Proceedings of the Twentieth International AAAI Conference on Web and Social Media (ICWSM 2026)

Detecting Divisive Language: A Concept-Grounded, LLM-Guided Pipeline for Polarizing Social Media Sphere

Yuting He,[1] **Tianhao Li,**[2] **Jiebiao Wang,**[3] **Yongjun Zhang**[4]

[1]School of Journalism and Media, The University of Texas at Austin
[2]School of Information, The University of Texas at Austin
[3]Crown Family School of Social Work, Policy, and Practice, University of Chicago
[4]Department of Sociology, Stony Brook University

Abstract

Political polarization poses a growing global challenge, yet existing NLP approaches typically rely on indirect proxies such as toxicity or negative sentiment, which fail to capture identity-based antagonism that is central to polarizing discourse. We address this gap by conceptualizing identity-related polarization discourse as *divisive language*: language that explains political or social disagreement by attributing it to group-based identities. Building on this definition, we propose a staged training pipeline that uses large language models (LLMs) to generate definition-grounded supervision and progressively distills it into lightweight classifiers suitable for large-scale analysis. Experiments on social media data show that the resulting models substantially outperform zero-shot prompting and small-scale supervised baselines, while detecting forms of polarization that are not captured by toxicity- or sentiment-based methods. Our findings demonstrate that divisive language can be treated as a distinct, computable linguistic construct, enabling scalable and theoretically grounded analysis of political polarization.

Introduction

Polarized discourse, rather than promoting connective engagement (Lukito et al. 2024), deepens societal divisions, externalizes conflicts by blaming opposing groups, and prioritizes opposition over dialogue-building (Doornbosch, van Vuuren, and de Jong 2025). This discourse extends beyond elite political circles into the broader public sphere (Bail et al. 2018), notably influencing contentious social issues such as immigration (Johnston, Newman, and Velez 2015), climate change (Chinn, Hart, and Soroka 2020), and abortion rights (Mouw and Sobel 2001). Such discourse radicalizes opinions, erodes social cohesion, and threatens democracy (Baldassarri and Gelman 2008).

Computational social science has sought linguistic indicators capable of characterizing polarization at scale. Prior work largely operationalized polarization through undesirable or harmful language categories such as hate speech (Parker and Ruths 2023), toxicity (Kaati, Shrestha, and Akrami 2022), or profanity (Teh and Cheng 2020) to infer political polarization. We argue that these constructs provide only indirect and largely context-independent signals of political polarization. At its core, polarizing language lies in attributing attitudinal cleavages to group identities.

Thus, a conceptual asymmetry remains under-examined in current NLP approaches: (1) undesirable or harmful language (e.g., hate speech, toxicity) is not inherently polarizing, as such expressions may target individuals rather than social groups and therefore do not reinforce in-group/out-group boundaries; (2) conversely, polarizing discourse can manifest without using this language, which may appear in humorous, satirical, or seemingly neutral forms that attribute disagreement to identity categories without deploying explicit insults, profanity, or overtly uncivil language (Orthaber 2019; Lim 2025; Muti 2025). Particularly, the latter dynamic has become increasingly salient in contemporary political discourse (Boukes 2025).

To our knowledge, this study is among the first to develop an NLP classifier for polarizing discourse. We conceptualize such discourse as *divisive language*, defined as discourse that attributes political or social disagreement to group-based identities. Thus, group identity becomes the primary explanatory frame through which differences are understood, such that it transforms attitudinal differences into identity-based boundaries. Following this definition, we collected social media data where divisive language was likely to occur and constructed a human-annotated ground-truth dataset using an expert-validated codebook. We then developed an automated NLP pipeline and trained transformer-based models for detecting divisive language.

The goal of this study is to develop one of the first classifiers enabling automated and scalable detection of identity-based antagonism on social media and offering new methodological tools and linguistic perspectives for studying polarization beyond hate, toxicity, or sentiment frameworks. Such a classifier supports platforms and policymakers in distinguishing disagreement from identity-based division, observing polarization as it unfolds, and contributing to democratic and deliberative civic engagement on social media.

Related Work

NLP Efforts on Political Polarization

Political polarization refers to a process in which members of a political community increasingly align with opposing

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

camps in their political beliefs and issue positions (Schedler 2023). Polarization is not only an outcome reflected in the distribution of attitudes, but also a communicative process that is produced and reinforced through discourses (Dixit and Weibull 2007; Baldassarri and Bearman 2007). Identity-based divisive language is a key linguistic mechanism in this process: it explains attitudinal disagreement through us–versus–them framing, thereby widening perceived inter-group distance, hardening group boundaries, and eroding common ground for mutual deliberation.

Existing NLP tools often rely on "undesired language" tasks (e.g., toxicity, hate, offensiveness) as proxies for polarization, either by harmonizing these datasets or directly repurposing them. However, evidence suggests that polarization and undesired language are correlated but not equivalent: for example, many polarized discourses are non-abusive, while many abusive statements are not inherently polarized (Susanto et al. 2025). Moreover, polarization is highly context- and culture-dependent, making it particularly challenging to model computationally (Naseem et al. 2025). To address this gap, we develop a divisive language classifier that directly captures identity-based antagonism—semantically aligning with the core characteristic of polarization as opposition between social groups, thereby avoiding the limitations of sentiment-based proxies.

Divisive Language

To inductively derive a theoretical definition and support operationalization, we conducted a three-stage literature review: collecting 5,639 articles via OpenAlex, applying BERT-based topic modeling to identify fifteen thematic clusters, and reviewing highly cited works to synthesize recurring semantic features. Across clusters, divisiveness consistently centers on group cleavages between in-groups and out-groups.

Building on this inductive evidence, we define *divisive language* as discourse that explains political or social disagreement by attributing differences to fundamental distinctions between social groups. In such discourse, group identity becomes the primary explanatory frame, shifting attention from arguments, interests, or contextual factors to who people are. We identify two core characteristics of divisive language: (1) **identity-based opposition**, in which actors are positioned as members of opposing social groups; and (2) **attributional absolutism**, in which group identity is treated as the primary and often sufficient explanation for social or political outcomes (Miller 2014). Correspondingly, we operationalize divisive language through two observable signals: (1) the presence of identifiable social groups, and (2) the attribution of outcomes—often negative—to those group identities. Examples are provided in the Appendix.

LLM-Guided Annotation and Bootstrapped Classification

Recent work in NLP has increasingly examined the use of large language models (LLMs) in data construction and supervision, particularly in settings where target constructs are abstractive or context-dependent. Rather than relying exclusively on direct human annotation, recent work explores how LLMs can function as annotators, instructors, or sources of weak supervision to support downstream model training through synthetic labeling, self-training, or distillation-based approaches (Gilardi, Alizadeh, and Kubli 2023; Pangakis and Wolken 2024; Törnberg 2024).

Prior research suggests that LLM-generated annotations can closely approximate human judgments when prompts are carefully specified and constraints are explicitly enforced. Such findings have been reported in domains including sentiment analysis, stance detection, and language identification (Gilardi, Alizadeh, and Kubli 2023; He et al. 2025). Among them, LLMs are typically used either to label large collections of data or to generate additional training instances that complement smaller human-annotated datasets. The resulting pseudo-labeled corpora are then employed to train conventional supervised models, allowing classification systems to be scaled without proportional increases in annotation cost.

Some research introduces instruction tuning and teacher-student distillation frameworks, which focus on transferring task knowledge from large, computationally intensive models to smaller and more efficient architectures (Hinton, Vinyals, and Dean 2015; Sanh et al. 2019). More recent studies extend this paradigm to tasks that require structured reasoning, showing that supervision signals produced by large models, such as explicit rationales or stepwise analyses, can be distilled into smaller models with substantial performance retention (Wang et al. 2023; Magister et al. 2023). LLM-guided annotation and bootstrapped training have been established as a general methodological approach for large-scale text classification, particularly in domains where manual annotation is costly and abstract theoretical constructs are not easily reducible to lexical cues.

Method

This study proposes a staged training pipeline for scalable detection of divisive language that integrates LLMs with conventional transformer-based classifiers. The design of the method is guided by two constraints identified in prior sections. First, divisive language is an abstract, semantic construct that cannot be reliably operationalized through surface lexical cues. Second, large-scale analysis of social media discourse requires models that are computationally efficient and suitable for deployment at scale.

To address these constraints, we proposed a method that deliberately separates concept-level supervision from high-throughput classification. Rather than directly training a lightweight classifier on a limited set of labeled data, we first ensure alignment with the theoretical definition of divisive language through reasoning-intensive supervision provided by a large language model. This supervision is then progressively transferred to smaller and more efficient models through a staged process. The resulting pipeline enables large-scale classification while preserving fidelity to the underlying theoretical construct.

Concretely, this method consists of three stages: (1) a large proprietary LLM is used to generate synthetic annotations explicitly grounded in the definition of divisive language; (2) an instruction-tuned open-source LLM is fine-

tuned on this dataset and subsequently used to infer labels for a substantially larger unlabeled corpus; (3) a lightweight encoder model is trained on the resulting large pseudo-labeled dataset for efficient deployment. This staged design allows conceptually precise supervision to be distilled from powerful but costly models into models optimized for scale.

Stage I: Definition-Grounded Synthetic Annotation

The first stage constructs an initial labeled dataset using **GPT-5.2**. Data are annotated by prompting the model with a detailed definition of divisive language. The prompt instructs the model to first produce a natural-language analysis explicitly grounded in this definition and then assign a categorical label.

In this stage, **GPT-5.2** assigns labels using a three-category label space (`divisive_language`, `potentially_divisive_language`, and `non_divisive_language`), allowing the annotation process to explicitly represent uncertainty in cases where identity-based framing is present but not clearly dominant. This design choice preserves conceptual nuance at the supervision stage, prior to any large-scale inference.

Requiring explicit reasoning prior to label assignment serves two methodological purposes. First, it constrains the model to base its decisions on the theoretical construct rather than on latent heuristics or correlated surface features. Second, beyond improving interpretability, it provides a richer supervision signal that can be transferred to downstream models during subsequent training stages. All outputs are constrained using a strict JSON schema to ensure structural consistency and eliminate post-processing ambiguity.

The annotated data are partitioned into training, validation, and a held-out global test set. The global test set is independently annotated by human annotators, with disagreements adjudicated by a domain expert with PhD-level expertise in this research area. This procedure provides an external check on the conceptual alignment of the LLM-generated annotations.

Stage II: Instruction-Tuned LLM for Scalable Label Generation

Although **GPT-5.2** provides conceptually precise annotations, its computational cost and accessibility constraints limit its suitability for large-scale inference. To address this limitation, an open-source instruction-following model, **LLaMA-3.1-8B-Instruct**, is fine-tuned on the **GPT-5.2**-labeled training data.

At this stage, the task is formulated as binary classification by removing the intermediate `potentially_divisive_language` category to reduce ambiguity during large-scale inference. The fine-tuned model is trained using the same definition of divisive language and the same reasoning-first output format, again constrained by a JSON schema.

After fine-tuning, the instruction-tuned LLM is used to infer labels for a large unlabeled corpus. While the model generates both analyses and labels, only the labels are retained for subsequent training. This stage transfers the concept-aligned decision boundary learned from **GPT-5.2** into an open-source model capable of scalable data annotation.

Stage III: Lightweight Classifier Training

In the final stage, a **RoBERTa**-based classifier is trained on the large pseudo-labeled dataset produced in Stage II. **RoBERTa** is selected for its strong empirical performance, computational efficiency, and suitability for large-scale deployment scenarios.

To ensure comparability across models, the same held-out global test set is used for final evaluation across all experimental conditions. By isolating large-scale pseudo-labeling from evaluation design, this stage enables assessment of how concept-aligned supervision can be effectively distilled into lightweight classifiers suitable for high-throughput analysis.

Experiments

Dataset

The raw dataset consists of comments posted under YouTube videos related to the 2024 U.S. presidential election during the final stretch of the campaign (October 1–November 5, 2024; $N = 8{,}315{,}407$), collected via the YouTube Research API. This period corresponds to heightened political engagement and attitudinal competition. Comments are ranked by reply count, motivated by prior work showing that highly replied comments are more likely to provoke interaction, either by eliciting contestation associated with divisive language or by encouraging dialogue through connective language (Lukito et al. 2024).

Annotation

From the ranked corpus, we sample a subset of comments for annotation and model training. An initial set of 10,000 comments is selected and annotated using **GPT-5.2**, following the definition-grounded, reasoning-first procedure described in Stage I of the method. Each instance is annotated with both a natural-language analysis and a categorical label, with outputs constrained by a strict JSON schema.

The annotated dataset is partitioned into training, validation, and a held-out global test set. Specifically, 9,000 instances are used for training, 600 for validation, and 400 are reserved for global testing. The global test set is independently annotated by human annotators. In cases of disagreement, labels are adjudicated by a domain expert with PhD-level expertise in this research area. Human annotations show high agreement with **GPT-5.2** labels, indicating strong alignment with the underlying definition of divisive language; disagreements were resolved through expert adjudication.

To enable large-scale downstream training, the instruction-tuned **LLaMA-3.1-8B-Instruct** model described in Stage II is then applied to the ranked corpus to generate pseudo-labels for an additional 990,000 comments. These pseudo-labeled instances are used exclusively for training and are not included in evaluation. All evaluation results reported in this study are based solely on the held-out global test set.

Experimental Setup

All **GPT-5.2** and **LLaMA-3.1-8B-Instruct** inference steps are constrained using strict JSON schemas to enforce structured outputs and eliminate ambiguity during post-processing. During annotation with **GPT-5.2**, we retain the full three-label space (`divisive_language`, `potentially_divisive_language`, `non_divisive_language`). For all downstream models evaluated in this study, labels are mapped to a binary label space (divisive vs. non-divisive) following the formulation described in the Method section.

We evaluate multiple model configurations to isolate the contribution of each stage in the pipeline. Accuracy is used as the primary evaluation metric, as all models are evaluated on the same binary classification task using the global test set. For comparison, we also train a **RoBERTa** classifier using only the initial 9,000 labeled instances, without incorporating any pseudo-labeled data. To assess the contribution of explicit reasoning supervision, we conduct an ablation in which **LLaMA-3.1-8B-Instruct** is fine-tuned using the same training data and label space, but without the natural-language analysis component. All configurations are evaluated on the same 400-instance global test set.

Specifically, we compare the following model configurations:

- **LLaMA Zero-Shot**: **LLaMA-3.1-8B-Instruct** performs direct inference using the divisive language prompt without any task-specific fine-tuning.
- **LLaMA Fine-Tuned (w/o Analysis)**: **LLaMA-3.1-8B-Instruct** fine-tuned on the 9,000 **GPT-5.2**-labeled training instances and label space, but without the natural-language analysis component.
- **LLaMA Fine-Tuned (w/ Analysis)**: **LLaMA-3.1-8B-Instruct** fine-tuned on the 9,000 **GPT-5.2**-labeled training instances, using both the natural-language analysis and the categorical labels as supervision.
- **RoBERTa (Small-Scale Supervised)**: **RoBERTa** classifier trained on the initial 9,000 labeled instances only, without additional pseudo-labeled data.
- **RoBERTa (Large-Scale Pseudo-Labeled)**: **RoBERTa** classifier trained on the 990,000 instances generated by the instruction-tuned LLM.

In our experiments, we use the official OpenAI API for inference with **GPT-5.2** model and use NVIDIA RTX A6000 GPUs for both finetuning and inference with **LLaMA-3.1-8B-Instruct** and **RoBERTa** models.

Results

Table 1 reports accuracy scores for all model configurations on the global test set. Across all configurations, models that incorporate task-specific supervision consistently outperform zero-shot prompting, and large-scale pseudo-labeling allows lightweight classifiers to approach the performance of instruction-tuned LLMs. In addition, removing the natural-language analysis component from supervision leads to a systematic reduction in accuracy, indicating that reasoning-based supervision contributes meaningfully to the learned decision boundary. Taken together, these results emphasize the combined role of concept-aligned supervision, explicit reasoning, and data scale in enabling effective detection of divisive language.

Model Configuration	Accuracy
LLaMA (Zero-Shot)	0.535
LLaMA (Fine-Tuned, w/o Analysis)	0.765
LLaMA (Fine-Tuned, w/ Analysis)	**0.827**
RoBERTa (Small-Scale Supervised)	0.625
RoBERTa (Large-Scale Pseudo-Labeled)	**0.827**

Table 1: Accuracy on the 400-instance global test set.

Discussion

This study demonstrates that divisive language can be detected at scale without reducing the construct to adjacent proxies such as toxicity or sentiment. By grounding supervision in an explicit theoretical definition and combining it with staged model distillation, the proposed pipeline operationalizes an abstract, identity-based notion of disagreement in a form suitable for large-scale computational analysis.

The results underscore the value of separating concept-level supervision from high-throughput classification. Rather than relying on zero-shot prompting or directly training lightweight models on limited labeled data, the approach establishes conceptual alignment through reasoning-intensive supervision and progressively transfers this alignment to smaller, more efficient models. In this framework, LLMs serve as methodological intermediaries that translate theory-driven constructs into structured supervision signals, enabling scalable analysis while preserving fidelity to the underlying concept.

Several limitations and ethical considerations warrant attention. Given the limited size of the held-out test set, these results are best interpreted as evidence of methodological feasibility rather than as definitive performance benchmarks. Accuracy is reported to ensure comparability across model configurations; future work will explore precision-recall tradeoffs in application-specific settings. Although the annotation procedure shows high agreement between LLM-generated labels and expert judgments, automated classification of political discourse inevitably entails the risk of misclassification, which may affect downstream analyses. Moreover, like other language technologies applied to political content, the proposed method could be misused for purposes such as censorship or intrusive surveillance if deployed without safeguards. To mitigate these risks, we rely on data collected via approved research APIs, apply anonymization procedures, and provide detailed documentation of definitions, annotation protocols, and model configurations to support transparency and reproducibility. The intended use of the models is explicitly framed around research and diagnostic analysis rather than enforcement or content moderation. We have reviewed the relevant ethics guidelines and ensured that the study conforms to established standards for responsible research on political communication and social media data.

References

Bail, C. A.; Argyle, L. P.; Brown, T. W.; Bumpus, J. P.; Chen, H.; Hunzaker, M. F.; Lee, J.; Mann, M.; Merhout, F.; and Volfovsky, A. 2018. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, 115(37): 9216–9221.

Baldassarri, D.; and Bearman, P. 2007. Dynamics of political polarization. *American Sociological Review*, 72(5): 784–811.

Baldassarri, D.; and Gelman, A. 2008. Partisans without constraint: Political polarization and trends in American public opinion. *American Journal of Sociology*, 114(2): 408–446.

Boukes, M. 2025. Deliberation in online political talk: exploring interactivity, diversity, rationality, and incivility in the public spheres surrounding news vs. satire. *Journal of Communication*, 75(2): 125–136.

Chinn, S.; Hart, P. S.; and Soroka, S. 2020. Politicization and polarization in climate change news content, 1985-2017. *Science Communication*, 42(1): 112–129.

Dixit, A. K.; and Weibull, J. W. 2007. Political polarization. *Proceedings of the National Academy of sciences*, 104(18): 7351–7356.

Doornbosch, L. M.; van Vuuren, M.; and de Jong, M. D. 2025. Moving beyond us-versus-them polarization towards constructive conversations. *Democratization*, 32(1): 96–122.

Gilardi, F.; Alizadeh, M.; and Kubli, M. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30): e2305016120.

He, Q.; Jia, Y.; Li, W.; Liao, S.; Quan, R.; Cui, T.; and Qin, J. 2025. Large Models are Good Annotators for Zero-Shot Learning. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2764–2768.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Johnston, C. D.; Newman, B. J.; and Velez, Y. 2015. Ethnic change, personality, and polarization over immigration in the American public. *Public Opinion Quarterly*, 79(3): 662–686.

Kaati, L.; Shrestha, A.; and Akrami, N. 2022. A machine learning approach to identify toxic language in the online space. In *2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 396–402. IEEE.

Lim, E. C. N. 2025. Humour and Social Power: A Sociological Lens on Political Communication. *Open Journal of Social Sciences*, 13(10): 324–334.

Lukito, J.; Chen, B.; Masullo, G. M.; and Stroud, N. J. 2024. Comparing a BERT classifier and a GPT classifier for detecting connective language across multiple social media. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 19140–19153.

Magister, L. C.; Mallinson, J.; Adamek, J.; Malmi, E.; and Severyn, A. 2023. Teaching small language models to reason. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 2: short papers)*, 1773–1781.

Miller, E. C. 2014. Divisive Discourse: The Extreme Rhetoric of Contemporary American Politics. *Argumentation and Advocacy*, 51(2): 132–135.

Mouw, T.; and Sobel, M. E. 2001. Culture wars and opinion polarization: the case of abortion. *American Journal of Sociology*, 106(4): 913–943.

Muti, A. 2025. Hidden in plain sight: detecting misogyny beneath ambiguities and implicit bias in language.

Naseem, U.; Ren, J.; Anwar, S.; Kohail, S.; Veliz, R. A. G.; Geislinger, R.; Jabr, A.; Abdulmumin, I.; Qureshi, L.; Borkar, A. A.; et al. 2025. POLAR: A Benchmark for Multilingual, Multicultural, and Multi-Event Online Polarization. *arXiv preprint arXiv:2505.20624*.

Orthaber, S. 2019. Aggressive humour as a means of voicing customer dissatisfaction and creating in-group identity. *Journal of Pragmatics*, 152: 160–171.

Pangakis, N.; and Wolken, S. 2024. Knowledge distillation in automated annotation: Supervised text classification with LLM-generated training labels. In *Proceedings of the Sixth Workshop on Natural Language Processing and Computational Social Science (NLP+ CSS 2024)*, 113–131.

Parker, S.; and Ruths, D. 2023. Is hate speech detection the solution the world wants? *Proceedings of the National Academy of Sciences*, 120(10): e2209384120.

Sanh, V.; Debut, L.; Chaumond, J.; and Wolf, T. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Schedler, A. 2023. Rethinking political polarization. *Political science quarterly*, 138(3): 335–359.

Susanto, L.; Wijanarko, M. I.; Pratama, P. A.; Tang, Z.; Akyas, F.; Hong, T.; Idris, I. K.; Aji, A. F.; and Wijaya, D. T. 2025. A Multi-Labeled Dataset for Indonesian Discourse: Examining Toxicity, Polarization, and Demographics Information. In *Findings of the Association for Computational Linguistics: ACL 2025*, 18863–18890.

Teh, P. L.; and Cheng, C.-B. 2020. Profanity and hate speech detection. *International Journal of Information and Management Sciences*, 31(3): 227–246.

Törnberg, P. 2024. Best practices for text annotation with large language models. *arXiv preprint arXiv:2402.05129*.

Wang, Y.; Kordi, Y.; Mishra, S.; Liu, A.; Smith, N. A.; Khashabi, D.; and Hajishirzi, H. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, 13484–13508.

Paper Checklist

For most authors...

1. Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Answer: Yes. No privacy violation or targeting of sensitive populations.
2. Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Answer: Yes
3. Do you clarify how the proposed methodological approach is appropriate for the claims made? Answer: Yes
4. Do you clarify what are possible artifacts in the data used, given population-specific distributions? Answer: Yes
5. Did you describe the limitations of your work? Answer: Yes
6. Did you discuss any potential negative societal impacts of your work? Answer: Yes. Negative societal impact is possible but limited; risks mainly relate to misclassification of political content.
7. Did you discuss any potential misuse of your work? Answer: Yes. Potential misuse (censorship/surveillance) is noted.
8. Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Answer: Yes. We mitigate risks via documentation, anonymization, restricted-use guidance, and reproducible release.
9. Have you read the ethics review guidelines and ensured that your paper conforms to them? Answer: Yes

Additionally, if your study involves hypotheses testing...

1. Did you clearly state the assumptions underlying all theoretical results? Answer: N/A. No formal hypotheses or theoretical assumptions are tested. This section of questions is not fit for our study.
2. Have you provided justifications for all theoretical results? Answer: N/A.
3. Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? Answer: N/A.
4. Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? Answer: N/A.
5. Did you address potential biases or limitations in your theoretical framework? Answer: N/A.
6. Have you related your theoretical results to the existing literature in social science? Answer: N/A.
7. Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? Answer: N/A.

Additionally, if you are including theoretical proofs...

1. Did you state the full set of assumptions of all theoretical results? Answer: N/A. No theoretical assumptions or formal results are stated. This section of questions is not fit for our study.
2. Did you include complete proofs of all theoretical results? Answer: N/A.

Additionally, if you ran machine learning experiments...

1. Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Answer: No. Due to privacy constraints, code and data are available upon request under agreement.
2. Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Answer: Yes. Despite space constraints, we provide critical training details in the paper.
3. Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? Answer: No. We report results based on a single run. We acknowledge that reporting variance across multiple runs would strengthen the robustness of our findings and leave this for future work.
4. Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Answer: Yes.
5. Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Answer: Yes.
6. Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Answer: Yes. We explicitly consider asymmetric misclassification costs. False positives (over-labeling non-divisive content) may inflate estimates of polarization, while false negatives may obscure identity-based antagonism. Given our goal of detection rather than intervention, we prioritize reducing false positives, while acknowledging the importance of capturing subtle divisive expressions. The appropriate trade-off depends on downstream use.

Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...

1. If your work uses existing assets, did you cite the creators? Answer: Yes
2. Did you mention the license of the assets? Answer: Yes
3. Did you include any new assets in the supplemental material or as a URL? Answer: No. Due to platform policies and ethical constraints, the original dataset cannot be publicly released. However, we provide illustrative examples and annotation details in the appendix.
4. Did you discuss whether and how consent was obtained from people whose data you're using/curating? Answer: Yes
5. Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Answer: Yes. No PII; potential offensive political content is documented.
6. If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? Answer: Yes. We include representative human-coded examples derived from the original data in the appendix

to illustrate the annotation framework and support transparency.

7. If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? Answer: No. We did not create a formal Datasheet for the Dataset. While our annotated dataset includes rich metadata (e.g., timestamps and engagement metrics) derived from YouTube comments, we instead document key aspects of data collection, annotation, examples, and sampling in the paper and appendix.

Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

1. Did you include the full text of instructions given to participants and screenshots? Answer: N/A.
2. Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? Answer: N/A.
3. Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? Answer: N/A.
4. Did you discuss how data is stored, shared, and deidentified? Answer: N/A.

Appendix: Illustrative Examples

The following examples are drawn directly from the dataset and are presented without modification.

Divisive Language

Discourse that attributes political or social differences to group identity as the key causal explanation.

- "No real Christian would vote for Harris. God is not to be mocked."
- "Latino and blacks for Trump should be ashamed of themselves."
- "Make no mistake, you can't be a Christian and vote Democrat this year."

Potentially Divisive Language

Discourse that associates social groups with certain behaviors or characteristics, but does not explicitly attribute these outcomes to group identity as their primary or exclusive explanation.

- "Why do black people always focus on that they're black. I've literally never seen any other race segregate themselves as much"
- "Beware politicians peddling simplistic solutions to complex problems."
- "I can tell you as a black cop in Texas, not all of us, but most of us lost faith in the democratic party after George Floyd."

Non-Divisive Language

Discourse that either does not reference social groups or does not use group identity to explain behaviors or outcomes.

- "That Judge needs to be removed from the courts."
- "Blaming others for your failures is the hallmark of narcissism."
- "Why are people mad? All of this is real stuff that has happened"