presentation



Touché: Argument and Causal Retrieval

Special Session

Moderator: Léo Hemamou

- ❑ **Best of Touché 2022**: Neural Image Retrieval for Argumentation (Tobias Schreieder and Jan Braker)

Touché: Argument and Causal Retrieval

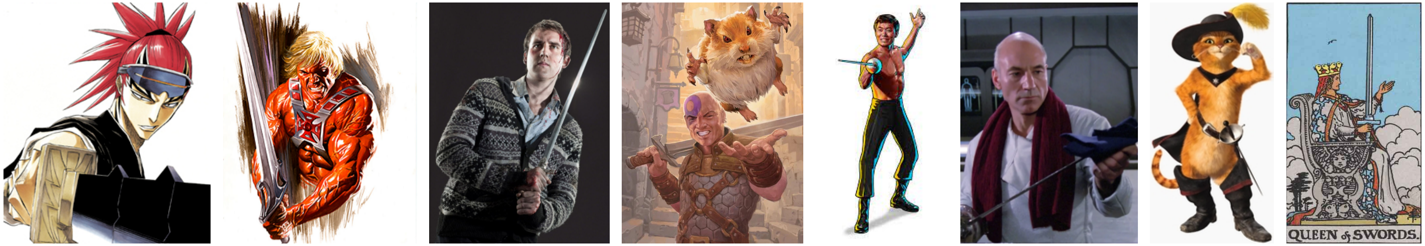
Panel discussion, closing remarks, future plans

Moderators: Alexander Bondarenko and Johannes Kiesel

Touché: Argument and Causal Retrieval

Statistics over 4 Years

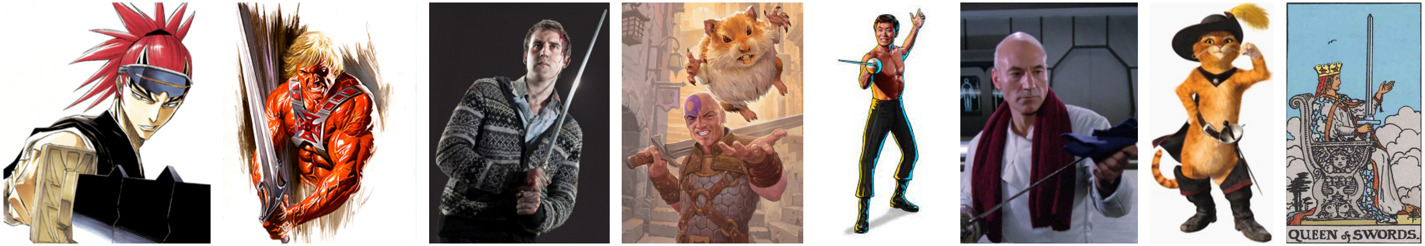
❑ Registrations:	163 teams (avg. 41 per year)
❑ Submissions:	74 participating teams (avg. 19 per year)
❑ Approaches:	243 valid runs were evaluated (avg. 61 per year)
❑ Evaluation:	> 30,000 manual judgments



Touché: Argument and Causal Retrieval

Statistics over 4 Years

- Registrations: 163 teams (avg. 41 per year)
- Submissions: 74 participating teams (avg. 19 per year)
- Approaches: 243 valid runs were evaluated (avg. 61 per year)
- Evaluation: > 30,000 manual judgments
- Tasks: Argument Retrieval for Controversial Questions
  Argument Retrieval for Comparative Questions
  Image Retrieval for Arguments
  Evidence Retrieval for Causal Questions
  Multilingual Multi-Target Stance Classification

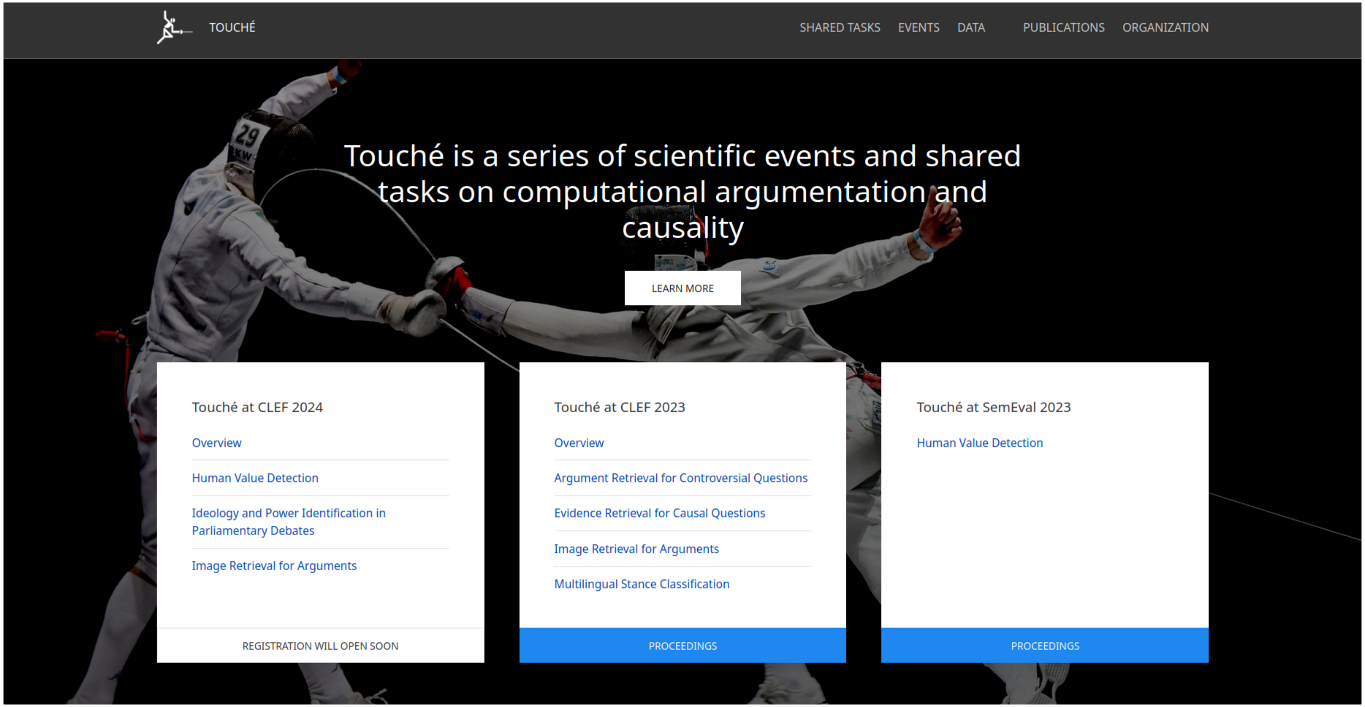


Touché: Argument and Causal Retrieval

Summary

TOUCHÉ 2023

- ❑ Platform for argument and causal retrieval and analysis [touche.webis.de]
- ❑ Relevance / quality / stance corpora and runs
- ❑ Tools for submission and evaluation [tira.io]



Touché is a series of scientific events and shared tasks on computational argumentation and causality

Touché at CLEF 2024

- Overview
- Human Value Detection
- Ideology and Power Identification in Parliamentary Debates
- Image Retrieval for Arguments

Touché at CLEF 2023

- Overview
- Argument Retrieval for Controversial Questions
- Evidence Retrieval for Causal Questions
- Image Retrieval for Arguments
- Multilingual Stance Classification

Touché at SemEval 2023

- Human Value Detection

Touché 2024: Argumentation Systems

TOUCHÉ
2024

Task 1: Human Value Detection (ValueEval)

Johannes Kiesel, Milad Alshomary, Nailia Mirzakhmedova, Nicolas Handke, Nicolas Stefanovitch,
Bertrand De Longueville Mario Scharfbillig, Henning Wachsmuth, Benno Stein

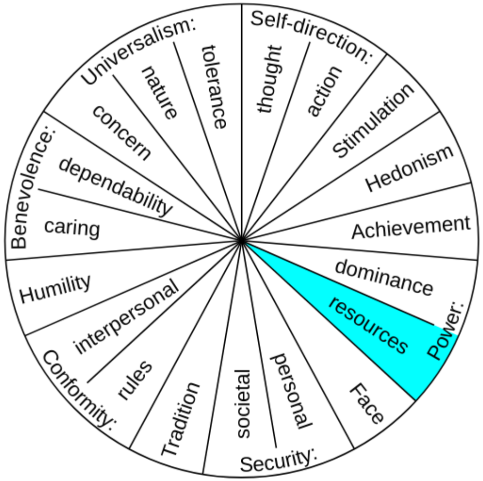
- ❑ Scenario: Users want to find different views (expressed by values) in texts
- ❑ Task: Given a text, detect for each sentence
  Subtask 1: which human values it refers to and
  Subtask 2: whether it signals (partial) attainment or constraint of the value
- ❑ Data: > 3 000 news+manifestos, 8 languages, 400 to 800 words each

Example:

The budget for last year's government policies on defence went out of control.

Value (Subtask 1): Power: Resources

Attainment (Subtask 2): (Partially) constrained



Schwartz value system

Touché 2024: Argumentation Systems

Task 2: Multilingual Ideology and Power Identification in Parliamentary Debates

Çağrı Çöltekin, Nikola Ljubešić, Katja Meden, Tomaž Erjavec, Vaidas Morkevičius, Matyáš Kopp

- Scenario: To better understand how political ideology the position of the speaker affects parliamentary debates

- Task: Given a transcribed speech in some language, detect
  Subtask 1: the ideology of the speaker's party
  Subtask 2: whether the speaker belongs to a governing party (coalition)

- Data: Speech samples from multiple national/regional parliaments from the ParlaMint project, and their automatic translations to English

Dataset: https://www.clarin.eu/parlamint

©touche.webis.de 2023

Touché 2024: Argumentation Systems

Task 3: Image Retrieval/Generation for Arguments (joint task with ImageCLEF)

Maximilian Heinrich, Johannes Kiesel, Martin Potthast, Benno Stein

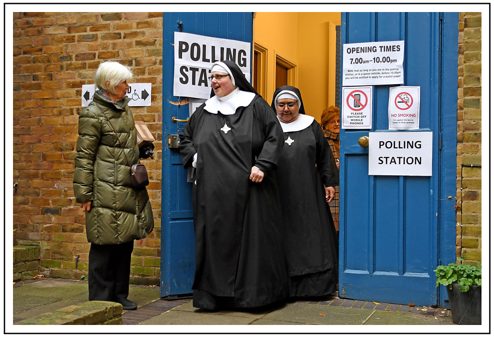
- ❑ Scenario: Users want to better convey arguments (with images)
- ❑ Task: Retrieve/generate images to reinforce an argument's premise
- ❑ Data: > 10 000 web images and Stable Diffusion API

Example:

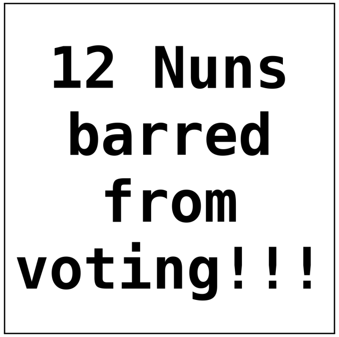
Claim: *Legislation to impose restrictive photo ID requirements has the potential to block millions of American voters*

Premise: *Indiana's photo ID law barred twelve retired nuns from voting*

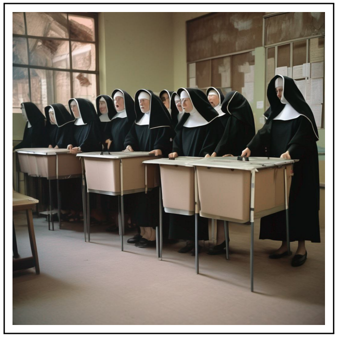
Submission:



retrieved from dataset


12 Nuns barred from voting!!!

generated "text-image"



generated via Stable Diffusion

Touché: Argument and Causal Retrieval

Open discussion

thank you!

© touche.webis.de 2023