

# Climate and Migration on German Political Party YouTube: A Corpus with Stance, Propaganda, and Divisiveness Annotation

Zaher Alkaei<sup>1</sup>, Manfred Stede<sup>1</sup>, Christoph M. Abels<sup>2</sup>

<sup>1</sup>Applied Computational Linguistics, University of Potsdam, Potsdam, Germany

<sup>2</sup>Department of Psychology, University of Potsdam, Potsdam, Germany

Correspondence: [alkaei@uni-potsdam.de](mailto:alkaei@uni-potsdam.de)

## Abstract

We present a German-language corpus for studying divisiveness in political social-media discourse: 7,277 parent–child comment pairs sampled from the YouTube channels of two ideologically opposed German parties, the AfD and Die Linke, on two contested policy domains, climate change and migration. Each pair is annotated on three dimensions: interaction quality (a five-class constructive–destructive  $\times$  agreement–disagreement scale), stance (of parent and child toward the topic), and the presence of 14 propaganda techniques. We detail the full construction pipeline – channel selection, comment and transcript collection, pseudonymization, conversation-tree reconstruction, topic filtering, and stratified sampling – and a reproducible three-task annotation protocol: a 700-pair human gold standard validates an LLM-assisted pipeline that scales annotation to the full corpus. Descriptive analysis reveals pronounced party- and topic-specific patterns – most notably that the Migration  $\times$  AfD stratum is the most hostile (54.5% of its pairs are Destructive Agreement) – while thread-level quality does not differ significantly between the two topics ( $\chi^2(3) = 5.15$ ,  $p = 0.16$ ). Only on climate are exchanges between opposing sides (a skeptic replying to a supporter) constructive on average; the same clash on migration remains destructive, and the most destructive exchanges overall occur between two opponents of the policy at issue.

## 1 Introduction

Online political discourse is increasingly polarized, yet modeling its dynamics remains methodologically challenging. Interlocutors negotiate disagreement within a pragmatic “disagreement space” (Jackson, 1992; Schumann and Oswald, 2024) where communication extends beyond propositional truth to the legitimacy, relevance, and appropriateness of speech acts. To capture both what interlocutors contest and how they contest it, we

adopt the five-class interaction quality schema of Bassi et al. (2025), mapping comment-reply pairs along joint axes of agreement-disagreement and constructiveness-destructiveness.

A central motivation for this resource is that the surface act of disagreement does not by itself make an exchange destructive: opposing sides can still engage constructively. Consider a representative exchange from our corpus (translated from German):

### Parent Comment

“Climate crisis, crisis, crisis. . . nag, nag, nag.” (stance: against climate action)

### Child Comment (Reply)

“Then why are crop failures from extreme weather getting more frequent, especially in Europe? The steady decline in birds, bees, and insects should be obvious to anyone . . .” (stance: supports climate action)

The child disagrees with the parent’s dismissiveness, yet does so by pointing to concrete evidence rather than attacking. We label this pair *Constructive Disagreement* (interaction-quality score +1.0), with opposing parent and child stances and two techniques flagged on the child: *Appeal to Fear* (its warning of ecological consequences) and *Exaggeration/Minimisation* (the sweeping “obvious to anyone”). This is exactly the distinction that conventional toxicity-only or agreement-only annotation misses and that our three-layer scheme is designed to capture (see Table 5 for the full five-class spectrum). Conversely, agreement can be destructive: users who endorse one another while demeaning outsiders produce *Destructive Agreement*, which is in fact the dominant pattern in the corpus’s most adversarial stratum, Migration  $\times$  AfD (Section 5).

We build this resource from German-language YouTube comments on the channels of two ideologically opposed parties: the right-wing populist Alternative für Deutschland (AfD) and the left-wing Die Linke, focusing on two highly contested pol-

icy domains, climate/energy (Coffé et al., 2026; Drecker, 2025) and migration/asylum. These channels concentrate exactly the kind of adversarial partisan exchange in which divisiveness is most salient, yet no German-language corpus annotates such discourse for fine-grained interaction quality.

The two parties anchor opposite poles of German partisan conflict. The AfD occupies the anti-migration extreme while left parties such as Die Linke hold pro-migration positions, structuring a polarized competition over immigration (Atzpodien, 2022); the same parties increasingly diverge on climate (Drecker, 2025), an issue that has grown deeply divisive in Germany (Zilles and Marg, 2023), with support for climate action falling fastest on the cultural right (Coffé et al., 2026). Social media, in turn, amplify the most hostile migration narratives well beyond their private support base (Levi et al., 2025). This makes these channels a fertile setting for studying how disagreement turns constructive or destructive.

This paper makes three primary contributions. First, we introduce a new German-language corpus of 7,277 YouTube comment pairs (2,000 threads) spanning two parties and two topics, labeled for interaction quality, stance, and propaganda techniques. Second, we describe a reproducible, LLM-assisted annotation methodology validated against a 700-pair human gold standard, and make the full annotation guidelines and prompts available. Third, we provide a descriptive analysis characterizing how constructive and destructive dynamics, stance, and rhetorical techniques distribute across party and topic channels.

## 2 Related Work

### 2.1 Interaction Quality in Polarized Platform Spaces

Computational approaches to online controversy have transitioned from macro-level network segregation metrics (Conover et al., 2011; Garimella et al., 2018) to fine-grained conversational processing. Micro-level studies analyze textual affect to show that controversial comments express more anger than non-controversial ones (Chen et al., 2023), or quantify controversy intensity by combining stance antagonism with sentiment, offensiveness, and sarcasm (Wang et al., 2023). Grounded in pragma-dialectical theory, we take the quality of disagreement to turn on the balance of hedging and certainty (De Kock and Vlachos, 2021), resis-

tance to toxic escalation (Seckin et al., 2026), and argument and deliberative quality (Falk, 2025).

These conversational dynamics are heavily shaped by platform architecture. Asynchronous, text-based spaces like YouTube systematically exhibit lower mutual understanding and higher perceived conflict than spoken discourse (Bevis et al., 2026), and their architecture drives radicalization dynamics through user migration toward more extreme channels (Ribeiro et al., 2020). In the German far right, the comment sections of AfD, Pegida, and Identitarian channels increasingly converge on shared user communities (Rauchfleisch and Kaiser, 2020). Within such preference-driven, like-minded spaces, partisan forums exhibit lower toxic outrage than issue-driven mainstream arenas (Jakob et al., 2023). To capture these dynamics, we operationalize “divisiveness” as the destructive pole of the agreement and constructiveness axes in Bassi et al. (2025)’s schema.

### 2.2 Stance and Rhetorical Technique

We annotate stance and propaganda techniques as complementary layers to interaction quality rather than in isolation, since both are theoretically tied to divisiveness. Stance – target-specific favorability (Mohammad et al., 2016) – is often orthogonal to conversational style: ideological opponents can engage in constructive debate, while aligned users frequently reinforce one another’s hostility. We therefore annotate stance (Against, Neutral, Support) at both the parent and child level, capturing configurations such as constructive exchanges between opposing stances (Bassi et al., 2025).

Rhetorical manipulation is likewise tightly coupled with destructive interaction. While explicit toxicity is easily flagged by lexical cues, covert ideological appeals and democratic norm violations require finer rhetorical distinctions (Stoll et al., 2020). Grounded in political communication theory (Zompetti, 2015), we adopt the 14-technique propaganda taxonomy of Bassi et al. (2025) and apply it to German partisan discourse, annotating the divisive rhetorical devices present in each child comment.

### 2.3 LLM-in-the-Loop Annotation

LLMs have emerged as scalable annotators for political text, matching human performance on tasks from stance and frame detection (Gilardi et al., 2023) to propaganda labeling (Jose et al., 2026). However, relying on LLMs introduces systematic biases and domain-generalization limits, particu-

larly for lower-resource languages (Jadhav et al., 2025). This raises the question of when LLM annotations can responsibly substitute for human ones (Calderon et al., 2025). We therefore adopt a hybrid strategy: we quantify LLM-to-human agreement on a manually annotated 700-pair gold standard before using claude-opus-4-6 to scale annotations to the remaining 6,577 pairs.

## 2.4 Related Corpora

Prior corpora typically target one dimension at a time: constructiveness on Wikipedia talk pages (De Kock and Vlachos, 2021) or news comments (Kolhatkar et al., 2020), multilingual stance in voting-advice applications (Vamvas and Sennrich, 2020), or offensive language in German social media (Wiegand et al., 2018). Multi-label resources exist – GermEval 2021 scores toxicity, engagement, and fact-claiming (Risch et al., 2021), and Wang et al. (2023) fuse stance, sentiment, offensiveness, and sarcasm into one controversy score for Weibo – but neither labels the quality of an individual exchange. The closest corpus is that of Bassi et al. (2025), whose three-task schema we adopt; theirs covers English-language US immigration discourse, a scope its authors note is confined to one language and topic. Ours annotates German comment pairs across two policy domains and two parties’ channels.

## 3 Dataset

### 3.1 Channel Selection and Collection

We target two German parties at opposite ends of the parliamentary spectrum, the right-wing populist *Alternative für Deutschland* (AfD) and the left-wing *Die Linke*, and collect from each party’s official channel and its parliamentary-group (*Fraktion*) channel: @AfDTV and @AfDFraktionimBundestag for the AfD, and @DIELINKE and @dielinkebt for Die Linke. Using the YouTube Data API v3, we retrieved for every uploaded video its metadata, all comments, and, where available, the German transcript, over 06–13 March 2026. This yielded 14,280 videos (10,856 AfD; 3,424 Die Linke) and 4,638,722 comments; transcripts were available for 12,621 videos. A per-channel breakdown of the topic-relevant subset is given in Appendix A.1.

### 3.2 Pseudonymization

All author names were collected, sorted, and replaced with sequential pseudonyms (User\_00001,

User\_00002, ...); the same deterministic mapping was applied to @-mentions inside comment text, and erroneous double-@@ patterns introduced by YouTube’s formatting were normalized. Names were mapped separately for each party’s channels, so an author is identified by the (party, pseudonym) pair: 476,330 identities were pseudonymized, 386,213 on the AfD channels and 90,117 on the Die Linke channels. All subsequent processing and the public release use only the pseudonymized text.

### 3.3 Conversation Reconstruction

The YouTube API returns every reply as a direct child of the thread root, discarding the true addressee when users reply to one another deep in a thread. Following Bassi et al. (2024), we reconstruct the reply structure: processing each video’s comments chronologically, we maintain a global registry mapping every author to their most recent comment, and for each reply we take its last @-mention as the intended addressee, re-attaching the reply to that author’s latest comment (falling back to the raw parent when the mention cannot be resolved). Each comment thereby receives a resolved parent and a conversation depth. From every resolved link we form a parent–child pair, discarding *mismatched-target* pairs, in which the child’s @-mentions refer only to third parties, so that every retained pair is an unambiguous dyadic exchange. Bassi et al. (2025) manually validated this reconstruction procedure on English YouTube data, finding only 30 incorrect links among 2,387 child comments (98.7% accuracy).

### 3.4 Thematic Filtering via Topic Modeling

To isolate climate and migration discussions, we fit a BERTopic model (Grootendorst, 2022) on the 12,621 available transcripts, using paraphrase-multilingual-MiniLM-L12-v2 embeddings and class-based TF-IDF weighting, yielding 52 topics plus an outlier class. We manually mapped topics to two domains by their top keywords (Table 1); each video inherits its dominant topic, and videos assigned to neither domain were discarded.

A blind manual check of 75 videos found 75.0% of assigned videos substantially on-topic, with errors drawn mainly from adjacent policy domains; the off-topic remainder inflates the neutral-stance mass but contributed no opposing-stance pairs in that sample (Appendix A.2).

Table 1: Thematic categories and representative BERTopic keywords (translated from German).

| Category  | Topic IDs        | Representative Keywords                               |
|-----------|------------------|-------------------------------------------------------|
| Migration | 1, 9, 11, 33, 35 | <i>migrants, refugees, asylum, borders</i>            |
| Climate   | 3, 4, 36         | <i>CO2, climate, energy transition, nuclear power</i> |

### 3.5 Pair Extraction and Stratified Sampling

After topic filtering and pair extraction, the topic-relevant corpus comprises 1,143 videos and 385,320 threads (full per-channel statistics in Appendix A.1). Because the large majority of threads are single root comments with no replies, pair extraction yields 125,967 clean parent–child pairs, concentrated in just 39,861 threads ( $\approx 10\%$  of all threads), which form the pool eligible for sampling. This pool is highly imbalanced: 91.7% of eligible threads come from AfD channels, and Climate  $\times$  Die Linke contributes only 723 (1.8%). Annotating it in full would therefore yield a corpus dominated by a single stratum, confounding party and topic effects with channel volume, while the smallest stratum caps any balanced design at 723 threads per cell. To keep the party  $\times$  topic contrasts that motivate this resource directly comparable, we instead drew 2,000 threads with a fixed random seed, 500 per party–topic stratum, from this eligible pool (Table 2). This produced 7,299 pairs; removing 22 duplicate pairs (all in the Climate  $\times$  Linke stratum, arising from comments shared across reconstructed threads) leaves a final corpus of 7,277 unique, clean pairs.

Table 2: Stratified sampling design and yield of clean, unique pairs.

| Stratum                  | Threads | Comments | Clean Pairs |
|--------------------------|---------|----------|-------------|
| Climate $\times$ AfD     | 500     | 2,632    | 1,999       |
| Climate $\times$ Linke   | 500     | 2,204    | 1,420       |
| Migration $\times$ AfD   | 500     | 2,691    | 1,984       |
| Migration $\times$ Linke | 500     | 2,683    | 1,874       |
| Total                    | 2,000   | 10,210   | 7,277       |

## 4 Annotation

### 4.1 Annotation Schema

Following Bassi et al. (2025), every comment pair is annotated on three tasks. **Interaction quality** is a five-class ordinal label on the child’s contribution relative to its parent, jointly encoding communication style (constructive vs. destructive) and

relation (agreement vs. disagreement): Constructive Disagreement (+1.0), Constructive Agreement (+0.5), Neutral/Rephrase (0.0), Destructive Agreement (−0.5), and Destructive Disagreement (−1.0). **Stance** toward the video’s topic (Against, Neutral, Support) is annotated independently for the parent and the child comment. **Propaganda** is a 14-label multi-label set of divisive rhetorical techniques (Zompetti, 2015; Bassi et al., 2025), applied to the child comment. Table 3 gives the operational definitions of the five classes.

Table 3: Interaction Quality category definitions and score mappings.

| Category                         | Score | Definition and Linguistic Criteria                                                                                               |
|----------------------------------|-------|----------------------------------------------------------------------------------------------------------------------------------|
| <b>Constructive Disagreement</b> | +1.0  | Expresses disagreement while preserving conditions for mutual understanding; evaluates specific claims with reasons or examples. |
| <b>Constructive Agreement</b>    | +0.5  | Supports the parent while offering substantive reasons, data, or arguments that advance the dialogue.                            |
| <b>Neutral/Rephrase</b>          | 0.0   | Neutral entries serving conversational flow without taking sides, including information requests or topic-shifting.              |
| <b>Destructive Agreement</b>     | −0.5  | Agrees with the parent but demeans/invalidates opposing viewpoints to build exclusive in-group solidarity.                       |
| <b>Destructive Disagreement</b>  | −1.0  | Obstructs dialogue using hostile tone, personal attacks, or aggressive language; dehumanization triggers this score.             |

### 4.2 Annotation Guidelines

The guidelines adapt those of Bassi et al. (2025) to the climate and migration domains, adding supplementary decision rules for both the human and the LLM annotators. Interaction quality is judged along two axes: functional alignment (agreement, disagreement, neutral) and communication style (constructive, destructive). Stance is assessed against topic-specific policy targets, independently for parent and child. Propaganda flags the presence of 14 rhetorical devices in the child comment, with disambiguation rules for closely related techniques (e.g. Exaggeration/Minimisation vs. Loaded Language, or Repetition vs. parallel sentence structure). A condensed version of the guidelines and system prompts is provided in Appendix A.3; the complete prompts are released with the corpus.

### 4.3 Human Gold Standard

Two annotators independently labeled a stratified sample of 700 pairs (175 per stratum) through a custom local tool that presented each pair with full context (party, topic, thread, comment levels, and timestamps) and captured all three tasks simultaneously; pairs in which either comment was shorter than 20 characters, or which were flagged as mismatched-target, were filtered out beforehand. The 75 pairs (10.7%) on which the annotators agreed exactly across all three tasks entered the gold standard directly; the remaining 625 (89.3%) were adjudicated to a single label per task through discussion. Table 4 reports the resulting label distribution.

Table 4: Gold-standard label distribution ( $n = 700$ ).

| Task / Label                     | Gold Count | Gold % |
|----------------------------------|------------|--------|
| <b>Interaction Quality (IQ)</b>  |            |        |
| Constructive Disagreement (+1.0) | 155        | 22.1%  |
| Constructive Agreement (+0.5)    | 69         | 9.9%   |
| Neutral/Rephrase (0.0)           | 119        | 17.0%  |
| Destructive Agreement (−0.5)     | 160        | 22.9%  |
| Destructive Disagreement (−1.0)  | 197        | 28.1%  |
| <b>Child Stance</b>              |            |        |
| Support                          | 54         | 7.7%   |
| Neutral                          | 568        | 81.1%  |
| Against                          | 78         | 11.1%  |
| <b>Propaganda (14 classes)</b>   |            |        |
| Top: <i>Loaded Language</i>      | 332        | 47.4%  |

Prior to adjudication, exact agreement on the 5-class ordinal IQ task reached 56.9% (unweighted Cohen’s  $\kappa = 0.456$ , quadratic-weighted  $\kappa = 0.494$ ). Parent and child stance agreement reached 87.0% ( $\kappa = 0.666$ ) and 87.4% ( $\kappa = 0.640$ ). Multi-label propaganda detection yielded a mean Jaccard index of 0.335 (mean multi-label  $F_1 = 0.392$ ). The interaction-quality agreement exceeds the  $\kappa = 0.37$  that Bassi et al. (2025) report for the same schema on English YouTube comments; both figures reflect the difficulty of operationalizing constructive versus destructive communication in informal online discourse. See Appendix A.6 for expanded agreement matrices.

### 4.4 Representative Examples

To illustrate the 5-class interaction quality spectrum, Table 5 provides representative gold-standard comment pairs alongside their operationalized stance and rhetorical markers.

### 4.5 LLM-Assisted Annotation

To scale annotation beyond the gold standard, we label the remaining pairs with claude-opus-4-6 via the Anthropic API, using four independent task-specific calls per pair: a structured tool-use call for interaction quality, separate system-prompted calls for child and parent stance, and a free-text call for propaganda techniques validated against the 14-label vocabulary. Before scaling, we validated the pipeline on the 700-pair gold standard, measuring the LLM’s agreement with the human labels across all three tasks (Table 6).

The LLM achieved moderate agreement on interaction quality (quadratic-weighted  $\kappa = 0.560$ ; unweighted 0.550) and fair-to-moderate agreement on both stance layers: child stance ( $\kappa = 0.428$ ) and parent stance ( $\kappa = 0.435$ ). For child stance, the prevalent *Neutral* class was substantially easier to classify ( $F_1 = 0.791$ ) than the polarized *Against* ( $F_1 = 0.515$ ) or *Support* ( $F_1 = 0.537$ ) postures. Re-annotating a random 200-pair sample under identical settings reproduced 93.0% of interaction-quality labels (quadratic-weighted  $\kappa = 0.906$ ) and 96.0% of child stance labels ( $\kappa = 0.930$ ), so the moderate agreement with the gold standard reflects the subjectivity of the task rather than instability in the model.

Following validation, the pipeline was run on the remaining pairs, yielding a final corpus of 7,277 pairs (700 human gold standard, 6,577 LLM-annotated).

## 5 Thread-Level Score Aggregation and Corpus Analysis

We aggregated pair-level interaction quality scores to the thread level by calculating the mean InteractionScore for each (VideoID, ThreadID) cluster. Following Bassi et al. (2025), average thread scores were mapped into four quality bins (Table 7).

Breaking these thread-quality bins down by stratum (Table 8) exposes a pattern the aggregate hides: the AfD strata concentrate in the *Moderately Destructive* bin (54–60%) with comparatively few *Highly Destructive* threads, whereas the Die Linke strata are more polarized, carrying both more Constructive and more Highly Destructive threads. The mean thread scores nonetheless confirm that the AfD strata are the more destructive overall. Bin names and thresholds follow Bassi et al. (2025), who define them on a divisiveness scale, so they

Table 5: Representative comment-reply pairs illustrating the five-class interaction quality (IQ) schema (translated from German). Topic: M = migration, C = climate.

| Topic | IQ   | Stance             | Parent Comment                                                                                  | Child Comment                                                                           | Rhetorical Markers & Notes                                                                                                                                                                    |
|-------|------|--------------------|-------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| M     | -1.0 | P: Supp., C: Agst. | "I'm with the SPD and would take in refugees. Unfortunately rejected due to lack of space."     | "@User_10960 That's not a problem – you could just move out and sleep on the street..." | <b>Destructive Disagreement.</b> Uses sarcasm to mock the parent's welcome stance. ( <i>Loaded Language</i> ).                                                                                |
| M     | -0.5 | P: Agst., C: Agst. | "Bed, bread, and soap. Then they'll leave on their own, and we won't have to spend billions..." | "I'd be in favor of that too, but I'm worried they'd become even more criminal..."      | <b>Destructive Agreement.</b> Affirms parent but escalates threat with unevicenced outgroup stereotypes. ( <i>Appeal to Fear, Causal Oversimplification, Loaded Language, Name-Calling</i> ). |
| M     | 0.0  | P: Neut., C: Neut. | "Thanks for not giving in to the rightward shift, unlike the Greens."                           | "Please explain that."                                                                  | <b>Neutral/Rephrase.</b> Requests clarification without taking a clear side or shifting tone. ( <i>Doubt</i> ).                                                                               |
| C     | +0.5 | P: Supp., C: Supp. | "The 1.5°C target can no longer be held? 1.5 has been dead for a long time already."            | "And no longer reversible! 19 degrees in November, yet another sad new record."         | <b>Constructive Agreement.</b> Validates parent and introduces local weather anomaly data. ( <i>Appeal to Fear, Loaded Language</i> ).                                                        |
| C     | +1.0 | P: Agst., C: Supp. | "What qualifications does she actually have to make this claim? Is this backed up by data?"     | "Yes, it's in the IPCC report and many other studies on this..."                        | <b>Constructive Disagreement.</b> Challenges parent directly by referencing institutional consensus. ( <i>Appeal to Authority</i> ).                                                          |

Table 6: LLM alignment with the human gold standard ( $n = 700$ ). IQ: interaction quality; Prop.: propaganda.

| Task                      | Metric                                       | Score                 |
|---------------------------|----------------------------------------------|-----------------------|
| <b>IQ</b>                 | Quad-weighted $\kappa$ / Unweighted $\kappa$ | 0.560 / 0.550         |
|                           | Mean Absolute Error (MAE)                    | 0.366                 |
|                           | Pearson correlation ( $r$ )                  | 0.560                 |
|                           |                                              |                       |
| <b>Stance</b><br>(child)  | Unweighted $\kappa$ / Macro- $F_1$           | 0.428 / 0.615         |
|                           | Class $F_1$ (Against / Neutral / Support)    | 0.515 / 0.791 / 0.537 |
| <b>Stance</b><br>(parent) | Unweighted $\kappa$ / Macro- $F_1$           | 0.435 / 0.645         |
|                           | Class $F_1$ (Against / Neutral / Support)    | 0.544 / 0.761 / 0.631 |
| <b>Prop.</b>              | Macro- $F_1$ / Hamming Loss                  | 0.460 / 0.118         |

index the composite score rather than destructiveness alone. Because Destructive Agreement carries  $-0.5$ , a thread of purely destructive agreement cannot fall below that value: the Destructive-

Table 7: Thread-level quality bins and corpus distribution ( $n = 2,000$  threads).

| Mean Score       | Quality Bin        | Threads | %     |
|------------------|--------------------|---------|-------|
| $\geq +0.25$     | Constructive       | 412     | 20.6% |
| $[-0.25, +0.25)$ | Slight / Neutral   | 487     | 24.4% |
| $[-0.75, -0.25)$ | Mod. Destructive   | 829     | 41.5% |
| $< -0.75$        | Highly Destructive | 272     | 13.6% |

Agreement-dominated AfD strata therefore concentrate in the *Moderately Destructive* bin, while the Die Linke strata, richer in Destructive Disagreement, reach *Highly Destructive* more often despite fewer destructive pairs overall.

## 5.1 Stratum and Rhetorical Dynamics

By tracking interaction quality and stance across our stratified design, we observed stark contrasts between political channels (Table 9). The same distribution is visualized in Appendix Figure 1.

A structural divergence emerged along the agreement axis: discussions on AfD channels are dom-

Table 8: Thread-level quality-bin composition per stratum (% of the 500 threads in each stratum) and mean thread interaction score. The mean ranges from  $-1.0$  (every reply Destructive Disagreement, maximally divisive) through  $0$  (neutral) to  $+1.0$  (every reply Constructive Disagreement, maximally constructive).

| Stratum                  | Con. | Sl./N. | Mod.D. | Hi.D. | Mean   |
|--------------------------|------|--------|--------|-------|--------|
| Climate $\times$ AfD     | 12.8 | 23.6   | 54.0   | 9.6   | -0.283 |
| Climate $\times$ Linke   | 31.6 | 26.6   | 26.0   | 15.8  | -0.077 |
| Migration $\times$ AfD   | 10.6 | 20.0   | 59.8   | 9.6   | -0.325 |
| Migration $\times$ Linke | 27.4 | 27.2   | 26.0   | 19.4  | -0.129 |

inated by agreement (50.3% for climate, 58.0% for migration), while comments on Die Linke are primarily disagreeing (54.2% for climate, 62.5% for migration). This suggests that AfD comment sections are spaces where users predominantly reinforce one another’s positions, whereas Die Linke channels host more contestation.

The two collapses (Table 10) show that party shapes both axes, and that agreement does not imply civility. On the *relation* axis, AfD channels skew toward agreement and Die Linke toward disagreement irrespective of topic; on the *communication style* axis, destructiveness peaks in the AfD strata (72.1% for Migration  $\times$  AfD) while the Die Linke strata are the most constructive (38.2% for Climate  $\times$  Linke). Migration  $\times$  AfD combines the highest agreement (58.0%) with the highest destructiveness, driven by Destructive Agreement – 54.5% of its pairs – in which users reinforce hostile shared positions rather than engage opposing ones. This tone is mirrored in the prevalence of specific propaganda techniques (Table 11).

Table 11 breaks technique prevalence down by stratum: emotive devices dominate everywhere (Loaded Language appears in roughly half of all pairs), while fear appeals and flag-waving concentrate in Migration  $\times$  AfD (17.4% and 11.1%), the stratum that also shows the highest Destructive Agreement.

Child-stance distributions mirror the party divide (Table 9). Neutral is the most common label in three of the four strata, from 51.9% in Migration  $\times$  Linke to 72.8% in Climate  $\times$  Linke, reflecting that many comments beneath a topic-labeled video take no explicit policy position of their own. The exception is Migration  $\times$  AfD, where Against-stance is by far most common: 56.8% of children oppose immigration (Neutral falls to 37.6%), whereas Support concentrates in the Die Linke

strata (16.5% for climate and 22.1% for migration, versus 7.1% and 5.7% on the AfD channels). Stance therefore tracks channel ideology closely, while the large neutral mass underscores that topic membership is defined at the video rather than the comment level.

**Climate: cross-cutting dialogue versus same-stance reinforcement.** Grouping pairs by the *stance relation* between parent and child reveals a pattern specific to climate (Table 12). Only on climate are *opposing*-stance exchanges – a skeptic engaging a supporter – constructive on average (mean  $+0.08$ , 53% constructive;  $n = 340$ ), the single positive relation in either topic; the analogous migration clash stays destructive (mean  $-0.07$ ), and the two differ significantly (Welch  $t = 2.28$ ,  $p = 0.02$ ). This holds within both channels: opposing-stance climate exchanges are constructive on average on the AfD ( $+0.04$ ) and Die Linke ( $+0.12$ ) channels alike (Appendix A.5). Divisiveness instead concentrates in *same-stance* exchanges, but unevenly across them (Table 12). The destructive engine is *hostile agreement between opponents*: pairs in which both sides are Against – skeptic replying to skeptic on climate, anti-migration to anti-migration – are the single most destructive configuration we report (mean  $-0.39$  on climate,  $-0.41$  on migration; roughly 88% destructive). Neutral-to-neutral pairs are only mildly destructive ( $-0.17 / -0.11$ ) but by far the most frequent, so their volume masks how hostile the same-stance relation otherwise is: excluding them, the same-stance mean falls to  $-0.29$  on climate and  $-0.38$  on migration. Support-to-support pairs, a sub-case of same stance rather than a relation in its own right, are by contrast mildly *constructive* ( $+0.13 / +0.03$ ) though rare: it is opponents reinforcing hostility, not allies agreeing, that produces divisiveness. This destructiveness is concentrated on the AfD side (same-stance mean  $-0.31$  vs.  $-0.03$  for climate), and not merely because AfD channels host more Against–Against exchanges: both Against–Against and Neutral–Neutral pairs are harsher there, the latter averaging  $-0.28$  against  $-0.05$  on Die Linke. Climate is thus the domain where cross-cutting dialogue still functions; on migration, the same skeptic–supporter encounter turns destructive. This tracks the two topics’ rhetorical signatures: climate discourse leans on *Doubt* (13.6% of pairs vs. 9.7% for migration) more than *Appeal to Fear* (9.6% vs. 13.8%), and climate skepticism is concentrated on the AfD chan-

Table 9: Pair-level interaction quality and child stance distributions across strata ( $n = 7,277$  unique pairs). IQ labels: CD = Constructive Disagreement (+1.0), CA = Constructive Agreement (+0.5), N = Neutral (0.0), DA = Destructive Agreement (−0.5), DD = Destructive Disagreement (−1.0).

| Stratum                  | $n$   | Interaction Quality (5-Class) |     |     |       |     | Child Stance |         |         |
|--------------------------|-------|-------------------------------|-----|-----|-------|-----|--------------|---------|---------|
|                          |       | CD                            | CA  | N   | DA    | DD  | Against      | Neutral | Support |
| Climate $\times$ AfD     | 1,999 | 301                           | 89  | 257 | 916   | 436 | 596          | 1,262   | 141     |
| Climate $\times$ Linke   | 1,420 | 387                           | 156 | 244 | 251   | 382 | 152          | 1,034   | 234     |
| Migration $\times$ AfD   | 1,984 | 244                           | 69  | 241 | 1,081 | 349 | 1,126        | 745     | 113     |
| Migration $\times$ Linke | 1,874 | 552                           | 139 | 249 | 315   | 619 | 487          | 973     | 414     |

Table 10: Two aggregations of the five-class interaction labels, by stratum (row % of pairs). *Relation*: agreement vs. disagreement; *Style* (communication style): constructive vs. destructive. Neut. (Neutral/Rephrase) is shared by both aggregations. Strata: C=Climate, M=Migration; A=AfD, L=Die Linke.

| Stratum      | Relation |        | Style   |        | Neut. |
|--------------|----------|--------|---------|--------|-------|
|              | Agree    | Disag. | Constr. | Destr. |       |
| C $\times$ A | 50.3     | 36.9   | 19.5    | 67.6   | 12.9  |
| C $\times$ L | 28.7     | 54.2   | 38.2    | 44.6   | 17.2  |
| M $\times$ A | 58.0     | 29.9   | 15.8    | 72.1   | 12.1  |
| M $\times$ L | 24.2     | 62.5   | 36.9    | 49.8   | 13.3  |

nels, where 29.8% of comments oppose climate action versus 10.7% on Die Linke.

## 5.2 Temporal and Thematic Trends

Analyzing thread quality by topic from 2019 onward (Table 13), climate threads carry a higher constructive share than migration threads in most years, though migration briefly overtakes climate in 2022–2023; both topics drop sharply in 2023 and recover by 2024, after which the gap narrows to within three points by 2026. A full breakdown by party  $\times$  topic stratum and year is given in Appendix A.4 (Table 19). Despite these fluctuations, the overall thread-quality distributions do not differ significantly between the two topics ( $\chi^2(3) = 5.15$ ,  $p = 0.16$ ). Because the sample is not balanced by party within years (it is balanced by party overall), these year-on-year movements describe the sampled threads rather than a development in the discourse itself (Section 8).

## 6 Corpus Format and Availability

The corpus is released as JSON Lines, one record per parent–child pair. Each record carries the pair identifier; the pseudonymized parent and child text with their author identifiers, conversation levels, and timestamps; the source video and thread identifiers; the party and topic stratum; and the three

annotation layers – the interaction-quality score and label, parent and child stance, and the set of propaganda techniques on the child comment. The human gold standard and the LLM-annotated remainder are released as separate files, so that the 700-pair gold subset can be used on its own. The full dataset, annotation guidelines, system prompts, and the construction and annotation code are publicly released.<sup>1</sup>

## 7 Conclusion

We presented a new German-language resource for studying divisiveness in political social-media discourse: a corpus of 7,277 parent–child comment pairs sampled from AfD and Die Linke YouTube channels on climate and migration, annotated for interaction quality, stance, and propaganda techniques. We described the full construction pipeline – channel selection, comment and transcript collection, pseudonymization, conversation reconstruction, topic filtering, and stratified sampling – and a reproducible three-task annotation protocol validated against a 700-pair human gold standard. Descriptive analysis reveals pronounced party- and topic-specific patterns, most notably the predominance of Destructive Agreement in the Migration  $\times$  AfD stratum. Zooming in on climate, we find that cross-cutting skeptic–supporter exchanges are the only stance relation that stays constructive on average, on both parties’ channels, while same-stance exchanges carry most of the divisiveness.

## 8 Limitations

First, 90.4% of the corpus is annotated with claude-opus-4-6 rather than by humans and its agreement with the human gold standard is moderate ( $\kappa = 0.550$ ). The human–human baseline is itself moderate ( $\kappa = 0.456$ ) and exceeds the

<sup>1</sup>[https://github.com/zaheralkaei/climate\\_migration\\_german\\_political\\_yt](https://github.com/zaheralkaei/climate_migration_german_political_yt)

Table 11: Propaganda technique prevalence per stratum (% of pairs); the highest value in each row is shown in bold.

| Technique                           | Climate $\times$ AfD | Climate $\times$ Linke | Migration $\times$ AfD | Migration $\times$ Linke |
|-------------------------------------|----------------------|------------------------|------------------------|--------------------------|
| Appeal_to_Authority                 | 3.2%                 | <b>4.7%</b>            | 2.3%                   | 2.6%                     |
| Appeal_to_Time                      | 4.4%                 | 2.9%                   | <b>6.1%</b>            | 2.7%                     |
| Appeal_to_fear-prejudice            | 11.8%                | 6.5%                   | <b>17.4%</b>           | 10.0%                    |
| Bandwagon,Reductio_ad_hitlerum      | 1.8%                 | 2.5%                   | 2.1%                   | <b>2.9%</b>              |
| Black-and-White_Fallacy             | 3.7%                 | 4.7%                   | 5.0%                   | <b>5.4%</b>              |
| Causal_Oversimplification           | 9.5%                 | 13.0%                  | 11.3%                  | <b>13.8%</b>             |
| Doubt                               | <b>13.6%</b>         | 13.5%                  | 8.6%                   | 10.8%                    |
| Exaggeration,Minimisation           | 25.2%                | 25.0%                  | 26.3%                  | <b>27.6%</b>             |
| Flag-Waving                         | 7.7%                 | 3.9%                   | <b>11.1%</b>           | 5.0%                     |
| Loaded_Language                     | 51.9%                | 44.2%                  | <b>53.6%</b>           | 48.2%                    |
| Name_Calling,Labeling               | <b>38.8%</b>         | 30.0%                  | 34.2%                  | 35.3%                    |
| Repetition                          | 3.3%                 | <b>3.9%</b>            | 3.3%                   | 2.2%                     |
| Slogans/Thought-terminating_Cliches | 9.6%                 | 10.0%                  | <b>10.2%</b>           | 9.1%                     |
| Whataboutism,Straw_Men              | 2.7%                 | 2.3%                   | 2.1%                   | <b>5.0%</b>              |

Table 12: Interaction quality by parent-child stance relation, aggregated across both parties, with the same-stance relation decomposed by shared stance. Mean interaction score (range  $-1$  to  $+1$ ) and pair count  $n$ . Opposing (skeptical vs. supporter) is the only relation constructive on average, and only on climate; hostile Against-Against agreement is the most destructive of the configurations reported here.

| Stance relation  | Climate |       | Migration |       |
|------------------|---------|-------|-----------|-------|
|                  | Mean    | $n$   | Mean      | $n$   |
| Opposing         | +0.08   | 340   | -0.07     | 583   |
| Same stance      | -0.19   | 2,276 | -0.24     | 2,377 |
| Against-Against  | -0.39   | 401   | -0.41     | 1,028 |
| Neutral-Neutral  | -0.17   | 1,780 | -0.11     | 1,263 |
| Support-Support  | +0.13   | 95    | +0.03     | 86    |
| One side neutral | -0.22   | 803   | -0.18     | 898   |

Table 13: Constructive-thread share (% of threads in the Constructive bin) by year and topic (year of video publication), 2019–2026; thread counts  $n$  in parentheses. Years before 2019 contain too few threads ( $< 15$ ) to be informative.

| Year | Climate ( $n$ ) | Migration ( $n$ ) |
|------|-----------------|-------------------|
| 2019 | 28.9 (45)       | 7.0 (43)          |
| 2020 | 11.8 (68)       | 6.8 (73)          |
| 2021 | 23.1 (104)      | 12.8 (47)         |
| 2022 | 11.9 (151)      | 14.3 (56)         |
| 2023 | 7.0 (129)       | 8.9 (123)         |
| 2024 | 34.2 (73)       | 19.3 (228)        |
| 2025 | 28.5 (361)      | 27.0 (352)        |
| 2026 | 25.0 (60)       | 24.6 (65)         |

$\kappa = 0.37$  reported for the same schema by Bassi et al. (2025); the model is highly self-consistent under re-annotation (93.0% exact on interaction quality, quadratic-weighted  $\kappa = 0.906$ ), so little of the residual gap is model noise; and the two annotators differ mainly by a uniform strictness offset rather than a content-linked bias (Appendix A.6). The

remaining disagreement therefore reflects the inherent subjectivity of interaction-quality annotation rather than instability in the pipeline. Fine-grained comparisons between individual classes should be read with that subjectivity in mind.

Second, topic filtering operates at the video level, and the blind check in Section 3.4 shows that even video-level assignment is imperfect; individual comments may drift further still. This labeling is therefore a rough filter for locating discussions related to climate and migration, not a precise topic label for each individual comment.

Third, the stratified sampling design imposes an artificial balance across parties and topics that does not reflect the natural class distribution of the source channels; we release the full eligible-thread pool so the pipeline can be scaled to the unbalanced corpus where representativeness matters more than cross-stratum comparability.

Fourth, aggregating the ordinal interaction-quality classes to the thread level requires numeric values. Following Bassi et al. (2025) we use  $(+1.0, +0.5, 0, -0.5, -1.0)$ , the equal-interval coding of the class ranks rescaled to  $[-1, +1]$ , so the larger magnitude of the disagreement classes follows from the class ordering rather than a separate weighting choice; the bin thresholds are likewise inherited. Our central result is robust to this: recoding by communication style alone (constructive =  $+1$ , neutral =  $0$ , destructive =  $-1$ ), which discards the agreement axis, still leaves opposing-stance climate exchanges the only relation with a positive mean. We release the categorical labels alongside the scores, so alternative aggregations can be derived.

Fifth, the sample is balanced across party and

topic overall but not within years, so its composition shifts over the period: pairs from 2020 are 99% AfD, whereas pairs from 2025 are 83% Die Linke. Corpus-wide year-on-year movements in destructiveness, agreement, and technique use largely track this shifting composition rather than a change in discourse. Some within-party series do move – the constructive share of Migration  $\times$  AfD threads rises from 4.9% in 2019 to 21.0% in 2025 – but within a party the sampled videos also change from year to year, so the temporal breakdowns in Table 13 and Appendix A.4 are descriptive summaries of the sample rather than evidence of development over time; establishing genuine trends would require a temporally stratified sample.

## 9 Ethical Considerations

**Data and privacy.** The corpus is built exclusively from publicly posted YouTube comments, collected through the official Data API v3 in accordance with its terms of service. Prior to any annotation or release, all author names were replaced with sequential pseudonyms and @-mentions rewritten with the same deterministic mapping, so that no usernames or account handles remain in the released text; we make no attempt to re-identify authors and release no author profile metadata.

**Intended use and risks.** The corpus contains, in parts, hostile, offensive, and propagandistic German-language political speech, and is released for research on online discourse, content moderation, and computational social science. It is not intended as training data for generating such content, and the propaganda-technique labels characterize rhetorical form, not the truth or endorsement of any claim.

**Annotation.** The gold standard was produced by two trained annotators familiar with German political discourse; no crowd workers were employed. Annotators saw each pair together with its thread context, including the source channel, because resolving sarcasm, deictic reference, and the target of a stance claim requires that context. Annotators were therefore not blind to party, which carries a risk of confirmation bias, with comments from a disfavored party read as more destructive by default. Three features of the design mitigate this. The guidelines instruct annotators to judge communication style rather than ideological alignment (Appendix A.3.1); 89.3% of pairs were reconciled

through joint adjudication, which surfaces systematic individual tendencies; and the LLM annotations that cover the remaining 90.4% of the corpus are party-blind, since each prompt receives only the parent and child comment text, with no channel, party, or video metadata. Party information is thus visible only for the 9.6% human-annotated subset. To test for a party-linked annotation effect directly, we compared the human gold labels against the party-blind LLM labels on the same pairs. Had annotators systematically read AfD comments as more destructive, the human scores would fall below the LLM scores on the AfD strata specifically. They do not: the mean human-minus-LLM interaction score is +0.013 on the AfD strata and −0.050 on the Die Linke strata (difference 0.063 on a  $[-1, +1]$  scale; Welch  $t = 1.19$ ,  $p = 0.24$ ), so the small observed difference runs opposite to the hypothesized direction and is not statistically significant. The party-blind LLM in fact labels a larger share of AfD pairs destructive (67.4%) than the human annotators do (58.3%). This is a consistency check rather than proof of absence, since the LLM is not a neutral reference and party affiliation is sometimes inferable from the comment text itself.

**Availability.** The annotations are released under CC BY-NC 4.0 and the pipeline code under the MIT license; copyright in the underlying comment text remains with its authors. A contact point for removal requests is provided with the release.

## Acknowledgments

We thank the anonymous reviewers for their careful feedback and constructive suggestions, which substantially improved this paper. We also thank Annika Blietz for her valuable contribution as the second annotator for the gold-standard dataset.

## References

- Dana Siobhan Atzpodien. 2022. [Party competition in migration debates: The influence of the AfD on party positions in German state parliaments](#). *German Politics*, 31(3):381–398.
- Davide Bassi, Michele Joshua Maggini, Renata Vieira, and Martín Pereira-Fariña. 2024. [A pipeline for the analysis of user interactions in YouTube comments: A hybridization of LLMs and rule-based methods](#). In *2024 11th International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 146–153.

- Davide Bassi, Erik Bran Marino, Renata Vieira, and Martin Pereira. 2025. [Old but gold: LLM-based features and shallow learning methods for fine-grained controversy analysis in YouTube comments](#). In *Proceedings of the 12th Argument mining Workshop*, pages 46–57, Vienna, Austria. Association for Computational Linguistics.
- Burint Bevis, Juliana Schroeder, and Michael Yeomans. 2026. [Spoken disagreement is more constructive than written disagreement](#). *Nature Communications*, 17(1):5792.
- Nitay Calderon, Roi Reichart, and Rotem Dror. 2025. [The alternative annotator test for LLM-as-a-judge: How to statistically justify replacing human annotators with LLMs](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16051–16081, Vienna, Austria. Association for Computational Linguistics.
- Kai Chen, Zihao He, Rong-Ching Chang, Jonathan May, and Kristina Lerman. 2023. [Anger breeds controversy: analyzing controversy and emotions on Reddit](#). In *International conference on social computing, behavioral-cultural modeling and prediction and behavior representation in modeling and simulation*, pages 44–53. Springer.
- Hilde Coffé, Sam Crawley, and Josh Givens. 2026. [Growing polarisation: ideology and attitudes towards climate change](#). *West European Politics*, 49(1):1–29.
- Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. 2011. [Political polarization on Twitter](#). In *Proceedings of the international aaai conference on web and social media*, volume 5, pages 89–96.
- Christine De Kock and Andreas Vlachos. 2021. [I beg to differ: A study of constructive disagreement in online conversations](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2017–2027, Online. Association for Computational Linguistics.
- Paul Drecker. 2025. [Dynamics of electoral polarisation in climate policy discourse: A temporal network analysis](#). *Politics and Governance*, 13.
- Neele Falk. 2025. [What makes discussions constructive?: modeling argument and deliberative quality](#). Ph.D. thesis, Universität Stuttgart, Stuttgart.
- Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2018. [Quantifying controversy on social media](#). *Trans. Soc. Comput.*, 1(1).
- Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. [Chatgpt outperforms crowd workers for text-annotation tasks](#). *Proceedings of the National Academy of Sciences*, 120(30).
- Maarten Grootendorst. 2022. [Bertopic: Neural topic modeling with a class-based tf-idf procedure](#). *Preprint*, arXiv:2203.05794.
- Sally Jackson. 1992. “virtual standpoints” and the pragmatics of conversational argument. In Frans H. van Eemeren, Rob Grootendorst, J. Anthony Blair, and Charles A. Willard, editors, *Argumentation Illuminated*, pages 260–269. SICSAT, Amsterdam.
- Suramya Jadhav, Abhay Shanbhag, Amogh Thakurdesai, Ridhima Sinare, and Raviraj Joshi. 2025. [On limitations of LLM as annotator for low resource languages](#). *Preprint*, arXiv:2411.17637.
- Julia Jakob, Timo Dobbrick, Rainer Freudenthaler, Patrik Haffner, and Hartmut Wessler. 2023. [Is constructive engagement online a lost cause? toxic outrage in online user comments across democratic political systems and discussion arenas](#). *Communication Research*, 50(4):508–531.
- Julia Jose, Meghna Manoj Nair, and Rachel Greenstadt. 2026. [Large-scale analysis of persuasive content on moltbook](#). *Preprint*, arXiv:2603.18349.
- Varada Kolhatkar, Nithum Thain, Jeffrey Sorensen, Lucas Dixon, and Maite Taboada. 2020. [Classifying constructive comments](#). *Preprint*, arXiv:2004.05476.
- Eugenio Levi, Michael Bayerlein, Gianluca Grimalda, and Tommaso Reggiani. 2025. [Narratives of migration and political polarization: Private preferences, public preferences and social media](#). Working paper.
- Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. [Semeval-2016 task 6: Detecting stance in tweets](#). In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics.
- Adrian Rauchfleisch and Jonas Kaiser. 2020. [The German far-right on YouTube: An analysis of user overlap and user comments](#). *Journal of Broadcasting & Electronic Media*, 64(3):373–396.
- Manoel Horta Ribeiro, Raphael Ottoni, Robert West, Virgílio A. F. Almeida, and Wagner Meira. 2020. [Auditing radicalization pathways on YouTube](#). In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAT\* ’20*, page 131–141, New York, NY, USA. Association for Computing Machinery.
- Julian Risch, Anke Stoll, Lena Wilms, and Michael Wiegand. 2021. [Overview of the GermEval 2021 shared task on the identification of toxic, engaging, and fact-claiming comments](#). In *Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments*, pages 1–12, Duesseldorf, Germany. Association for Computational Linguistics.

Jennifer Schumann and Steve Oswald. 2024. [Pragmatic perspectives on disagreement](#). *Journal of Language Aggression and Conflict*, 12(1):1–16.

Ozgur Can Seckin, Bao Tran Truong, Alessandro Flammini, and Filippo Menczer. 2026. [Identifying constructive conflict in online discussions through controversial yet toxicity resilient posts](#). In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 20, pages 2079–2093.

Anke Stoll, Marc Ziegele, and Oliver Quiring. 2020. [Detecting impoliteness and incivility in online discussions: Classification approaches for German user comments](#). *Computational Communication Research*, 2(1):109–134.

Jannis Vamvas and Rico Sennrich. 2020. [X-stance: A multilingual multi-target dataset for stance detection](#). *Preprint*, arXiv:2003.08385.

Haiyang Wang, Ye Wang, Xin Song, Bin Zhou, Xuechen Zhao, and Feng Xie. 2023. [Quantifying controversy from stance, sentiment, offensiveness and sarcasm: a fine-grained controversy intensity measurement framework on a Chinese dataset](#). *World Wide Web*, 26(5):3607–3632.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. [Overview of the GermEval 2018 shared task on the identification of offensive language](#). In *Proceedings of the GermEval 2018 Workshop on the Identification of Offensive Language*, pages 1–10, Vienna, Austria. Austrian Academy of Sciences Press.

Julia Zilles and Stine Marg. 2023. [Protest and polarisation in the context of energy transition and climate policy in Germany: Mindsets and collective identities](#). *German Politics*, 32(3):495–516.

Joseph P. Zompetti. 2015. *Divisive Discourse: The Extreme Rhetoric of Contemporary American Politics*. Cognella Academic Publishing, San Diego, CA.

## A Supplemental Material: Corpus, Guidelines, and Validation

This appendix provides an overview of the German partisan corpus composition, a condensed version of the system-prompt instructions and representative gold-standard examples, and the validation matrices for our inter-annotator agreement (IAA) trials.

### A.1 Corpus Demographics and Sampling Coverage

The per-channel composition of the topic-relevant corpus is presented in Table 14.

To ensure balanced representation across the discourse zones, a stratified sampling strategy was applied to generate the final 7,277-pair corpus. Table 15 details the final annotation coverage across

each of the four target party-topic strata, partitioning them into human-adjudicated gold standard pairs and automated LLM-annotated pairs.

Table 16 lists the most frequent propaganda techniques across the corpus as a whole; the per-stratum breakdown for all 14 techniques is given in Table 11.

### A.2 Topic-Assignment Validation

One author re-judged a random sample of 75 videos (30 climate, 30 migration, 15 other), drawn with a fixed seed, while blind to both the assigned category and the BERTopic keywords and seeing only the video title, channel, and publication date. Each video was marked as substantially about climate, about migration, about both, or about neither.

Of the 60 assigned videos, 75.0% were judged substantially on-topic (70.0% for climate, 80.0% for migration), and one of the 15 other videos was judged to contain climate or migration content. The errors are systematic rather than random: climate false positives concentrate in adjacent transport and energy debates (rail policy, fuel and gas prices), and migration false positives in general party-political content.

Because stance is annotated toward the domain, comments under an off-topic video cannot take a domain stance and are labeled Neutral. Videos assigned to a domain but judged off-topic carry 80.2% Neutral child stance against 55.0% for on-topic videos, and contributed none of the 27 opposing-stance pairs the sample produced. Topic-assignment error therefore inflates the neutral mass reported in Section 5.1 rather than contaminating the stance-relation analysis.

### A.3 Annotation Guidelines and System Prompts

This section documents the formal system instructions provided to our annotators and prompt pipelines.

#### A.3.1 Interaction Quality Categories and Coding Rules

The annotator classifies how a child reply relates to its parent, using the five interaction-quality categories defined in Table 3 (Section 4.1).

To ensure coding stability, annotators must adhere to six operational rules. First, annotators must classify the interaction relationship between the parent and child comments rather than evaluating comments in isolation. Second, the evaluation must

Table 14: Comprehensive corpus composition by political channel and topic.

| Topic                        | Party | Channel                | Videos | Comments | Threads | Pairs   |
|------------------------------|-------|------------------------|--------|----------|---------|---------|
| Climate                      | AfD   | AfD TV                 | 41     | 12,744   | 8,416   | 3,837   |
| Climate                      | AfD   | AfD-Fraktion Bundestag | 345    | 108,034  | 79,164  | 26,279  |
| Climate                      | Linke | Die Linke im Bundestag | 17     | 1,975    | 1,045   | 777     |
| Climate                      | Linke | DieLinke               | 41     | 3,162    | 1,767   | 1,179   |
| Migration                    | AfD   | AfD TV                 | 168    | 76,659   | 52,931  | 21,062  |
| Migration                    | AfD   | AfD-Fraktion Bundestag | 442    | 303,859  | 230,955 | 65,248  |
| Migration                    | Linke | Die Linke im Bundestag | 25     | 3,171    | 2,015   | 939     |
| Migration                    | Linke | DieLinke               | 64     | 16,655   | 9,027   | 6,646   |
| Climate × AfD (subtotal)     |       |                        | 386    | 120,778  | 87,580  | 30,116  |
| Climate × Linke (subtotal)   |       |                        | 58     | 5,137    | 2,812   | 1,956   |
| Migration × AfD (subtotal)   |       |                        | 610    | 380,518  | 283,886 | 86,310  |
| Migration × Linke (subtotal) |       |                        | 89     | 19,826   | 11,042  | 7,585   |
| Total (topic-relevant)       |       |                        | 1,143  | 526,259  | 385,320 | 125,967 |



Figure 1: Distribution of interaction quality labels by party and topic.

Table 15: Annotation coverage and gold standard distribution by stratum.

| Stratum           | Pairs | Gold | LLM   |
|-------------------|-------|------|-------|
| Climate × AfD     | 1,999 | 175  | 1,824 |
| Climate × Linke   | 1,420 | 175  | 1,245 |
| Migration × AfD   | 1,984 | 175  | 1,809 |
| Migration × Linke | 1,874 | 175  | 1,699 |
| Total Corpus      | 7,277 | 700  | 6,577 |

Note: Pairs = Gold + LLM per stratum. Gold = human-annotated (9.6% of the corpus); LLM = claude-opus-4-6-annotated (90.4%). Every pair carries labels on all three tasks.

Table 16: Most frequent propaganda techniques across all pairs ( $n = 7,277$ ).

| Rank | Propaganda technique        | Tech- Core Function         | Count |
|------|-----------------------------|-----------------------------|-------|
| 1    | Loaded Language             | Emotive, biased phrasing    | 3,632 |
| 2    | Name Calling / Labeling     | Stereotyping and ad hominem | 2,542 |
| 3    | Exaggeration / Minimisation | Distorting fact severity    | 1,897 |
| 4    | Appeal to Fear / Prejudice  | Invoking collective anxiety | 862   |
| 5    | Causal Oversimplification   | Reducing causal complexity  | 856   |
| 6    | Doubt                       | Questioning credibility     | 837   |
| 7    | Slogans                     | Thought-terminating clichés | 707   |
| 8    | Flag-Waving                 | Group/nationalist loyalty   | 524   |

focus strictly on communication style rather than ideological alignment. Third, any use of dehumanizing language automatically yields a score of  $-1.0$ . Fourth, sarcasm and irony must be evaluated by intent: playful sarcasm is classified by its underlying argument, whereas contemptuous or mocking sarcasm is treated as Destructive. Fifth, the guidelines prioritize substance over politeness, meaning that blunt or harsh arguments that address claims are marked as Constructive, while polite dismissals lacking reasoning are classified as Destructive. Sixth, *reductio ad absurdum* (extending arguments to extreme conclusions to expose logical flaws) counts as a valid argument and is classified as Constructive. All output is required as a single-line JSON structure matching the schema: `{"score": <float>, "label": <string>}`.

### A.3.2 Stance Classifications

The stance detection pipeline is divided into issue-specific prompts for Climate Action and Migration. Table 17 consolidates the stance boundaries, numerical labels, and core criteria across both domains, serving as the system definitions for both child and parent stance tasks.

Table 17: Stance category classification definitions across domains.

| Class                 | Label | Operational Criteria                                                                                                 |
|-----------------------|-------|----------------------------------------------------------------------------------------------------------------------|
| <b>Climate Action</b> |       |                                                                                                                      |
| Against / Skeptic     | 0     | Denies/downplays climate change, opposes renewable transition, mocks science, or critiques policy as destructive.    |
| Neutral               | 1     | Factual claims without clear positioning, neutral questions, or ambiguous/sarcastic text with unclear policy stance. |
| Support               | 2     | Advocates climate action, renewables, or environmental protection; challenges skeptics.                              |
| <b>Migration</b>      |       |                                                                                                                      |
| Against               | 0     | Opposes immigration, portrays integration negatively, criticizes migrants, or demands assimilation.                  |
| Neutral               | 1     | Factual, ambiguous, or irrelevant comments lacking clear stance; anger/sarcasm without policy stance.                |
| Support               | 2     | Defends immigrant rights, highlights positive integration, or criticizes border restrictions.                        |

Parent context must be utilized to resolve ambiguous, short, or ironic child statements. Annotators are instructed to watch for explicit policy

markers, such as references to *CO<sub>2</sub>-Steuer* (CO<sub>2</sub> tax) or *Energiewende* (energy transition). Framing climate policy as an ideology or religion is treated as a typical signal of a skeptic stance (0). Crucially, a Neutral (1) classification indicates that the stance is unidentifiable, rather than implying that the comment is emotionally neutral. Stance prompts output a JSON format: `{"label": <int>}`.

### A.3.3 Propaganda Techniques

For multi-label propaganda classification, the pipeline identifies the occurrence of 14 distinct rhetorical and manipulative devices within the comments. Table 18 maps each target propaganda technique to its operational definition, forcing an output schema of: `{"techniques": [<list of strings>]}` or an empty array `[]` if no techniques are detected.

### A.3.4 Illustrative Gold-Standard Examples

We show one representative gold-standard example per task, illustrating how the guidelines apply. The prompts are instruction-only: no examples were included in the model input.

#### Interaction Quality (Destructive Disagreement).

*Parent*: “Auf 26 kinder ist meine Tochter 1 von 4 Deutschen. Wenn ich sie von der Schule abhole, stehen da 8 Kopftücher und 2 normale. KOTZT MICH AN” (“*Out of 26 children, my daughter is 1 of 4 Germans. When I pick her up from school there are 8 headscarves and 2 normal ones. IT DISGUSTS ME*”)

*Child*: “Da freut sich Frau Weidel!” (“*Ms. Weidel will be pleased!*”)

*Answer*: `{"score": -1.0, "label": "Destructive Disagreement"}`

#### Climate Stance (Against/Skeptic). *Parent*:

“Herr Dr. Spaniel... [text condensed] System runter gefahren wegen Stromausfall!” (“*Mr. Dr. Spaniel... [text condensed] system shut down due to a power outage!*”)

*Child*: “Nicht zu vergessen Reichweite ich fahre mit unserem Airsoft Team 400 km...” (“*Not to forget range, I drive 400 km with our Airsoft team...*”)

*Answer*: `{"label": 0}`

#### Propaganda (Loaded Language & Labeling).

*Comment*: “Scheiß auf Integration.” (“*Screw integration.*”)

*Answer*: `{"techniques":`

Table 18: Definitions and criteria of the 14 multi-label propaganda techniques.

| Technique                                                      | Definition and Criteria                                                                                                                                                                  |
|----------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Loaded_Language                                                | Words or phrases with strong emotional connotations meant to influence feelings directly.                                                                                                |
| Name_Calling, Labeling<br>Repetition                           | Assigning labeling terms to discredit or praise without substantive argument. Deliberately repeating core claims or slogans within a single comment (excluding grammatical parallelism). |
| Exaggeration, Minimisation                                     | Distorting factual reality via absolute quantifiers ( <i>never, always, all</i> ) when evidence is mixed.                                                                                |
| Appeal_to_fear-prejudice                                       | Creating anxiety about potential negative consequences to win support for a position.                                                                                                    |
| Flag-Waving<br>Causal_Oversimplification                       | Exploiting group identity (national, political, religious) to validate a claim. Reducing complex, multi-causal socio-political issues to a single, simplistic cause.                     |
| Appeal_to_Authority                                            | Citing authority or expertise as sole validation without providing substantive arguments.                                                                                                |
| Slogans/Thought-terminating_Cliches<br>Whataboutism, Straw_Men | Using stock, simplistic catchphrases to discourage critical thought or end debate. Deflecting criticism by alleging opponent hypocrisy or distorting their actual position.              |
| Black-and-White_Fallacy                                        | Explicitly framing a complex issue as a false binary.                                                                                                                                    |
| Bandwagon, Reductio_ad_hitlerum                                | Leveraging popularity or negative historical associations to judge ideas.                                                                                                                |
| Doubt                                                          | Undermining the credibility of sources by questioning their motives or expertise.                                                                                                        |
| Appeal_to_Time                                                 | Substituting temporal assertions ( <i>it is only a matter of time</i> ) for logical proof.                                                                                               |

["Loaded\_Language",  
"Name\_Calling, Labeling"]}

#### A.4 Constructive-Thread Share by Year and Stratum

Table 19 gives the full temporal breakdown underlying the topic-level trend in Table 13: for each of the four party  $\times$  topic strata and every sampled year, it reports the *constructive-thread share* – the percentage of threads whose mean interaction score falls in the Constructive bin ( $\geq +0.25$ ; the top of the four thread-quality bins in Table 7). Many early-year and Die Linke cells rest on very few threads (see  $n$  in parentheses) and should be read with caution; the shares stabilize only in the better-sampled years (2023–2026), and from 2024 onward the AfD strata are consistently less constructive than the Die Linke strata.

#### A.5 Stance-Relation Interaction Quality by Party

Table 20 breaks the stance-relation analysis of Section 5.1 (Table 12) down by individual party  $\times$  topic stratum, confirming that the climate pattern holds within both channels: opposing-stance climate exchanges are constructive on average on the AfD and Die Linke channels alike, while same-stance destructiveness is markedly stronger on the AfD side.

Table 19: Constructive-thread share – the percentage of threads whose mean interaction score lands in the Constructive bin ( $\geq +0.25$ ) – by year and party  $\times$  topic stratum (year of video publication), with thread counts  $n$  in parentheses. Strata abbreviated: C = Climate, M = Migration; A = AfD, L = Die Linke. “–” marks strata with no sampled threads that year; low- $n$  cells are not reliable.

| Year | C $\times$ A | C $\times$ L | M $\times$ A | M $\times$ L |
|------|--------------|--------------|--------------|--------------|
| 2016 | –            | –            | –            | 100.0 (1)    |
| 2017 | –            | 100.0 (3)    | –            | 50.0 (2)     |
| 2018 | 33.3 (3)     | 100.0 (3)    | 0.0 (10)     | –            |
| 2019 | 19.4 (36)    | 66.7 (9)     | 4.9 (41)     | 50.0 (2)     |
| 2020 | 9.1 (66)     | 100.0 (2)    | 6.8 (73)     | –            |
| 2021 | 19.3 (88)    | 43.8 (16)    | 9.8 (41)     | 33.3 (6)     |
| 2022 | 11.3 (142)   | 22.2 (9)     | 9.4 (53)     | 100.0 (3)    |
| 2023 | 7.0 (86)     | 7.0 (43)     | 8.5 (118)    | 20.0 (5)     |
| 2024 | 20.7 (29)    | 43.2 (44)    | 13.7 (95)    | 23.3 (133)   |
| 2025 | 11.4 (44)    | 30.9 (317)   | 21.0 (62)    | 28.3 (290)   |
| 2026 | 0.0 (6)      | 27.8 (54)    | 14.3 (7)     | 25.9 (58)    |

#### A.6 Inter-Annotator Agreement and Quality Validation

To ensure our validation gold standard is reliable, this section presents the agreement metrics computed across our independent annotation sample ( $n = 700$ ) prior to final adjudication. Table 21 summarizes the exact agreement, Cohen’s Kappa variants, and Krippendorff’s Alpha coefficients across all four core tasks.

Table 22 reports two reliability views: (a) the cross-annotator interaction-quality confusion matrix between annotators A and Z, and (b) the LLM

Table 20: Interaction quality by parent–child stance relation, per party  $\times$  topic stratum. Mean interaction score (range  $-1$  to  $+1$ ); C/D = % constructive / % destructive. “Opposing” = skeptic vs. supporter; “One side neutral” = exactly one side neutral. Opposing-stance exchanges are constructive on average (positive mean) only on climate, and on both party channels.

| Stance relation  | Climate |       |           |       | Migration |       |           |       |
|------------------|---------|-------|-----------|-------|-----------|-------|-----------|-------|
|                  | AfD     |       | Die Linke |       | AfD       |       | Die Linke |       |
|                  | Mean    | C/D   | Mean      | C/D   | Mean      | C/D   | Mean      | C/D   |
| Opposing         | +0.04   | 51/48 | +0.12     | 55/45 | −0.11     | 43/55 | −0.06     | 45/53 |
| Same stance      | −0.31   | 15/71 | −0.03     | 36/43 | −0.34     | 12/77 | −0.08     | 35/47 |
| One side neutral | −0.28   | 19/67 | −0.10     | 36/50 | −0.25     | 20/62 | −0.12     | 33/52 |

Table 21: Inter-annotator agreement ( $n = 700$ , pre-adjudication).  $\kappa$ : unweighted Cohen’s  $\kappa$  (interaction also quadratic-weighted, 0.494);  $\alpha$ : Krippendorff’s (nominal); propaganda reports exact-set match and mean Jaccard.

| Task                | Exact | $\kappa$ | $\alpha$           |
|---------------------|-------|----------|--------------------|
| Interaction Quality | 56.9% | 0.456    | 0.453              |
| Stance (Child)      | 87.4% | 0.640    | 0.639              |
| Stance (Parent)     | 87.0% | 0.666    | 0.665              |
| Propaganda          | 19.3% | —        | 0.335 <sup>†</sup> |

<sup>†</sup> Mean Jaccard index, not Krippendorff’s  $\alpha$  (undefined for the multi-label task).

annotator (claude-opus-4-6) validated against the human gold standard, per propaganda label.

The two annotators differ mainly in strictness rather than in content-specific judgment. Annotator A is systematically harsher (mean interaction score  $-0.183$  vs.  $-0.057$ ), assigning Destructive Disagreement to 227 pairs against Z’s 142, while Z assigns more Constructive Agreement (98 vs. 65). This offset is stable across topic (climate  $-0.104$ , migration  $-0.147$ ;  $p = 0.44$ ) and across party (AfD  $-0.086$ , Die Linke  $-0.166$ ;  $p = 0.15$ ), so it is consistent with a difference in threshold rather than a topic- or party-linked bias. Agreement between the two annotators is higher on each axis separately (communication style 63.4%, functional relation 67.9%) than on the joint five-class label (56.9%), and the most frequent single disagreement is A labeling a pair Destructive Disagreement where Z labels it Constructive Disagreement (45 cases): both agree on labeling the disagreement but differ on whether its tone is hostile or engaged. For propaganda, Z labels more techniques per pair (1.76 vs. 1.46), with the largest divergences on Loaded\_Language and Causal\_Oversimplification (more frequent for Z) against Doubt (more frequent for A), which accounts for the low exact-set agreement on that task.

Table 22: Reliability and annotation validation. (a) Cross-annotator interaction-quality confusion matrix between annotators A and Z ( $n = 700$ ); classes DD ( $-1.0$ ), DA ( $-0.5$ ), N ( $0.0$ ), CA ( $+0.5$ ), CD ( $+1.0$ ). (b) LLM annotator (claude-opus-4-6) vs. the human gold standard, per-label propaganda  $F_1$  (global Hamming loss = 0.1180).

| A\Z   | DD         | DA         | N         | CA        | CD        | Total |
|-------|------------|------------|-----------|-----------|-----------|-------|
| DD    | <b>120</b> | 26         | 32        | 4         | 45        | 227   |
| DA    | 5          | <b>111</b> | 13        | 14        | 4         | 147   |
| N     | 7          | 24         | <b>50</b> | 36        | 4         | 121   |
| CA    | 1          | 9          | 14        | <b>35</b> | 6         | 65    |
| CD    | 9          | 6          | 34        | 9         | <b>82</b> | 140   |
| Total | 142        | 176        | 143       | 98        | 141       | 700   |

(a) Interaction-quality confusion matrix.

| Technique                           | $F_1$  |
|-------------------------------------|--------|
| Appeal_to_Authority                 | 0.3714 |
| Appeal_to_Time                      | 0.5753 |
| Appeal_to_fear-prejudice            | 0.5792 |
| Bandwagon, Reductio_ad_hitlerum     | 0.1702 |
| Black-and-White_Fallacy             | 0.3333 |
| Causal_Oversimplification           | 0.5472 |
| Doubt                               | 0.3874 |
| Exaggeration, Minimisation          | 0.4906 |
| Flag-Waving                         | 0.4865 |
| Loaded_Language                     | 0.7565 |
| Name_Calling, Labeling              | 0.7366 |
| Repetition                          | 0.3478 |
| Slogans/Thought-terminating_Cliches | 0.3500 |
| Whataboutism, Straw_Men             | 0.3095 |
| Macro $F_1$                         | 0.4601 |

(b) claude-opus-4-6 vs. gold per-label  $F_1$ .