Overview of Touché 2024: Argumentation Systems

Johannes Kiesel[1(✉)], Çağrı Çöltekin[2], Maximilian Heinrich[1], Maik Fröbe[3], Milad Alshomary[4], Bertrand De Longueville[5], Tomaž Erjavec[6], Nicolas Handke[7], Matyáš Kopp[8], Nikola Ljubešić[6], Katja Meden[6], Nailia Mirzhakhmedova[1], Vaidas Morkevičius[9], Theresa Reitis-Münstermann[10], Mario Scharfbillig[5], Nicolas Stefanovitch[5], Henning Wachsmuth[4], Martin Potthast[11,12,13], and Benno Stein[1]

[1] Bauhaus-Universität Weimar, Weimar, Germany
touche@webis.de
[2] University of Tübingen, Tübingen, Germany
[3] Friedrich-Schiller-Universität, Jena, Germany
[4] Leibniz University Hannover, Hanover, Germany
[5] European Commission, Joint Research Centre (JRC), Brussels, Belgium
[6] Jožef Stefan Institute, Ljubljana, Slovenia
[7] Leipzig University, Leipzig, Germany
[8] Charles University, Prague, Czech Republic
[9] Kaunas University of Technology, Kaunas, Lithuania
[10] Arcadia Sistemi Informativi Territoriali, Milano, Italy
[11] University of Kassel, Kassel, Germany
[12] hessian.AI, Darmstadt, Germany
[13] ScaDS.AI, Leipzig, Germany

Abstract. This paper is a condensed overview of Touché: the fifth edition of the lab on argumentation systems that was held at CLEF 2024. With the goal to foster the development of support-technologies for decision-making and opinion-forming, we organized three shared tasks: (1) Human value detection (ValueEval), where participants detect (implicit) references to human values and their attainment in text; (2) Multilingual Ideology and Power Identification in Parliamentary Debates, where participants identify from a speech the political leaning of the speaker's party and whether it was governing at the time of the speech (new task); and (3) Image retrieval or generation in order to convey the premise of an argument with visually. In this paper, we describe these tasks, their setup, and participating approaches in detail.

Keywords: Argumentation · Human values · Ideology · Image retrieval

L. Goeuriot et al. (Eds.): CLEF 2024, LNCS 14959, pp. 308–332, 2024.
https://doi.org/10.1007/978-3-031-71908-0_14

1 Introduction

Decision-making and opinion-forming are everyday tasks, for which everybody has the chance to acquire knowledge on the Web on almost every topic. However, conventional search engines are primarily optimized for returning *relevant* results, which is insufficient for collecting and weighing the pros and cons for a topic. To close this gap of technologies that support people in decision-making and opinion-forming, the Touché lab's shared tasks[1] (https://touche.webis.de) call for the research community to develop respective approaches. In 2024, we organized the three following shared tasks:

1. Human Value Detection (a continuation of ValueEval'23 @ SemEval [38]) features two subtasks in ethical argumentation of detecting human values in texts and their attainment, respectively.
2. Ideology and Power Identification in Parliamentary Debates features two subtasks in debate analysis of detecting the ideology and position of power of the speaker's party, respectively (new task).
3. Image Retrieval/Generation for Arguments (third edition, now joint task with ImageCLEF) is about the retrieval or generation of images to help convey an argument's premise.

In total, 20 teams participated in Touché in 2024. Nine teams participated in the human value detection task (cf. Sect. 4)—of which six submitted a notebook paper—and submitted 21 runs. Most teams integrated DeBERTa [32], RoBERTa [46], or the multi-lingual XLM-RoBERTa [12]. Only one team employed a generative approach (employing GPT-4o). Nine teams participated in the multilingual ideology and power identification task (cf. Sect. 5) and submitted 52 runs. The majority of teams participated in both subtasks. While traditional machine learning methods like support vector classifiers or logistic regression with n-gram features were more common among participating teams, higher-scores were typically obtained by teams using pretrained models. The two teams that participated in the image retrieval/generation task used similarity embeddings between images and text. One team used CLIP [58], the other a DPR [35] inspired approach. The corpora, topics, and judgments created at Touché are freely available to the research community on the lab's website.[2]

2 Related Work

Argumentation systems are diverse and are connected to many fields within and outside of computer science. The following sections review the related work for each Touché task of 2024.

[1] 'touché' confirms "a hit in fencing or the success or appropriateness of an argument, an accusation, or a witty point." [https://merriam-webster.com/dictionary/touche].
[2] https://touche.webis.de/.

2.1 Human Value Detection

Due to their outlined importance, human values have been studied both in the social sciences [66] and in formal argumentation [8] for decades. According to the former, a "value is a (1) belief (2) pertaining to desirable end states or modes of conduct, that (3) transcends specific situations, (4) guides selection or evaluation of behavior, people, and events, and (5) is ordered by importance relative to other values to form a system of value priorities." For cross-cultural analysis, Schwartz derived 48 value questions from universal individual and societal needs, including concepts such as *obeying all the laws* and *being humble* [67]. Based on these taxonomies are several studies in the social sciences, which could greatly benefit from the automated methods our task aims at [64]. See Scharfbillig et al. [65] for a recent overview and practical insights from the social sciences.

Moreover, several works in computer science utilize values. For example, in the context of interactive systems, to tune interactive chat-based agents or texts in general towards morally acceptable behavior [3,45]. A related dataset is ValueNet [57], which contains 21K one-sentence descriptions of social scenarios (taken from SOCIAL-CHEM-101 [23]) annotated for the 10 value categories of an earlier version of Schwartz' value taxonomy. A major difference to the Touché24-ValueEval dataset are the more ordinary situations in ValueNet (e.g., whether to say "I miss mom"). Our earlier work analyzed values in short arguments [37,38].

2.2 Ideology and Power Identification

Parliamentary data has a high societal impact and provides publicly available sources for analyzing (argumentative) language. Therefore, the number of resources based on parliamentary proceedings [22,42], and computational and linguistics analyses of parliamentary debates [1,28] increased in recent years.

The present task is about two important aspects of the political discourse, *ideology* and *power*. Although a simplification, political orientation on the left-to-right spectrum has been one of the defining properties of political ideology [5,74]. Power is another factor that shapes the political discourse [15,20,21]. Automatic identification of political orientation from texts has attracted considerable interest [10,13,27,55,56], including a few recent shared tasks [25,62]. The present task differs from the earlier ones, with respect to the source material (parliamentary debates, rather than the popular sources of social media or news) and multilinguality. Despite its central role in critical discourse analysis, to the best of our knowledge, power in parliamentary debates has not been studied computationally. There has been only a few recent computational studies providing indications of linguistic differences between governing and opposition parties [40,49,51,71]. The present shared task and associated data is likely to provide a reference for the future studies investigating power in political discourse.

2.3 Image Retrieval/Generation for Arguments

Images are a powerful tool for visual communication. They can provide contextual information and express, underline, or popularize an opinion [17], thereby

taking the form of subjective statements [18]. Some images express both a premise and a conclusion, making them full arguments [30,61]. Other images may provide contextual information only and have to be combined with a textual conclusion to form a complete argument. In this regard, a recent SemEval task distinguished a total of 22 persuasion techniques in memes alone [16]. Moreover, argument quality dimensions like acceptability, credibility, emotional appeal, and sufficiency [75] all apply to arguments that include images as well.

3 Lab Overview and Statistics

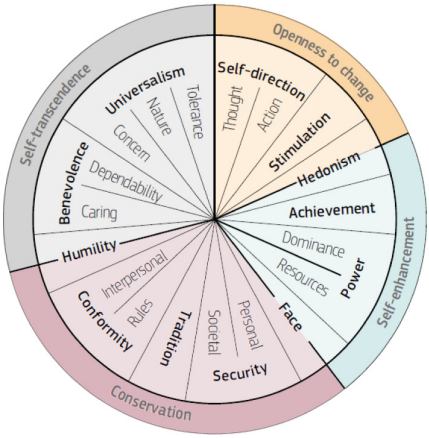
For the fifth edition of the Touché lab, we received 68 registrations from 22 countries (vs. 41 registrations in 2023). The most lab registrations came from India (24). Out of the 68 registered teams, 20 actively participated in this year's Touché edition (9, 9, and 2 teams submitting valid runs for Task 1, 2, and 3, respectively). Active teams in previous editions were: 7 in 2023, 23 in 2022, 27 in 2021, and 17 in 2020.

We used TIRA [24] as the submission platform for Touché 2024 through which participants could either submit code, software, or run files.[3] Code and software submissions increase reproducibility, as the software can later be executed on different data of the same format. To submit software, a team implemented their approach in a Docker image that they then uploaded to their dedicated Docker registry in TIRA. Software submissions in TIRA are immutable, and after the docker image had been submitted, the teams specified the to-be-executed command—the same Docker image can thus be used for multiple software submissions (e.g., by changing some parameters). A team could upload as many Docker images or software submissions as they liked; only they and TIRA had access to their dedicated Docker image registry (i.e., the images were not public while the shared task was ongoing). To improve reproducibility, TIRA executes software in a sandbox by removing the internet connection (ensuring that the software is fully installed in the Docker image which eases rerunning software later, as libraries and models must be installed in an image). For the execution, participants could select the resources that their software had available for execution, from 1 CPU core with 10 GB RAM up to 5 CPU cores with 50 GB RAM and 1 Nvidia A100 GPU with 40 GB RAM. Participants could run their software multiple times using different resources to study the scalability and reproducibility (e.g., whether the software executed on a GPU yields the same results as on a CPU). TIRA used a Kubernetes cluster with 1,620 CPU cores, 25.4 TB RAM, 24 GeForce GTX 1080 GPUs, and 4 A100 GPUs to schedule and execute the software submissions, to allocate the resources that the participants selected.

4 Task 1: Human Value Detection (ValueEval)

The goal of this task is to develop approaches that allow for the large-scale analysis of human values behind texts. In argumentation, one has to consider

[3] https://tira.io.



Inner circle: 19 human values
(see https://valueeval.webis.de)

Outer circle: four motivational directions
(not used in this task)

- **Openness to change**
  Being independent and exploring
- **Self-enhancement**
  Seeking pleasure, wealth, and esteem
- **Conservation**
  Preserving group cohesion, order, and security
- **Self-transcendence**
  Helping others, close ones, and nature

Fig. 1. The 19 values used in this task, shown in the Schwartz value taxonomy [67].

that people have different beliefs and priorities of what is generally worth striving for (e.g., personal achievements vs. humility) and how to do so (e.g., being self-directed vs. respecting traditions), referred to as (human) values. By analyzing corpora of texts, for example for news portals or political parties, one can develop an understanding of the values that the authors deem the most important.

4.1 Task Definition

The task is to identify the values of the widely accepted value taxonomy of Schwartz [67] (cf. Fig. 1) and their attainment in long texts of nine languages (Bulgarian, Dutch, English, French, German, Greek, Hebrew, Italian, and Turkish). This taxonomy has been replicated in over 200 samples in 80 countries and is the backbone of value research [65]. A value can either be mentioned as something that is or should be attained (i.e., lead towards fulfilling the value) or something that is constrained, i.e., not attained. For example, for Security, (partial) attainment would mean that something is made safer or healthier. In contrast, an event can be stated in a way that thwarts or constrains safety or health. Participating teams can submit software in one or both of two sub-tasks: (1) Given a text, for each sentence, detect which human values the sentence refers to; and (2) Given a text, for each sentence and value this sentence refers to, detect whether this reference (partially) attains or constrains the value.

4.2 Data Description

The task employs a collection of 2648 human-annotated texts in nine languages from news articles and political manifestos. Texts are sampled to reflect diverse opinions (different parties; mainstream news and others) from 2019 to 2023. The

Table 1. Overview of the Touché24-ValueEval dataset by language, with the respective number of texts, sentences, annotator agreement as measured by Krippendorf's α, and the thousandths of these sentences with any or a specific value (attained or constrained). Languages are Bulgarian (BE), German (DE), Greek (EL), English (EN), French (FR), Hebrew (HE), Italian (IT), Dutch (NL), and Turkish (TR).

				Sentences with value (‰)																		
Lang.	Texts	Sentences	α	Any value	Self-direction: thought	Self-direction: action	Stimulation	Hedonism	Achievement	Power: dominance	Power: resources	Face	Security: personal	Security: societal	Tradition	Conformity: rules	Conformity: interpersonal	Humility	Benevolence: caring	Benevolence: dependability	Universalism: concern	Universalism: nature	Universalism: tolerance
BG	260	6 919	.495	641	010	055	046	005	075	053	053	021	011	108	020	089	009	002	059	021	071	023	005
DE	261	9 183	.367	533	018	055	034	011	079	032	038	020	026	059	009	072	015	002	017	015	050	026	014
EL	328	7 349	.696	615	003	013	029	003	054	074	089	018	011	130	006	060	046	000	024	032	054	025	014
EN	408	10 305	.409	306	004	025	005	004	043	016	016	006	014	053	008	036	016	003	006	007	031	012	008
FR	219	4 650	.685	304	005	023	016	005	019	024	015	020	021	065	006	030	010	001	012	007	038	020	009
HE	250	7 331	.557	859	025	042	021	003	081	122	094	032	029	170	031	096	011	002	016	041	080	022	015
IT	276	6 379	.610	632	010	015	072	008	133	053	082	029	013	071	003	076	002	000	018	004	045	038	009
NL	323	10 982	.411	366	014	029	004	003	039	030	037	010	009	072	004	033	005	002	004	017	043	019	009
TR	323	11 133	.463	473	015	046	027	022	059	025	045	016	042	072	027	071	007	004	047	025	036	014	007
All	2 648	74 231	.546	512	012	035	026	008	063	045	050	018	020	086	013	061	013	002	022	019	048	021	010

data is annotated as part of the ValuesML project[4] by over 70 value scholars. The annotators marked segments in the texts, selected from 19 values the one that the segment refers to most, and selected whether the segment (partially) attains or constrains the value, or whether it is unclear if it attains or constrains it. Dedicated team leaders per language trained the respective annotators, consolidated annotations into a single ground truth, and discussed sentences were annotators disagreed (measured continuously by us) in their language teams. The team leaders discussed issues with us in bi-weekly meetings. Moreover, we discussed with the team leaders the current holistic inter-annotator agreement [70] and its change compared to the previous meeting to monitor annotation quality and coherence across documents and languages. To measure annotator agreement, we computed Krippendorf's α before curation for all language teams individually and overall (cf. Table 1). We see this agreement as sufficient, and belief that the curation process increased the annotation quality even further.

For Touché, the dataset is automatically split into sentences using Trankit version 1.1.1 [52] (cf. Table 2 for the sentence-based dataset format). The dataset is provided both in the original language and automatically translated to English,

[4] https://knowledge4policy.ec.europa.eu/projects-activities/valuesml-unravelling-expressed-values-media-informed-policy-making_en.

Table 2. Excerpt of the dataset for the human value detection task. The dataset comes in six directories: training, validation, and test data for both the original multilingual dataset and its automatic translation to English. Each directory contains a `sentences.tsv` where each row corresponds to one sentence. The training and validation directories also each contain a `labels.tsv` where each row corresponds to a sentence in `sentences.tsv` and columns 3–40 correspond to labels (attained and constrained for each of the 19 values). Label values in the `labels.tsv` are either 1.0 if the sentence refers to that value and attainment polarity, 0.0 if it does not, or 0.5 if the sentence refers to that value but the attainment polarity is unclear (0.2% of cases).

`sentences.tsv` (3 columns)

Text-ID	Sentence-ID	Text
EN_012	1	Who designed global guidelines for puberty blockers?
EN_012	2	More and more children and young people believe they have to question their gender ...
EN_012	3	Some 60 minors were treated in the Netherlands in 2010, but has increased to around ...

`labels.tsv` (40 columns)

Text-ID	Sentence-ID	Self-direction: thought attained	Self-direction: thought constrained	...
EN_012	1	0.0	0.0	...
EN_012	2	1.0	0.0	...
EN_012	3	0.0	0.0	...

either using DeepL or, for Hebrew, Google Translate.[5] The dataset is split into sets by texts, so that 60% of sentences are in the training set, 20% in the validation set, and 20% in the test set.[6]

Table 1 shows the size of the dataset for each language and the value distribution. The number of texts per language are between 219 (French) and 408 (English). The number of sentences per language are between 4 650 (French) and 11 133 (Turkish). Only 30.4% of the French sentences are annotated as referring to a value, but 85.9% of Hebrew sentences. The least frequent value overall is *Humility* (0.2%) and the most frequent one is *Security: societal* (8.6%). This in-balance between languages and values makes the multi-label classification problem especially challenging.

4.3 Participant Approaches

In 2024, nine teams participated in this task (of which six submitted a notebook paper) and submitted 21 runs. Moreover, we added two baseline runs for comparison. Five of the six teams that submitted a paper relied on DeBERTa [32], RoBERTa [46], or the multi-lingual XLM-RoBERTa [12]. The other team (Eric Fromm) used GPT-4o.[7] Two teams work with the multi-lingual dataset (Arthur Schopenhauer, Hierocles of Alexandria) whereas the others use the English trans-

[5] https://www.deepl.com/pro-api and https://cloud.google.com/translate.
[6] Dataset: https://zenodo.org/doi/10.5281/zenodo.10396293.
[7] https://openai.com/index/hello-gpt-4o/.

lations only. Only one team (Hierocles of Alexandria) used the sentence sequence, whereas the other teams classified each sentence individually.

Baselines. We provide two baselines, that also served to kickstart the participants' approaches:[8] (1) a random baseline that assigns a (uniformly) random value "confidence" to each value for each sentence in subtask 1 and randomly distributes this confidence between attained and constrained for subtask 2; and (2) a BERT [14] baseline with a multi-label classification head for all 38 combinations of value and attainment.

Team Arthur Schopenhauer [77].[9] The team used the multi-lingual dataset and analyzed the sentences independently. They approached subtask 1 as a classification problem. A *no-label* class was added for sentences without assigned value, and sentences with *Humility* were ignored due to the scarcity of that value. The 6% of sentences with more than one assigned value were ignored, as well. Different models were fine-tuned for English texts (deberta-v2-xxlarge [32]) and others (xlm-roberta-large [12]). In both cases, an ensemble with a thresholded soft voting scheme of four models was employed: one model for each combination of two seeds and two loss functions. For loss functions the authors report that cross entropy lead to higher results in their preliminary tests for frequent values but weighted cross entropy did so for infrequent values. The team approached subtask 2 as a binary classification problem, ignoring the few sentences with *unknown* attainment. Their approach is otherwise the same as for subtask 1, except that only a single model was employed instead of an ensemble (with cross entropy loss) based on results from their preliminary tests.

Team Edward Said [7]. The team used the English translations of the dataset and analyzed the sentences independently. To counter the label imbalance, the team upsampled sentences by a factor of four if the associated label is one of 14 underrepresented labels (value + attainment). They selected these 14 labels out of the 38 labels if the label was infrequent in total or in comparison to the other label for the same value (but different attainment). They then fine-tuned a RoBERTa [46] and DeBERTa [32] model for multi-label classification.

Team Eric Fromm [50]. The team used the English translations of the dataset and analyzed the sentences independently. They employed GPT-4o for zero-shot classification, prompting it with the 19 value descriptions from the annotator's guide to select one or none for each sentence. They did not tackle subtask 2.

Team Hierocles of Alexandria [41].[10] The team used both the multi-lingual dataset and English translations and incorporated sentence sequence information. More specifically, their approach predicts values for a sentence from an

[8] https://github.com/touche-webis-de/touche-code/tree/main/clef24/human-value-detection/approaches.

[9] Code: https://github.com/h-uns/clef2024-human-value-detection.

[10] Code: https://github.com/SotirisLegkas/Touche-ValueEval24-Hierocles-of-Alexandria.

input text that consists of the previous two sentences concatenated with the target sentence. The two preceding sentences contained special tokens to represent any values assigned to them. During training and validation the true labels were employed, but during testing the predicted labels of the previous sentences were leveraged. The team fine-tuned different RoBERTa [46] and DeBERTa [32] models for English and XLM-RoBERTa [12] models for the multi-lingual dataset, with the best performing one being XLM-RoBERTa-xl [29]. Moreover, they developed a custom model architecture for multi-label text classification consisting of multiple classification heads. Each classification head focused on a different language for the multi-lingual dataset. The custom model architecture was adapted and employed for the English-translated dataset as well. After preliminary experiments concerning loss functions, class weights and various thresholds, they used the binary cross-entropy loss with logits as their loss function and selected an optimal classification threshold for each value. The approach is trained to tackle both subtasks 1 and 2.

Team Philo of Alexandria [76].[11] The team used the English translations of the dataset and analyzed the sentences independently. They approached subtask 1 as a multi-label problem and fine-tuned DeBERTa (deberta-base [32]) after initial experiments with several models. They employ the same base model for subtask 2 and fine-tune it to classify each text pair of sentence and human value name into either attaining or constraining.

Team SCaLAR NITK (code name: Peter Abelard) [34]. The team used the English translations of the dataset and analyzed the sentences independently. They experimented with SVMs, KNNs, decision trees, hierarchical classification, transformer models and large language models. Based on preliminary experiments, they fine-tuned a RoBERTa [46] model for both subtasks (multi-label and binary classification, respectively).

4.4 Task Evaluation

Following ValueEval'23 [38], submissions are evaluated using standard macro F_1-score over all values. The same metric is used for the new subtask 2. The submission format has been designed so that participants submit only one run file for both subtasks (same format as the `labels.tsv`), but the scores for the subtasks are calculated independently of each other from the same file as follows. Each submission includes for each sentence and value a confidence score (between 0 and 1) for both attained and constrained polarity. If the sum of the two numbers is above 0.5, the submission is evaluated as having predicted that the sentence refers to that value (subtask 1). For subtask 2, only the sentence-value pairs are considered for which the sentence refers to the value according

[11] Code: https://github.com/VictorMYeste/touche-human-value-detection
Models: https://huggingface.co/VictorYeste/deberta-based-human-value-detection
https://huggingface.co/VictorYeste/deberta-based-human-value-stance-detection
Image: `docker pull victoryeste/valueeval24-philo-of-alexandria-deberta-cascading`.

to the ground-truth. For these pairs, the submission is evaluated as having predicted the attainment polarity for which it produced the larger confidence score.

Table 3 shows the results for the best-performing approaches per team for both subtasks. The best-performing approach for subtask 1 is the one of team Hierocles of Alexandria that uses XLM-RoBERTa-xl, the previous sentences, and is trained specifically for subtask 1. Overall, multilingual models performed best, with also the second-in-place employing such a model. Rarer values are overall detected worse, with the exception of the zero-shot approach by team Eric Fromm (especially Humility), indicating insufficient training data. Several teams achieved top scores for subtask 2. Overall, this binary classification task is, as once can expect, much easier than subtask 1. However, most teams clearly focused their efforts on subtask 1, so there is likely more room for improvement.

5 Task 2: Multilingual Ideology and Power Identification in Parliamentary Debates

The study of parliamentary debates is crucial to understand the decision processes in the parliaments and their societal impacts. The goal of this task is to automatically identify two important aspects of parliamentary debates: the political orientation of the party of the speaker, and the role of the party of the speaker in the governance of the country or the region. Identifying these underlying aspects of parliamentary debates enables automated comprehension of these discussions, the decisions that these discussions lead to, and their consequences.

5.1 Task Definition

Both subtasks were defined as binary classification tasks: Given a parliamentary speech, (1) predict the political orientation of the party of the speaker on the *left–right* spectrum, and (2) predict whether the speaker belongs to one of the governing parties or the opposition. The first task is relatively well studied, and there have been some recent shared tasks on identifying political orientation [25,62]. Unlike the earlier tasks, our data set includes multiple parliaments and languages, and is based on parliamentary debates. To the best of our knowledge, automatic identification of governing role—power—has not been studied earlier.

5.2 Data Description

The source of the data for this task is the ParlaMint [19], a uniformly encoded and annotated corpus of transcripts of parliamentary speeches from multiple national and regional parliaments.[12] The transcripts are The ParlaMint version 4.0 used for the task includes data from the following national and regional

[12] Although all transcripts are obtained thorough the data published by the respective parliaments, the method for obtaining the transcripts vary, such as scraping the web site of the parliament, extracting from published PDF files, and obtaining through an API provided by the parliament. For details, we refer to [19].

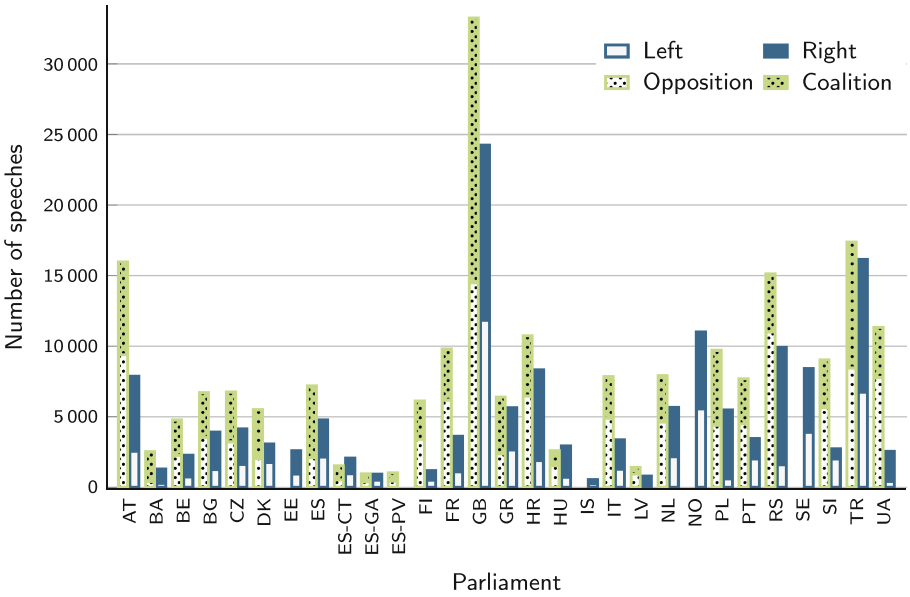


Fig. 2. Overview of the Touché24 ideology and power identification dataset. The bars show the training set for both subtasks for each parliament. Test set sizes are approximately 2 000 speeches for all parliaments.

parliaments: Austria (AT), Bosnia and Herzegovina (BA), Belgium (BE), Bulgaria (BG), Czechia (CZ), Denmark (DK), Estonia (EE), Spain (ES), Catalonia (ES-CT), Galicia (ES-GA), Basque Country (ES-PV), Finland (FI), France (FR), Great Britain (GB), Greece (GR), Croatia (HR), Hungary (HU), Iceland (IS), Italy (IT), Latvia (LV), The Netherlands (NL), Norway (NO), Poland (PL), Portugal (PT), Serbia (RS), Sweden (SE), Slovenia (SI), Turkey (TR) and Ukraine (UA). The labels for both subtasks are also coded in the ParlaMint corpora. For the sake of simplicity, we formulate both tasks as binary classification tasks. For both tasks, the main challenge in the creation of a dataset is to minimize the effects of covariates. Even though the instances to classify are speeches, the annotations are based on the party membership of the speaker. As a result, underlying variables like party membership, or speaker identity perfectly covary with ideology and power in most cases.

As a trade-off between data size, and for reducing the effect of covariates, we opt for a speaker-based sampling. First, to discourage, to some extent, the classifiers from relying on author identification, we sample at most 20 speeches of a single speaker. This is also important for introducing variation into the dataset, as the number of speeches from each speaker follows a power-law distribution: While a small number of speakers tend to deliver most of the speeches, e.g., party or party group leaders, most speakers have relatively few speeches. The distribution of speeches or speakers to include in training and test sets is also

Table 3. Achieved F_1-score of the best submission per team (as measured by overall F_1-score) on the test dataset for subtasks 1 and 2, and whether the submission used the original multilingual dataset or the automatic translation to English (EN). Baseline submissions ("Aristotle") are shown in gray.

Subtask 1

		F_1-score																			
Team	**Lang.**	**Overall**	**Self-direction: thought**	**Self-direction: action**	**Stimulation**	**Hedonism**	**Achievement**	**Power: dominance**	**Power: resources**	**Face**	**Security: personal**	**Security: societal**	**Tradition**	**Conformity: rules**	**Conformity: interpersonal**	**Humility**	**Benevolence: caring**	**Benevolence: dependability**	**Universalism: concern**	**Universalism: nature**	**Universalism: tolerance**
Hierocles of Alexandria [41]	multil.	39	15	27	30	37	45	42	49	31	42	49	46	51	24	00	34	33	47	63	27
Arthur Schopenhauer [77]	multil.	35	12	24	33	35	40	37	47	24	38	46	49	50	19	00	32	31	46	60	27
Philo of Alexandria [76]	EN	28	08	22	27	31	35	31	34	17	33	40	47	42	09	00	21	28	40	57	21
SCaLAR NITK [34]	EN	28	05	17	27	27	38	34	38	15	34	40	41	43	07	00	23	26	37	56	16
Edward Said [7]	EN	28	05	17	11	15	25	31	34	16	32	41	45	44	06	05	10	23	41	57	27
Erich Fromm [50]	EN	25	15	10	10	18	25	18	09	24	21	30	46	33	09	15	26	15	41	55	20
Lawrence Kohlberg	EN	25	08	11	19	23	31	22	31	11	28	37	34	42	09	00	21	23	34	54	18
Aristotle (BERT)	EN	24	00	13	24	16	32	27	35	08	24	40	46	42	00	00	18	22	37	55	02
John Shelby Spong	EN	07	00	00	02	00	16	05	11	00	01	28	00	15	00	00	00	00	13	27	00
Alain Badiou	EN	07	00	00	02	00	16	05	11	00	01	28	00	15	00	00	00	00	13	27	00
Aristotle (random)	EN	06	02	07	05	02	11	08	10	03	04	14	03	11	03	00	05	04	09	04	02

Subtask 2

Arthur Schopenhauer [77]	multil.	83	77	83	85	88	87	73	84	80	82	84	78	80	79	74	91	89	86	85	81
Edward Said [7]	EN	83	77	82	85	88	88	79	80	77	84	84	85	80	80	76	90	86	85	85	78
Philo of Alexandria [76]	EN	82	85	80	85	91	86	79	80	78	85	80	82	77	78	77	93	89	84	83	79
Aristotle (BERT)	EN	81	83	79	86	88	84	77	80	74	84	81	78	78	79	87	89	86	85	81	78
John Shelby Spong	EN	81	81	77	83	88	88	77	79	76	83	82	85	76	81	84	90	85	81	81	79
Alain Badiou	EN	81	81	77	83	88	88	77	79	76	83	82	85	76	81	84	90	85	81	81	79
Hierocles of Alexandria [41]	multil.	77	73	73	77	75	78	77	79	71	78	79	77	78	74	25	74	77	78	84	71
SCaLAR NITK [34]	EN	77	69	72	78	73	79	77	79	71	78	81	79	77	70	70	77	76	79	80	71
Erich Fromm [50]	EN	70	71	69	73	70	72	74	73	67	60	66	76	70	68	73	75	71	70	73	67
Lawrence Kohlberg	EN	66	81	77	83	80	70	76	63	56	33	45	85	63	46	84	90	79	69	70	60
Aristotle (random)	EN	52	51	47	54	52	53	55	53	52	52	50	54	53	49	45	53	56	52	49	56

important for proper evaluation. For the ideology task, the set of speakers in the training and test sets are disjoint. The ideal dataset split for the power identification task requires a different constraint: training and test sets should include speeches from the same speaker with different power roles. To come as close as possible to this ideal split, we opt for a best-effort training–test split.

When possible, we make sure that the speakers in the test set are also available in the training set with the opposite power role. Otherwise, we randomly sample more speakers to obtain the test set.

For evaluation, we set the test set size to 2 000 instances for both subtasks (100 to 200 speakers depending on the individual corpus and the task). Despite multiple speeches from each speaker, due to missing annotations and the lack of diversity of orientation in some parliaments, the disjoint speakers constraint mentioned above results in a small number of instances in the training set for some of the parliaments. Not all parliamentary data provides both labels. Some countries do not have the opposition–governing party distinction, and for the Galician parliament, the number and distribution of orientation labels did not result in a test set that was large enough. Figure 2 shows the training set sizes for each parliament. The test set size for all parliaments is approximately 2000 speeches. We do not provide a validation set. We provide further details on the data set and the sampling procedure in a separate publication [11].[13]

In addition ot the original speech transcripts and labels, we also provide automatic English translations, an anonymized speaker ID and the speaker's sex in the data for both tasks. Except the speaker ID, which is not in the test sets.

Both data sets exhibit a mild class and text length imbalance between parliaments. The data set's size was a technical challenge for some participants. The average text length is approximately 600 space-separated tokens, which is larger than the maximum accepted by many of the pretrained language models. Moreover, the data set is also large overall (more than 3GB uncompressed).

5.3 Participant Approaches

In 2024, 9 teams participated in this task and submitted 52 runs. We added a baseline for comparison. Unlike the ValueEval task, where pretrained language models were the dominant classifiers, for this task many participants preferred traditional, 'computationally light' approaches. A possible reason may be the large text size which is more costly to process with larger systems. Most teams, even the teams that used language models with large context sizes, truncated the texts to alleviate computational requirements. Some of the interesting improvements include ensemble of classifiers, data augmentation through back-translation and synonym replacement, multi-task learning, additional features, such as sentiment scores, and the use of domain-specific models.

Baselines. We provided only a single logistic regression baseline with tf-idf weighted character n-grams. The baseline is intentionally kept simple to encourage participation by early researchers, and reduce the computation requirements.

Team Policy Parsing Panthers [54]. The team did a set of experiments with original transcripts and their English translations, using various deep pretrained models, including BERT [14], mBERT [14], RoBERTa [46], XLM-RoBERTa [12],

[13] Training and test data are available at https://zenodo.org/doi/10.5281/zenodo.10450640, and https://zenodo.org/doi/10.5281/zenodo.11061649 respectively.

DeBERTa-v3 [32] Gemma [47] and ensembles of these models. This team presents an extensive set of approaches, and their analyses. A few interesting approaches worth mentioning in this short summary includes (1) Data augmentation and balancing through back-translation, (2) experiments with additional metadata, (3) multi-task learning, (4) the use of automatically obtained polarity labels, and increasing the number of instances in the training set of the orientation subtask by using the matching speaker IDs in the power dataset. This team participated in both subtasks for all parliaments.

Team Trojan Horses [48]. The team experimented with improving the logistic regression baseline, as well as fine-tuning BERT. They used the English translations and participated in both subtasks for the majority of the parliaments.

Team Pixel Phantoms [31]. The team experimented with some of the traditional classifiers (SVMs, logistic regression and decision trees) using the English translations provided. As well as tf-idf weighted features, they also extracted text embeddings from DistilBERT [63], through Sentence BERT [60]. They participated in both subtasks for the majority of the parliaments.

Team Ssnites [73]. The team fine-tuned BERT for the majority of parliaments and both subtasks. They relied on the English translations provided, and participated in both subtasks for the majority of the parliaments.

Team Hale Lab [68]. After some initial experiments with BERT, the team used a variety of classification methods including simple feed-forward networks, and LSTMs. The features for the models were either bag-of-words features weighted with tf-idf, or the multilingual LASER [6] embeddings. They used the original (untranslated) data, using various libraries for tokenization and preprocessing, and participated in both subtasks for the majority of the parliaments.

Team Vayam Solve Kurmaha [69]. This team also experimented with multiple traditional classification methods (SVM, kNN, random forests) and their ensembles, using the English translations. The team also used data augmentation through synonym replacement. They participated in both subtasks for the majority of the parliaments.

Team Gerber [26]. The team used a convolutional neural network (CNN) for the task without any pretrained embeddings. They used the original transcripts only, and participated in both subtasks for the majority of the parliaments.

Team JU_NLP_DID [36]. The team used SVM classifiers with tf-idf features, participating in both subtasks for the majority of the parliaments. They also make use of automatic sentiment labels as an additional feature.

Team INSA Passau [4]. The team also experimented with multiple approaches, where some of their submissions were focused on orientation identification and a smaller number of parliaments. The methods used included training SVMs, fine-tuning BERT-based models (pre)trained on legal documents [9,79] and fine-tuning and zero- and few-shot prompting the Llama [72] version 3 models with varying sizes (which were released during while the shared task was running).

Table 4. F_1-scores of the best submissions per team (as measured by overall F_1-score) on ideology identification task. Baseline scores are shown in gray.

	F_1-score																										
Team	**Overall**	**AT**	**BA**	**BE**	**BG**	**CZ**	**DK**	**EE**	**ES**	**ES-CT**	**ES-GA**	**FI**	**FR**	**GB**	**GR**	**HR**	**HU**	**IS**	**IT**	**LV**	**NL**	**NO**	**PL**	**PT**	**RS**	**SE**	**SI**	**TR**	**UA**
Policy Parsing Panthers	79	77	51	71	77	63	84	64	94	80	98	77	75	92	89	65	87	71	77	67	71	82	88	95	79	95	78	93	83
gerber	63	60	45	54	62	52	56	00	77	66	76	54	58	76	72	51	69	00	60	49	59	00	72	69	64	00	58	84	73
HALE Lab	61	56	44	59	60	52	56	52	76	69	84	52	48	74	71	43	67	57	60	49	53	61	62	67	55	77	49	83	60
Pixel Phantoms	59	58	49	56	56	47	56	54	72	64	75	59	58	72	71	55	68	57	57	54	60	54	59	54	51	61	47	78	56
Ssnites	59	50	53	55	53	50	61	52	61	58	64	55	56	64	59	53	60	58	53	51	56	66	71	64	64	75	58	79	53
Trojan Horses	59	61	25	57	61	51	60	57	72	67	00	33	60	73	74	53	71	55	66	00	60	61	68	63	00	74	00	80	68
INSA Passau	59	60	53	54	61	47	57	53	63	61	66	34	58	69	59	56	66	56	56	54	56	58	69	55	61	66	51	80	62
JU_NLP_DID	57	53	42	42	55	51	60	57	69	57	70	00	50	71	63	43	60	55	61	47	56	59	51	67	48	73	46	77	57
Baseline	56	52	42	45	53	52	56	47	72	65	67	54	43	74	74	43	57	39	56	45	51	62	46	63	53	75	39	84	58

5.4 Task Evaluation

We use macro-averaged F_1-score as the main evaluation metric for both subtasks. Similar to the ValueEval task, the participants were encouraged to submit confidence scores, where a score over 0.5 is interpreted as class 1 and otherwise 0.

Table 4 and Table 5 present the overall best-performing approaches per team for the ideology and power subtasks respectively. The best scores for both tasks are from the team Policy Parsing Panthers. The team used an ensemble of multiple models, with multiple improvements including data augmentation and multi-task learning. Results on the tables do not include approaches that were focused on only one or a small number of parliaments. A noteworthy focused submission for only GB and ideology subtask by the team INSA Passau based on fine-tuning the most recent Llama 3 model achieved the second-best result for this parliament. Although the results on both tasks are higher than the baseline we provided, the variation in the scores indicate that there is quite some room for improvement for each of the approaches.

We also observe that, as formulated in this task, identifying orientation is slightly more difficult than identifying power. The overall success of the systems on a particular parliament depends on, among others, size and class distribution of the training data, and composition of the parliament. For example, we observe a general trend (with some exceptions) that for parliaments with few or no government and opposition role changes in the data (e.g., HU, PL, and TR) the roles are easier to predict than for parliaments with more varied composition and more role changes (e.g., AT, BA, and UA).

Table 5. F_1-scores of the best submissions per team (as measured by overall F_1-score) on power identification task for each parliament. Baseline scores are shown in gray.

	F_1-score																							
Team	**Overall**	**AT**	**BA**	**BE**	**BG**	**CZ**	**DK**	**ES**	**ES-CT**	**ES-GA**	**ES-PV**	**FI**	**FR**	**GB**	**GR**	**HR**	**HU**	**IT**	**LV**	**NL**	**PL**	**PT**	**RS**	**SI**	**TR**	**UA**
Policy Parsing Panthers	83	88	56	74	81	78	87	88	91	98	90	80	82	83	95	75	97	78	75	74	90	85	84	81	94	65
HALE Lab	70	69	46	61	68	69	70	65	85	88	78	65	67	75	82	68	88	69	62	64	78	65	69	61	84	49
Trojan Horses	69	72	57	63	67	63	68	69	82	85	74	39	66	72	83	67	86	72	64	64	74	65	75	62	83	56
gerber	68	68	51	60	66	64	63	72	80	86	74	60	71	72	68	63	87	52	63	64	77	66	73	58	84	48
Vayam Solve Kurmaha	68	48	48	65	69	68	69	72	83	87	76	35	66	47	85	67	88	72	62	68	75	67	75	63	85	48
Pixel Phantoms	66	70	50	59	63	65	69	65	64	77	69	61	64	73	72	57	80	69	58	62	70	66	69	60	80	52
Baseline	64	66	45	61	68	64	56	65	78	83	71	56	66	71	63	60	86	43	51	62	76	62	65	53	83	46
JU_NLP_DID	63	68	47	55	58	57	67	60	78	55	72	00	59	00	77	65	83	71	47	63	70	63	54	56	78	43
INSA Passau	62	67	45	60	66	65	54	65	00	00	00	56	66	72	56	61	85	45	52	64	77	62	63	54	84	47
Ssnites	60	66	45	58	60	61	61	62	58	62	60	60	65	60	69	65	79	62	54	57	62	58	60	57	61	46

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

Images provide powerful visual communication, are usually perceived before text is read, and can appeal directly to our emotions. The goal of this task is to find images that convey premises. The proper use of an image can increase the persuasiveness of an argument. In this regard, images can increase the pathos [59], which is the effect an argument has on its audience.

6.1 Task Definition

This observation leads to our task, in which participants are asked to find images based on an argument that help to convey the premise of the argument. In this context, "convey" is meant in broad terms; it can represent what is described in the argument, but it can also show a generalization (e.g., a symbolic image that illustrates a related abstract concept) or a specialization (e.g., a concrete example). There is a difference between verbal language and images. Verbal language provides clear but limited information, while images provide more information than written words, but are not as precise [39]. Therefore, images alone can be ambiguous and difficult to understand without context, e.g. when they refer to symbolism. For this reason, we offer the option of submitting a rationale together with the image. The rationale is an explanatory statement that assists in understanding the picture. For example, it can be a caption or contextual information about the image. The image and the rationale are evaluated together to see how this combination conveys the premise. Participants can choose to use a retrieval approach, where they submit images from a provided dataset, or a generation-based approach, where suitable images can be generated using a model of their choice. In each submission, a participant can submit up to 10 images in a ranking order for an argument.

```
<argument>
  <id>36062-a-3</id>
  <topic>Should boxing be banned?</topic>
  <premise>
    The idea of winning through intentional infliction of pain and harm
    to another person can nurture a violent and destructive mentality.
  </premise>
  <claim>
    Boxing poses both physical and psychological threats to
    participants, hence it should be banned.
  </claim>
  <stance>pro</stance>
  <type>ANECDOTAL</type>
</argument>
```

Fig. 3. Example argument from the data set. The argument consists of an id, a premise and a claim. We also indicate the topic of the argument, as well as the argument's stance on the topic. The type element indicates that the arguments relies on anecdotal evidence. Only arguments of this type are used in our dataset.

6.2 Data Description

For the task we prepared a dataset[14] containing 136 arguments and over 9000 images. The arguments were generated with GPT-4 [2] and correspond to 24 topics. The topics were taken from various IBM datasets[15] and previous Touché Shared Tasks[16]. Each generated argument consists of a premise and a claim, and can take a pro or con stance on the topic. An example of an argument can be seen in Fig. 3. Each of the images in the dataset is tagged with additional information, such as the URL and content of the corresponding website. In addition, we have provided an analysis of each image using the Google Cloud Vision API, as well as an automatically generated caption using LLaVA [44].

6.3 Participant Approaches

In 2024, 2 teams participated in this task and submitted 8 runs. All teams chose the retrieval-approach. Moreover, we added 2 baseline runs for comparison.

Baselines. The first baseline is BM25, where the corresponding documents are the image captions from the data set and the query is the premise of the argument. In the second baseline, keywords are first extracted from the image captions. Then embeddings for the premise of an argument and the keywords are generated with SBERT [60]. A corresponding relevance score is calculated based on the cosine similarity between the embeddings and averaging them. The most relevant images are selected for submission.

[14] https://zenodo.org/records/11045831.
[15] https://research.ibm.com/haifa/dept/vst/debating_data.shtml.
[16] https://touche.webis.de/shared-tasks.html.

DS@GT [53]. The team uses CLIP [58] to embed each argument and each image in a common embedding space. The first approach ranks images by cosine similarity of the embeddings. The second approach compares for each argument the 40 highest ranked images to images that are generated to support or attack the argument. The most similar images are submitted.

HTW-DIL [33]. The team has chosen an approach inspired by DPR [35]. It applies a fine-tuned multimodal Moondream model based on the Phi 1.5 LLM [43] and uses SigLIP [78] for its vision capabilities. To generate synthetic training data, the team uses GPT-4 to generate arguments from the available image/web page data. Combinations of positive and negative argument-image pairs are used for training. The results are obtained by maximising the cosine similarity for argument and image embeddings.

6.4 Task Evaluation

For each argument and each submission, the best 5 images together with the rationales are evaluated by a human expert. This expert knows neither the rank of the image nor the team that submitted it. To facilitate the annotation, we prepared a narrative for each argument that describes what a conveying image should generally show. Therefore, each combination of image, argument and rationale is rated on a three-point Likert scale from 0 to 2, where 0 means that the image does not convey the premise at all, 1 stands for partial conveyance and 2 means that the image conveys the premise completely. For seven topics, only very few relevant images could be submitted by the participating teams, so we removed these topics, resulting in a total number of 104 arguments for the evaluation. For each submission, we first calculated the NDCG score for each argument. For the required IDCG, we have considered all submitted image, argument and justification triples submitted for the corresponding argument. The final score of a submission is the average of all NDCG scores for all arguments. The results of the shared task can be seen in Table 6. To conclude, it can be said that the relevance of an image is often determined by implicit assumptions and is subject to interpretation. Therefore, the identification of conveying images is still a very challenging task.

7 Conclusion

The fifth edition of the Touché lab on argumentation systems featured three tasks: (1) Human Value Detection, (2) Ideology and Power Identification in Parliamentary Debates, and (3) Image Retrieval/Generation for Arguments. In contrast to previous years, the focus this year was more on classification than retrieval tasks. Furthermore, two of the three tasks were multilingual, although automatic English transcriptions were provided to facilitate participation. We expanded the scope of Touché with the new tasks on human values and political power and orientation. In addition, we methodically extended the retrieval task

Table 6. NDCG values for the top 5, top 3, and most relevant image(s). The approachesare sorted according to the NDCG@5 score.

Rank	Team	Approach	NDCG@5	NDCG@3	NDCG@1
1	HTW-DIL	Ada-Summary	0.428	0.409	0.404
2	HTW-DIL	Moondream-Text	0.363	0.355	0.356
3	HTW-DIL	Moondream-Default-Image-Text	0.293	0.302	0.317
4	Baseline	BM25	0.284	0.273	0.293
5	Baseline	SBERT	0.232	0.225	0.221
6	DS@GT	Generated-Image-Clip	0.180	0.178	0.197
7	HTW-DIL	Moondream-Image-Text-EP3	0.150	0.163	0.183
8	HTW-DIL	Moondream-Image	0.146	0.155	0.178
9	DS@GT	Base-Clip-Submission	0.123	0.111	0.106
10	HTW-DIL	Moondream-Image-Text	0.120	0.140	0.178

by allowing participants to generate images instead of retrieving them. Unfortunately, no team submitted generated images in the end.

Of the 68 registered teams, 20 participated in the tasks and submitted a total of 81 runs. Participants mainly used classification architectures, with BERT and variants still very dominant, although more classical machine learning models were also used in the Ideology and Power Identification in Parliamentary Debates task. Generative models, on the other hand, were rarely used. Although the Image Retrieval/Generation for Arguments task changed to seeking images for a specific argument rather than a topic, the approaches submitted were similar to previous years. They embedded the images from the collection and then used the similarity to the query for ranking, either by embedding the query directly or generating images for the query and embedding those.

We plan to continue Touché as a collaborative platform for researchers in argumentation systems. All Touché resources are freely available, including topics, manual relevance, argument quality, and stance judgments, and submitted runs from participating teams. These resources and other events such as workshops will help to further foster the community working on argumentation systems.

Acknowledgments. This work was partially supported by the European Commission under grant agreement GA 101070014 (https://openwebsearch.eu) and the German Research Foundation under project 455911521 (LARGA) as part of the SPP 1999 (RATIO). The ideology and power identification shared task has been supported by CLARIN ERIC, under the ParlaMint project (https://www.clarin.eu/parlamint).

References

1. Abercrombie, G., Batista-Navarro, R.: Sentiment and position-taking analysis of parliamentary debates: a systematic literature review. J. Comput. Soc. Sci. **3**(1), 245–270 (2020)
2. Achiam, J., et al.: GPT-4 Technical Report (2024)
3. Ammanabrolu, P., Jiang, L., Sap, M., Hajishirzi, H., Choi, Y.: Aligning to social norms and values in interactive narratives. In: Carpuat, M., de Marneffe, M., Ruíz, I.V.M. (eds.) Proceedings of NAACL-HLT 2022, pp. 5994–6017. ACL (2022). https://doi.org/10.18653/v1/2022.naacl-main.439
4. Andruszak, M., Alhamzeh, A., Egyed-Zsigmond, E., Carlsson, A., Leydet, J., Otiefy, Y.: Team INSA Passau at Touché: multi-lingual parliamentary speech classification. In: Faggioli, G., Ferro, N., Galuščáková, P., de Herrera, A.G.S. (eds.) Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024). CEUR Workshop Proceedings, CEUR-WS.org (2024)
5. Arian, A., Shamir, M.: The primarily political functions of the left-right continuum. Comp. Polit. **15**(2), 139–158 (1983)
6. Artetxe, M., Schwenk, H.: Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. Trans. Assoc. Comput. Linguist. **7**, 597–610 (2019). https://doi.org/10.1162/tacl_a_00288
7. Aydin, A., Shaar, S., Cardie, C.: Edward said at touché: human values classification. In: Faggioli, G., Ferro, N., Galuščáková, P., de Herrera, A.G.S. (eds.) Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024). CEUR Workshop Proceedings, CEUR-WS.org (2024)
8. Bench-Capon, T.: Persuasion in practical argument using value-based argumentation frameworks. J. Logic Comput. **13**(3), 429–448 (2003). https://doi.org/10.1093/logcom/13.3.429
9. Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., Androutsopoulos, I.: LEGAL-BERT: the muppets straight out of law school. In: Cohn, T., He, Y., Liu, Y. (eds.) Findings of ACL: EMNLP 2020, pp. 2898–2904. ACL (2020). https://doi.org/10.18653/v1/2020.findings-emnlp.261
10. Chen, C., Walker, D., Saligrama, V.: Ideology prediction from scarce and biased supervision: learn to disregard the "what" and focus on the "how"! In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Proceedings of ACL (Volume 1: Long Papers), Toronto, Canada, pp. 9529–9549. ACL (2023). https://doi.org/10.18653/v1/2023.acl-long.530
11. Çöltekin, Ç., Kopp, M., Katja, M., Morkevicius, V., Ljubešić, N., Erjavec, T.: Multilingual power and ideology identification in the parliament: a reference dataset and simple baselines. In: Fiser, D., Eskevich, M., Bordon, D. (eds.) 4th Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora, pp. 94–100. ELRA and ICCL (2024). https://aclanthology.org/2024.parlaclarin-1.14
12. Conneau, A., et al.: Unsupervised cross-lingual representation learning at scale. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J.R. (eds.) Proceedings of ACL, pp. 8440–8451. ACL (2020). https://doi.org/10.18653/v1/2020.acl-main.747
13. Conover, M.D., Gonçalves, B., Ratkiewicz, J., Flammini, A., Menczer, F.: Predicting the political alignment of Twitter users. In: Proceedings of PASSAT and SocialCom, pp. 192–199. IEEE (2011). https://doi.org/10.1109/PASSAT/SocialCom.2011.34
14. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio,

T. (eds.) Proceedings of NAACL-HLT, pp. 4171–4186. ACL (2019). https://doi.org/10.18653/V1/N19-1423
15. van Dijk, T.: Discourse and Power. Bloomsbury Publishing (2008)
16. Dimitrov, D., et al.: SemEval-2021 task 6: detection of persuasion techniques in texts and images. In: Proceedings of SemEval, pp. 70–98. ACL (2021). https://doi.org/10.18653/v1/2021.semeval-1.7. https://aclanthology.org/2021.semeval-1.7
17. Dove, I.J.: On images as evidence and arguments. In: van Eemeren, F., Garssen, B. (eds.) Topical Themes in Argumentation Theory, vol. 22, pp. 223–238. Springer, Dordrecht (2012). https://doi.org/10.1007/978-94-007-4041-9_15
18. Dunaway, F.: Images, emotions, politics. Mod. Am. Hist. **1**(3), 369–376 (2018). https://doi.org/10.1017/mah.2018.17
19. Erjavec, T., Ogrodniczuk, M., et al.: The ParlaMint corpora of parliamentary proceedings. LREC **57**, 415–448 (2022). https://doi.org/10.1007/s10579-021-09574-0
20. Fairclough, N.: Critical Discourse Analysis: The Critical Study of Language. Longman Applied Linguistics. Taylor & Francis (2013). https://doi.org/10.4324/9781315834368
21. Fairclough, N.: Language and Power. Language In Social Life. Taylor & Francis (2013). https://doi.org/10.4324/9781315838250
22. Fišer, D., Lenardič, J.: CLARIN resources for parliamentary discourse research. In: Fišer, D., Eskevich, M., de Jong, F. (eds.) Proceedings of LREC. ELRA (2018)
23. Forbes, M., Hwang, J.D., Shwartz, V., Sap, M., Choi, Y.: Social chemistry 101: learning to reason about social and moral norms. In: Webber, B., Cohn, T., He, Y., Liu, Y. (eds.) Proceedings of EMNLP, pp. 653–670. ACL (2020). https://doi.org/10.18653/v1/2020.emnlp-main.48
24. Fröbe, M., et al.: Continuous integration for reproducible shared tasks with TIRA.io. In: Kamps, J., et al. (eds.) ECIR 2023. LNCS, vol. 13982, pp. 236–241. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-28241-6_20
25. García-Díaz, J.A., et al.: Overview of PoliticES 2022: Spanish author profiling for political ideology. Procesamiento del Lenguaje Natural **69**, 265–272 (2022). https://doi.org/10.26342/2022-69-23
26. Gerber, C.: Gerber at touché: ideology and power identification in parliamentary debates 2024. In: Faggioli, G., Ferro, N., Galuščáková, P., de Herrera, A.G.S. (eds.) Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024). CEUR Workshop Proceedings, CEUR-WS.org (2024)
27. Gerrish, S., Blei, D.M.: Predicting legislative roll calls from text. In: Getoor, L., Scheffer, T. (eds.) Proceedings of ICML, pp. 489–496. Omnipress (2011)
28. Glavaš, G., Nanni, F., Ponzetto, S.P.: Computational analysis of political texts: bridging research efforts across communities. In: 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts, pp. 18–23. ACL (2019). https://doi.org/10.18653/v1/P19-4004
29. Goyal, N., Du, J., Ott, M., Anantharaman, G., Conneau, A.: Larger-scale transformers for multilingual masked language modeling. In: Rogers, A., et al. (eds.) Proceedings of RepL4NLP@ACL-IJCNLP, pp. 29–33. ACL (2021). https://doi.org/10.18653/V1/2021.REPL4NLP-1.4
30. Grancea, I.: Types of visual arguments. Argumentum. J. Seminar Discursive Logic Argumentation Theory Rhetoric **15**(2), 16–34 (2017)
31. Hariharakrishnan, J., Mirunalini, P.: Pixel phantoms at touché: ideology and power identification in parliamentary debates using linear SVC. In: Faggioli, G., Ferro, N., Galuščáková, P., de Herrera, A.G.S. (eds.) Working Notes of the Conference

and Labs of the Evaluation Forum (CLEF 2024). CEUR Workshop Proceedings, CEUR-WS.org (2024)
32. He, P., Liu, X., Gao, J., Chen, W.: DeBERTa: decoding-enhanced BERT with disentangled attention. In: Proceedings of ICLR (2021). https://openreview.net/forum?id=XPZIaotutsD
33. Janusko, T., Kämpf, A., Keiling, D., Knick, J., Thiele, D.S.M.: Htw-dil at touché: multimodal dense information retrieval for arguments. In: Faggioli, G., Ferro, N., Galuščáková, P., de Herrera, A.G.S. (eds.) Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024). CEUR Workshop Proceedings, CEUR-WS.org (2024)
34. Kiesel, J., et al.: SCaLAR NITK at touché: human value detection. In: Faggioli, G., Ferro, N., Galuščáková, P., de Herrera, A.G.S. (eds.) Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024). CEUR Workshop Proceedings, CEUR-WS.org (2024)
35. Karpukhin, V., et al.: Dense passage retrieval for open-domain question answering. In: Proceedings of EMNLP, pp. 6769–6781. ACL (2020). https://doi.org/10.18653/v1/2020.emnlp-main.550
36. Khurshid, A., Das, D., Khaskel, R., Datta, S.: JU_NLP_DID at touché. In: Faggioli, G., Ferro, N., Galuščáková, P., de Herrera, A.G.S. (eds.) Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024). CEUR Workshop Proceedings, CEUR-WS.org (2024)
37. Kiesel, J., Alshomary, M., Handke, N., Cai, X., Wachsmuth, H., Stein, B.: Identifying the human values behind arguments. In: Muresan, S., Nakov, P., Villavicencio, A. (eds.) Proceedings of ACL, pp. 4459–4471. ACL (2022). https://doi.org/10.18653/v1/2022.acl-long.306
38. Kiesel, J., et al.: SemEval-2023 task 4: ValueEval: identification of human values behind arguments. In: Kumar, R., Ojha, A.K., Doğruöz, A.S., Martino, G.D.S., Madabushi, H.T. (eds.) Proceedings of SemEval, pp. 2287–2303. ACL (2023). https://doi.org/10.18653/v1/2023.semeval-1.313
39. Kjeldsen, J.E.: Virtues of visual argumentation: how pictures make the importance and strength of an argument salient (2013)
40. Kurtoğlu Eskişar, G.M., Çöltekin, Ç.: Emotions running high? A synopsis of the state of Turkish politics through the ParlaMint corpus. In: Fišer, D., Eskevich, M., Lenardič, J., de Jong, F. (eds.) Proceedings of ParlaCLARIN, pp. 61–70. ELRA (2022). https://aclanthology.org/2022.parlaclarin-1.10
41. Legkas, S., Christodoulou, C., Zidianakis, M., Koutrintzes, D., Petasis, G., Dagioglou, M.: Hierocles of alexandria at touché: multi-task & multi-head custom architecture with transformer-based models for human value detection. In: Faggioli, G., Ferro, N., Galuščáková, P., de Herrera, A.G.S. (eds.) Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024). CEUR Workshop Proceedings, CEUR-WS.org (2024)
42. Lenardič, J., Fišer, D.: CLARIN Resource Families: Parliamentary Corpora (2023). https://www.clarin.eu/resource-families/parliamentary-corpora. Accessed 09 July 2024
43. Li, Y., Bubeck, S., Eldan, R., Giorno, A.D., Gunasekar, S., Lee, Y.T.: Textbooks are All You Need II: phi-1.5 technical report (2023). https://arxiv.org/abs/2309.05463
44. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning (2023)
45. Liu, R., Jia, C., Zhang, G., Zhuang, Z., Liu, T.X., Vosoughi, S.: Second thoughts are best: learning to re-align with human values from text edits. In: Advances in Neural Information Processing Systems, vol. 35, pp. 181–196 (2022)

46. Liu, Y., et al.: RoBERTa: a robustly optimized BERT pretraining approach. CoRR (2019). http://arxiv.org/abs/1907.11692
47. Mesnard, T., et al.: Gemma: open models based on Gemini research and technology (2024). https://doi.org/10.48550/arXiv.2403.08295
48. Mirunalini, P., Koushik, A., Seshan, D.: Trojan horses at touché: logistic regression for classification of political debates. In: Faggioli, G., Ferro, N., Galuščáková, P., de Herrera, A.G.S. (eds.) Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024). CEUR Workshop Proceedings, CEUR-WS.org (2024)
49. Mochtak, M., Rupnik, P., Ljubešić, N.: The ParlaSent multilingual training dataset for sentiment identification in parliamentary proceedings. In: Calzolari, N., Kan, M.Y., Hoste, V., Lenci, A., Sakti, S., Xue, N. (eds.) Proceedings of LREC, pp. 16024–16036. ELRA and ICCL (2024). https://aclanthology.org/2024.lrec-main.1393
50. Morren, M., Mishra, R.: Eric from at touché: prompts vs finetuning. In: Faggioli, G., Ferro, N., Galuščáková, P., de Herrera, A.G.S. (eds.) Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024). CEUR Workshop Proceedings, CEUR-WS.org (2024)
51. Navarretta, C., Haltrup Hansen, D.: Government and opposition in Danish parliamentary debates. In: Fiser, D., Eskevich, M., Bordon, D. (eds.) Proceedings of ParlaCLARIN, pp. 154–162. ELRA and ICCL (2024). https://aclanthology.org/2024.parlaclarin-1.23
52. Nguyen, M.V., Lai, V.D., Veyseh, A.P.B., Nguyen, T.H.: Trankit: a light-weight transformer-based toolkit for multilingual natural language processing. In: Gkatzia, D., Seddah, D. (eds.) Proceedings of EACL, pp. 80–90. ACL (2021). https://doi.org/10.18653/v1/2021.eacl-demos.10
53. Ostrower, B., Aphiwetsa, P.: Ds@gt at touché: image search and ranking via clip and image generation. In: Faggioli, G., Ferro, N., Galuščáková, P., de Herrera, A.G.S. (eds.) Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024). CEUR Workshop Proceedings, CEUR-WS.org (2024)
54. Palmqvist, O., Jiremalm, J., Picazo-Sanchez, P.: Policy parsing panthers at touché: ideology and power identification in parliamentary debates. In: Faggioli, G., Ferro, N., Galuščáková, P., de Herrera, A.G.S. (eds.) Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024). CEUR Workshop Proceedings, CEUR-WS.org (2024)
55. Pla, F., Hurtado, L.F.: Political tendency identification in Twitter using sentiment analysis techniques. In: Tsujii, J., Hajic, J. (eds.) Proceedings of Coling, pp. 183–192. Dublin City University and ACL (2014). urlhttps://aclanthology.org/C14-1019
56. Preoţiuc-Pietro, D., Liu, Y., Hopkins, D., Ungar, L.: Beyond binary labels: political ideology prediction of Twitter users. In: Barzilay, R., Kan, M.Y. (eds.) Proceedings of ACL, pp. 729–740. ACL (2017). https://doi.org/10.18653/v1/P17-1068
57. Qiu, L., et al.: ValueNet: a new dataset for human value driven dialogue system. In: Proceedings of AAAI, pp. 11183–11191. AAAI Press (2022). https://doi.org/10.1609/aaai.v36i10.21368
58. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (eds.) Proceedings of ICML, vol. 139, pp. 8748–8763. PMLR (2021). https://proceedings.mlr.press/v139/radford21a.html
59. Rapp, C.: Aristotle's rhetoric. In: Zalta, E.N., Nodelman, U. (eds.) The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University (2023)

60. Reimers, N., Gurevych, I.: Sentence-BERT: sentence embeddings using Siamese BERT-networks. In: Proceedings of EMNLP, pp. 3982–3992. ACL (2019). https://doi.org/10.18653/v1/D19-1410
61. Roque, G.: Visual argumentation: a further reappraisal. In: van Eemeren, F.H., Garssen, B. (eds.) Topical Themes in Argumentation Theory, vol. 22, pp. 273–288. Springer, Cham (2012). https://doi.org/10.1007/978-94-007-4041-9_18
62. Russo, D., et al.: PoliticIT at EVALITA 2023: overview of the political ideology detection in italian texts task. In: Proceedings of EVALITA. CEUR Workshop Proceedings, vol. 3473. CEUR-WS.org (2023). https://ceur-ws.org/Vol-3473/paper7.pdf
63. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter (2020)
64. Scharfbillig, M., Ponizovskiy, V., Pasztor, Z., Keimer, J., Tirone, G.: Monitoring social values in online media articles on child vaccinations. Technical report, European Commission's Joint Research Centre, Luxembourg (2022). https://doi.org/10.2760/86884
65. Scharfbillig, M., et al.: Values and identities - a policymaker's guide. Technical report, European Commission's Joint Research Centre, Luxembourg (2021). https://doi.org/10.2760/349527
66. Schwartz, S.H.: Are there universal aspects in the structure and contents of human values? J. Soc. Issues 19–45 (1994). https://doi.org/10.1111/j.1540-4560.1994.tb01196.x
67. Schwartz, S.H., et al.: Refining the theory of basic individual values. J. Pers. Soc. Psychol. (2012). https://doi.org/10.1037/a0029393
68. Sevitha, S., Patel, M., Shevgoor, S.: Team hale lab at touché 2024: ideology and power identification in parliamentary debates. In: Faggioli, G., Ferro, N., Galuščáková, P., de Herrera, A.G.S. (eds.) Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024). CEUR Workshop Proceedings, CEUR-WS.org (2024)
69. Shwetha, S., Kamath, S., Balaji, S., Narayanan, S.: Vayam solve Kurmaha at touché: power identification in parliamentry speeches using TFIDF vectorizer and SVM classifier. In: Faggioli, G., Ferro, N., Galuščáková, P., de Herrera, A.G.S. (eds.) Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024). CEUR Workshop Proceedings, CEUR-WS.org (2024)
70. Stefanovitch, N., Piskorski, J.: Holistic inter-annotator agreement and corpus coherence estimation in a large-scale multilingual annotation campaign. In: Bouamor, H., Pino, J., Bali, K. (eds.) Proceedings of EMNLP, pp. 71–86. ACL (2023). https://doi.org/10.18653/v1/2023.emnlp-main.6
71. Tarkka, O., et al.: Automated emotion annotation of Finnish parliamentary speeches using GPT-4. In: Fiser, D., Eskevich, M., Bordon, D. (eds.) Proceedings of ParlaCLARIN, pp. 70–76. ELRA and ICCL (2024). https://aclanthology.org/2024.parlaclarin-1.11
72. Touvron, H., et al.: LLaMA: Open and Efficient Foundation Language Models (2023). https://doi.org/10.48550/arxiv.2302.13971
73. Kiesel, J., et al.: Ssnites at touché: ideology and power identification in parliamentary debates using BERT model. In: Faggioli, G., Ferro, N., Galuščáková, P., de Herrera, A.G.S. (eds.) Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024). CEUR Workshop Proceedings, CEUR-WS.org (2024)
74. Vegetti, F., Širinić, D.: Left-right categorization and perceptions of party ideologies. Polit. Behav. **41**(1), 257–280 (2019)

75. Wachsmuth, H., et al.: Computational argumentation quality assessment in natural language. In: Proceedings of EACL, pp. 176–187 (2017). https://aclanthology.org/E17-1017
76. Yeste, V., Ardanuy, M.C., Rosso, P.: Philo of Alexandria at touché: a cascade model approach to human value detection. In: Faggioli, G., Ferro, N., Galuščáková, P., de Herrera, A.G.S. (eds.) Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024). CEUR Workshop Proceedings, CEUR-WS.org (2024)
77. Yunis, H.: Arthur schopenhauer at touché 2024: multi-lingual text classification using ensembles of large language models. In: Faggioli, G., Ferro, N., Galuščáková, P., de Herrera, A.G.S. (eds.) Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024). CEUR Workshop Proceedings, CEUR-WS.org (2024)
78. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of ICCV, pp. 11941–11952. IEEE Computer Society (2023). https://doi.org/10.1109/iccv51070.2023.01100
79. Zheng, L., Guha, N., Anderson, B.R., Henderson, P., Ho, D.E.: When does pretraining help?: assessing self-supervised learning for law and the CaseHOLD dataset of 53,000+ legal holdings. In: Proceedings of ICAIL, pp. 159–168. ACM (2021). https://doi.org/10.1145/3462757.3466088