PAIRSCALE: Analyzing Attitude Change in Online Communities

Rupak Sarkar[1], **Patrick Y. Wu**[2], **Kristina Miler**[1],
Alexander Hoyle*[3], and **Philip Resnik***[1]

[1]University of Maryland, College Park
[2]American University
[3]ETH Zürich
{rupak,kmiler,resnik}@umd.edu, hoylea@ethz.ch, patrickwu@american.edu

Abstract

We introduce a text-based framework for measuring attitudes in communities toward issues of interest, going beyond the pro/con/neutral of conventional stance detection to characterize attitudes on a continuous scale using both implicit and explicit evidence in language. The framework exploits LLMs both to extract attitude-related evidence and to perform pairwise comparisons that yield unidimensional attitude scores via the classic Bradley and Terry (1952) model. We validate the LLM-based steps using human judgments, and illustrate the utility of the approach for social science by examining the evolution of attitudes on two high-profile issues in U.S. politics in two political communities on Reddit over the period spanning from the 2016 presidential campaign to the 2022 mid-term elections. [WARNING: Potentially sensitive political content.]

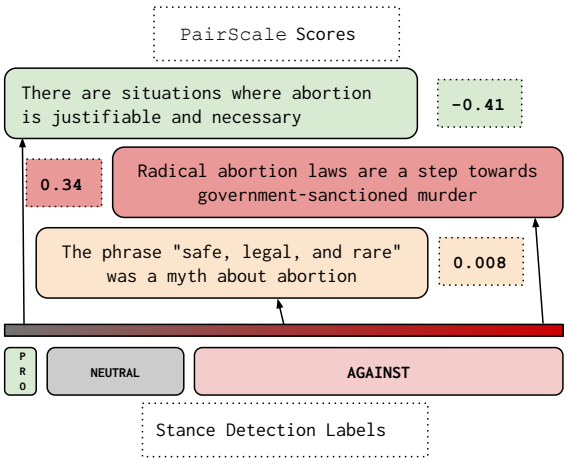


Figure 1: Comments in the r/Conservative subreddit about abortion are overwhelmingly "against" abortion. We derive a continuous scale that makes it possible to relate individual attitude expressions to the attitudes in the community as a whole.

1 Introduction

Measuring and understanding a community's attitudes on issues is notoriously difficult (Mastroianni and Dana, 2022). Practically, public opinion surveys require great care and effort to design, deploy, and analyze (Atkeson and Alvarez, 2018; Krupnikov and Findley, 2018). Looking at attitudes across time raises further challenges—survey changes, e.g. in format or wording, can cause problems in comparability between time periods (Bishop et al., 1978; Krosnick and Berent, 1993). In addition, the meaning of terms in surveys can change over time: to identify as a U.S. conservative in 2024, for example, implies a different set of beliefs than in 1984 (Lewis, 2021; Amira, 2022).

In this work, we introduce a new framework for measuring a community's attitudes toward issues and analyzing the dynamics of those attitudes over time. The framework requires three main steps. Given an issue of interest and a community engaged in perhaps wide-ranging discussion, the first step involves identifying statements in language that are relevant to that issue; we call these *attitude expressions*. Second, we must use those statements to characterize *valence* with regard to that issue. Finally, we need to use that information to characterize the collection of attitudes at the community level, and to do so over time.

Standard NLP methods lack nuance with respect to these problems. For extraction, NLP approaches in computational social science settings include parsing subject-predicate-object triples (Bamman and Smith, 2015) or semantic roles (Ash et al., 2024), but these rely only on the surface forms of text, failing to capture context and neglecting issue-relevant beliefs the author may hold even if those are not expressed overtly. For valence, the for/against/neutral categories afforded by traditional stance detection (Siddiqua et al., 2019; Allaway and McKeown, 2020; Li et al., 2021)

*The last two authors advised this project equally.

Findings of the Association for Computational Linguistics:
NAACL 2025, pages 1722–1738

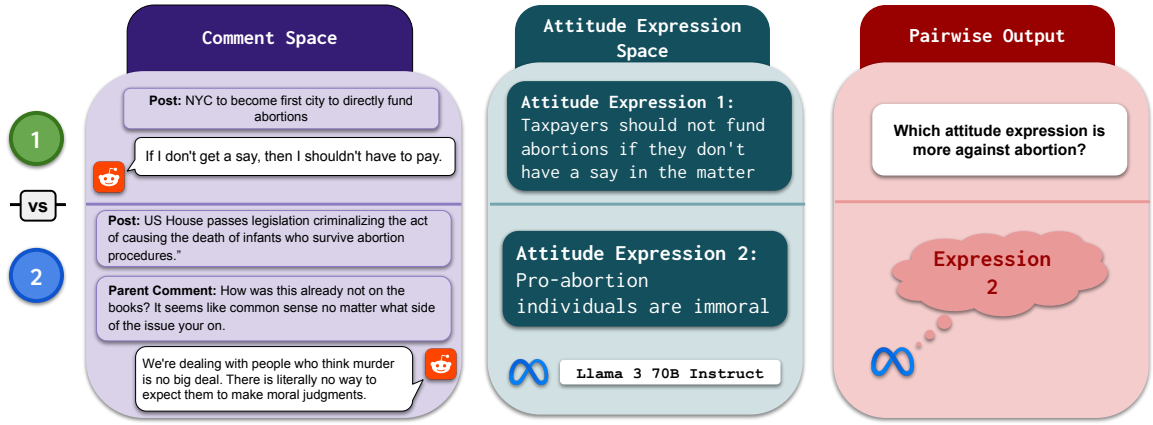


Figure 2: Illustration of our method, showing (a) pairing of contextualized comments, (b) extraction of attitude expressions from those comments, and (c) pairwise comparison of attitude expressions. Model-fitting based on pairwise comparisons yields a continuous scale that makes it possible to compare attitudes between communities (Table 2) and across time (Figures 3, 4, 5).

are too broad and reductive from a practitioner's standpoint. For example, in a real-world dataset consisting of abortion-related comments in the `r/Conservative` subreddit, `gpt-4o` marks the majority of comments as "against" abortion (Table 1)—a relatively trivial finding for a conservative community. For community-level characterizations, practitioners are often interested in how attitudes in a community change over time—e.g. how the legalization of gay marriage became mainstream within the Democratic Party (Dimock et al., 2013)—or how a specific issue-related belief might stand relative to the community's set of beliefs as a whole.

The framework we introduce addresses these issues. First, we extract issue-specific attitude expressions by adapting the method presented in Hoyle et al. (2023), transforming contextually observed texts into collections of issue-relevant attitude expressions. Hoyle et al. refer to these as *(inferential) decompositions*, since they are inferred and capture components of a text's contextually derived meaning (Borg and Fisher, 2021). They may be overt in the text, entirely implicit, or a mix of the two. Representing them in natural language casts them as objects amenable to analysis by downstream NLP methods that might not themselves be sensitive to implicit content (including in-group language common to political conversation, KhudaBukhsh et al. 2020; Holtgraves and Bray 2023). The importance of implicit attitude expressions is made evident by an inspection of stance detection outputs for one issue we considered: a failure to *explicitly* mention opposition to abortion often leads to a "neutral" label. For example, on a post lamenting a lack of Republican push to ban abortion, a reply saying *"That ship has sailed. We need a new GOP"* was labeled as `neutral` although it indirectly expresses a negative attitude on the subject.[1]

Second, we characterize attitude expressions about an issue on a scalar support–oppose (for–against) continuum. Here, we are motivated by longstanding literature in social science where items are placed on a unidimensional *scale* (Goertz, 2020); e.g., ideal point models using votes to infer legislators' position on a liberal–conservative axis (Poole and Rosenthal, 1985). To do so, we follow Wu et al.'s (2024)'s comparison-based approach, passing sampled pairs of extracted attitude expressions to an LLM that compares each pair with regard to perceived support of the issue. Values for the items on a unidimensional scale are then inferred by fitting a Bradley and Terry (1952) model for the outcomes of pairwise comparisons.

Third, we track the relative position of an attitude expression on that scale across time. Using pairwise comparisons based on language makes it straightforward to characterize attitudes in *pre-existing discussions*, where they can be placed on a scale that is comparable across time periods (even including issues that were never probed in traditional surveys at the time). We refer to this idea with the term "retroactive surveys" (§6). Our technique also allows us to examine emergent (and diminishing) attitude expressions over time (§7).

[1]Based on findings of Cruickshank and Ng (2024), we posit that zero-shot use of GPT-4o reasonably represents what can be accomplished by stance detection in the absence of the significant per-issue data annotation investment needed to support supervised model training or LLM fine-tuning.

	Favor	Neutral	Against
2015	5.33	30.67	64.0
2019	5.33	26.67	68.0

Table 1: Illustrative stance detection results from GPT-4o. While most comments are labeled as "neutral" or "against", the ontology of stance detection provides a coarse view of the attitude landscape.

We call our overall approach PAIRSCALE[2]. In this paper, our substantive use of the new method focuses on two salient and contentious issues in U.S. politics: abortion and immigration. We do so across two Reddit communities involving self-identified Conservatives and Democrats; we compare attitudes across communities and investigate shifts over time, covering a wide period from the beginning of Donald Trump's first presidential campaign in 2015 through to the 2022 midterms.

To summarize our contributions,

- We introduce a new framework, PAIRSCALE, that places attitude expressions about an issue, identified in a community within a time frame, on a polar scale.
- We validate key component steps using human judgments.[3]
- Using our framework, we quantifiably compare attitudes about two important political issues, looking at two different communities at different points in time.
- We use PAIRSCALE scores to investigate the dynamics of community attitudes over time.

2 What is an attitude?

We follow Adcock and Collier (2001) in distinguishing between *background* and *systematized* versions of a concept when defining constructs. The background concept at the heart of our paper is the opinion expressed toward an issue based on content. As a general scheme for systematization, this can be formulated as a triple of ⟨item, target, polar variable⟩, where *item* refers to the evidence/content, *target* refers to the issue towards which this attitude is directed, and the *polar variable* ranges over possible values expressing polarity according to a particular interpretation.[4]

In a popular systematization, stance detection, the item is a text (e.g. a tweet Glandt et al., 2021), the target is a topic or an entity, such as MASKING, or ANTHONY FAUCI, and the *polar variable* is ordinal, taking values like "pro", "neutral", and "against". The systematized interpretation of the polar variable is typically application dependent; e.g., Gilardi et al. (2023) use repeal/neutral/keep for stance toward legislation. In a different systematization, classical vote-based ideological scaling in political science (Poole and Rosenthal, 1985) has been modified to accommodate distinct continuously-valued ideal points (polar variable) for different issues (targets) (Gerrish and Blei, 2012; Lauderdale and Clark, 2014; Shin, 2024), underlying sets of votes (content).

For attitude as systematized in this paper, the *content* is a piece of text, the *target* is an issue under discussion, and, like ideal points, the *polar variable* is continuous. Like ideal points, its interpretation is relative to the entire population being analyzed, and its interpretation depends on how a *comparison construct* relevantly defines polarity (§ 4.3)

3 Related Work

Measuring a construct through pairwise comparisons is common in the political science and social science literature (Benoit et al., 2019; Carlson and Montgomery, 2017; Chen et al., 2013). Gienapp et al. (2020) used the Bradley-Terry model to obtain high-quality labels from data with crowdsourced pairwise comparisons. Using LLMs to perform pairwise comparisons instead of crowdworkers is relatively recent. Wu et al. (2023) used pairwise comparisons made using LLMs to recreate traditional ideological scales, such as DW-NOMINATE (Poole and Rosenthal, 1985).

Closest to our approach, Wu et al. (2024) used scores obtained from pairwise comparisons to measure affective polarization. However, their unit of analysis is at the level of comments rather than attitude expressions. So, for example, the method can measure how anti-abortion attitudes have shifted over time, but it is not able to track temporal shifts for a specific, issue-related expression of attitude, such as "Abortion is evil'.

[2]Code and data available at https://github.com/styx97/pairscale

[3]Our approach is intended to support social science research, which places a premium on validity (Christian Baden and van der Velden, 2022), and we consider human validation of component steps to be an essential part of the work.

[4]Notice that in this formulation, attitude is a property of evidence, e.g. a text, not a mental state of the person who produced that evidence, since our focus here is on attitudes in a community, not attitudes of individual users. The content, in our case, may be observed or inferred.

	Abortion		Immigration	
	Text	**Score**	**Text**	**Score**
r/Conservative	Abortion takes the life of a human being	0.20	Advocating for amnesty for illegal immigrants is treason	0.20
r/Conservative	Abortion supporters manipulate public opinion with false stories	0.13	Illegal immigrants will increase crime in the cities they are sent to	0.122
r/Conservative	Abortion should be regulated, not banned	-0.08	Many female and child migrants are victims of sex trafficking	-0.04
r/democrats	Abortion bans can cause immense harm and suffering.	-0.09	Human trafficking in the US is a serious problem	-0.046
r/democrats	The pro-life movement is about controlling women's bodies	-0.11	Resources to alleviate issues at the border are a better use of funds than a wall	-0.075
r/democrats	Giving women the right to choose is a common sense issue	-0.18	Asylum seekers are in the US legally	-0.09

Table 2: Comparison of abortion and immigration attitude expressions found in the two communities from a single time period in 2022. The overlap in PAIRSCALE scores in Fig.5 corresponds to attitudes with similar polarity. For the middle two rows with the purple background, we show similar attitude expressions that have received similar scores.

4 Approach

We first collect comments about two contentious topics, abortion, and immigration, from a popular online conservative community, `r/Conservative` (§ 4.1). We then decompose the comments into attitude expressions such as, "illegal immigrants should be deported" with an LLM (§ 4.2). Then, for each issue, an LLM compares pairs of sampled attitude expressions to identify which is more or less in favor. Last, scalar scores are inferred for each item from these ranked comparisons (§ 4.3).[5]

4.1 Data

While polarization in the U.S. Congress is well-studied, our goal is to study attitude shifts among the public via its participation in online communities. We pick `r/Conservative` as our community of focus, as it is one of the few conservative communities on Reddit that has enjoyed significant membership across several years, with 1.1M members as of March 2024. We picked `r/democrats` as its liberal equivalent to situate and contrast our results.[6] To quantify the shift in attitude before, during, and after Donald Trump's presidency, we select comments from 2015 until the end of 2022. We break down this period into sections of six months, and sample 300 comments from each six-month period to study the shift in attitudes for each topic.

In our periods of focus, we notice a high rate of new active users in each six-month slice (Figure 7 in Appendix). This lines up with findings by Waller and Anderson (2021), who claim that the shift in Reddit in 2016 towards conservative views was driven by "new and newly political" users. It is important to note that rather than studying how *individuals* shift in their attitudes over time, we aim to study the shifts in a *community*, where much like Congress, the ideological space is defined by the ideologies of the current members.

During sampling, we allow more than one comment from a particular user. While this could risk the sample being dominated by a single user, this does not happen in practice. For the topic of abortion, over the 19 six-month periods considered, an average of 86% (4.5 s.d.) comments came from unique users. For immigration, that number was 81% (6.6 s.d.). This shows that an overwhelming majority of the comments in each time period are posted by unique users. Moreover, considering one comment per user might inaccurately represent the state of a community, as it is possible that a vocal minority who post more content (supported by the majority through their upvotes) may be the ones steering the discourse, and consequently policy decisions (Knoblock, 2020).

To ensure the expressed comments have community support, we select comments with at least

[5] For both decomposition and comparisons we use LLAMA-3.1-70B-INSTRUCT with 4-bit quantization and keeping temperature at 1.0. Cumulatively, our experiments take around 250 hours to run on four NVIDIA A6000 GPUs.

[6] `r/democrats` has an order of magnitude less engagement, making it hard to analyze in isolation (400k). `r/Conservative` doesn't exactly have a liberal equivalent, which can be attributed to Reddit's overall liberal leaning.

five upvotes. We remove comments with fewer than ten tokens, which rarely express attitudes. In social media (and in general), understanding the context in which a comment was posted is crucial to reconstruct the user's communicative intent faithfully. We incorporate all preceding context for decomposing a comment, considering only top-level comments and replies to top-level comments to limit the total length of each text item.

Issue Selection. We focus on two key issues discussed frequently in contemporary political literature — abortion and immigration. To obtain comments about these topics, we estimate an LDA topic model with Gibbs sampling (Griffiths and Steyvers, 2004) on data across the entirety of our dataset (2014-2022). Reddit membership has increased exponentially, so we must ensure topics are not dominated by documents that are more recent. We divide the data into 6-month slices and create a data set where each slice can contribute up to 30k comments. We train a 100-topic topic model following best practices (Hoyle et al., 2022). We then identify topics that pertain to the issue under consideration, and select comment threads whose highest document-topic probability is one of these topics.[7]

4.2 Extracting Attitude Expressions

Treating Reddit comments as text units is problematic for several reasons: they are highly contextual; expressions of political views often involve complex presuppositions or implications; and comments usually contain pronominal references. We instead extract propositions from comments using an LLM, adapting Hoyle et al.'s (2023) "inferential decompositions" method to generate natural language propositions that are explicitly or implicitly communicated by the comment. Unlike their approach, we extract only inferential decompositions that are salient with respect to the issue being investigated, and instead of generating decompositions from comments alone, we augment the prompt to condition on preceding context as needed.[8] To extract salient attitudes about a topic from a six-month window, we sample 300 of the most topical comments from each six-month time slice in our dataset. This ensures the number of data points representing each time frame remains comparable.

4.3 Using Pairwise Comparisons to Obtain an Attitude Scale

Measuring where an attitude lies on a polar scale is a challenging problem. We build on Wu et al.'s (2024) concept-guided chain-of-thought (CGCoT), which uses an LLM to generate structured summaries of text items for use in downstream pairwise comparisons. We use the LLM to compare the decomposed attitudes pairwise according to a *comparison construct* (e.g., see prompt A.3.2). Given two attitude expressions about a topic such as abortion, the model picks the one that expresses a greater opposition to abortion or speaks more in favor of anti-abortion legislation (similarly for immigration). Repeating the process over sampled attitude pairs from the dataset, we use Bradley and Terry (1952) to scale the comparison results onto a single axis, following (Wu et al., 2024). In our dataset, attitudes that express a stronger view obtain a more positive score, and ones that express a more moderate or left-accommodating view receive a more negative score. Hence, all attitudes in a community are placed on a relative scale, where the attitudes that receive the highest and lowest scores after scaling can be interpreted as the most extreme viewpoints in that specific dataset.

Figure 2 outlines our approach of pairwise comparing attitudes from the same topic. In each pairwise comparison, the model has four options: `attitude1`, `attitude2`, `tie`, and `noncomp` for decomposition pairs that the model deems non-comparable.[9] We sample pairs of attitude expressions using replacement sampling until every attitude is part of at least 50 comparisons, dropping all expressions participating in < 10 comparisons after dropping `noncomp` pairs.

5 Validation

Before we focus on our substantive contributions, we first ensure the validity of the two phases of our approach. If not validated properly, text-based approaches to quantify political views can easily be used to make unsubstantiated claims from data. We verify that attitude expressions extracted by an

[7] Further details in A.5. Topic models are not inherent to the method; other text selection techniques would suffice.

[8] We also let the model output `<Insufficient Context>` if the context and utterance are not enough to tease out the attitudes of a user about a topic and `<Not Topical>` when the comment is not about the topic.

[9] Even after topical filtering, certain decompositions in our pool express sentiments that may not be comparable. For example, the claim "Democrats are inconsistent in their moral stances" is not directly comparable to an abortion-related attitude such as "Taxpayers should not fund abortions".

LLM are reasonable and that pairwise comparisons made by the LLM correlate with human scores.

5.1 Validating Attitude Extraction

Attitude expressions extracted from Reddit comments form our linguistic unit of analysis. As such, it is necessary to ensure that these model-generated texts are plausible expressions of users' attitudes.

We sample 150 inferential decompositions from 50 comments that were not part of the selected dataset for pairwise comparisons, and judge the validity of each extracted attitude on its plausibility. Following Hoyle et al. (2023), two crowdworkers are shown comment-decomposition pairs, and must determine, on a 5-point scale, whether it is reasonable to conclude that a user who wrote the comment would also state the decomposition. 1 denoted "definitely reasonable" and 5 denoted "not reasonable at all". Along with the comment, the crowdworkers are also shown the submission text and the top-level comment, if applicable, as context to help ground their judgments. Each item is then coded with a majority vote (breaking ties with the rounded mean). Crowdworkers are also asked whether the model-generated decomposition is an underlying belief or is explicitly stated. 23 US-based crowdworkers were recruited via Prolific.[10] Each crowdworker annotated 15 items and was paid an average of 15.31 USD per hour. Median survey completion time was roughly 15 minutes. The detailed annotation instructions can be found in Appendix A.8.3.

79% of items are coded as either "definitely" or "probably" reasonable, 12% ambiguous, and 6% "probably not" reasonable. Only 4 items (3%) are deemed "not reasonable." 38% of generated items are implicit beliefs, 33% are explicit in the comment, and the rest received tied votes. As we will show, using this generated text in downstream stance comparisons improves the agreement with ground-truth annotations.

5.2 Validating the Pairwise Comparisons

Next, we validate the pairwise comparisons made by the model. We collect pairwise outcome judgments from 43 annotators on 150 pairs of comments on abortion, ensuring each pair has been annotated at least 3 times (as before, we pay $15USD/hour based on an estimated 15-minute survey of 15 questions).

[10] https://www.prolific.com/

Pairwise comparisons in our task operate over attitudes expressed as inferential decompositions from comments. To compare pairwise outcomes over comments, we first gather the outcomes of all possible pairs of attitude expressions decomposed from the two comments. In a comparison between the two comments, the majority outcome wins. If attitude expressions from both comments score an equal amount of wins, we declare a tie. We also declare a tie when the models deem them incomparable.

As a point of comparison to our approach, we recreate the comment-level method proposed by Wu et al. (2024) on comments about abortion. Wu et al. (2024) only had a win, lose, or tie as possible outcomes, so we only keep pairs where no annotator deemed the pair non-comparable while having a majority outcome. This results in 77 pairs. Our induced comment scores had a higher macro-averaged F1 score with human annotations (0.55 vs 0.50), and comparable Krippendorff's α (Krippendorff, 2004) (ours 0.53 vs. Wu et al. (2024)'s 0.61), indicating that both methods have moderate agreement with majority-vote human annotations. The annotators had an inter-rater agreement of $\alpha = 0.38$, underscoring the task's subjectivity.

6 Retroactive "Surveys"

We begin our analyses with a unique study made possible through treating attitude expressions as our unit of focus. Since these expressions in our framework are represented in natural language, they are inherently comparable with related attitude expressions across time. This grants us the ability to gauge the relative ideological position of an attitude not necessarily present in the dataset and track its position across time.

We sample questions that are either inspired or paraphrased from surveys administered by the Pew Research Center.[11] We run pairwise comparisons of these expressions with abortion-related attitude expressions extracted from comments on `r/Conservative` in each time period. While these questions were asked to human participants, surveys involving humans are expensive and time-consuming to administer. Our approach lets us

[11] The expressions in "In what circumstances should abortion be illegal" (right) was inspired from https://www.pewresearch.org/religion/fact-sheet/public-opinion-on-abortion/. The others were from https://www.pewresearch.org/religion/2022/05/06/americas-abortion-quandary/

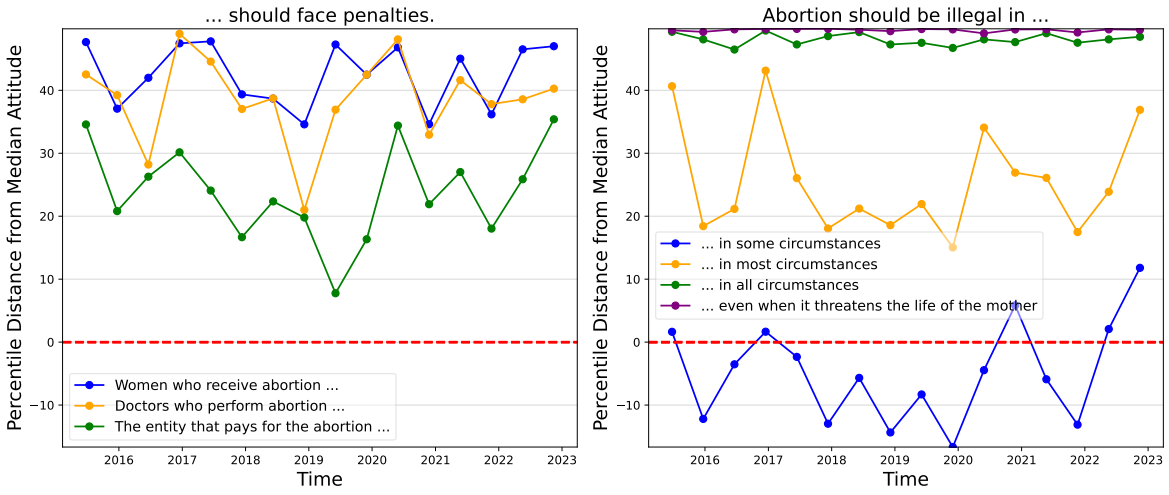


Figure 3: Retroactive "survey" of attitudes about abortion. The y-axis determines how many percentile points away an attitude expression is from the median opinion. From 2015 to 2020, the movement of expressions against abortion towards the median shows how they became relatively less "extreme" to the community.

essentially perform an opinion "survey" over the space of public attitude in a community purely through their text, where a large portion of those attitudes are implicit.

In Figure 3, the y-axis represents the distance (in percentile points) from a hypothetical "median" attitude in a time period, where half of the attitudes take a more opposing stance against abortion, and the other half takes a less opposing stance. The y-axis has been origin shifted to 50 to show that the effect goes both ways—while the attitude "Abortion should be illegal in all circumstances" (green line on the right) remains an extreme viewpoint, abortion being legal in *some* circumstances (blue) is an example of a "median" opinion. In 2020, it shifted downward, which means most other attitudes at the time were more extreme. Its score shifts away from the "median" opinion of that time in the other direction to become a relatively pro-abortion attitude. One way to visualize how this is a median opinion is that its percentile movement looks like an inverted version of abortion scores in Fig. 5.

When this value decreases, it indicates a *higher acceptability* of the attitude in the community (they are not as extreme anymore). For "The entity who pays for the abortion should face penalties", 30% of attitudes in 2017 received a lower score (its distance from the top) denoting that this was still a contentious attitude). Between 2019 and 2020, more than 40% of attitude expressions received a higher score, suggesting that this is now a relatively less extreme attitude against abortion. Overall, we see a common trend of acceptability of previously extreme opinions from 2015 to 2020.

7 Tracing Attitudes over Time

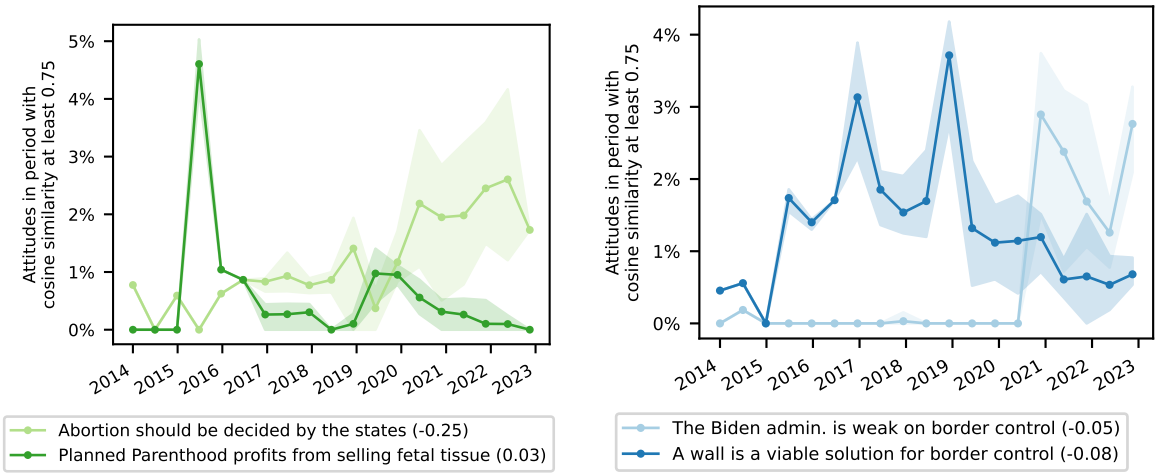
One benefit of our framework is that the decomposed attitude expressions are in simple language, making them easier to aggregate and interpret than the original text. In this section, we identify attitudes that show meaningful changes in *prevalence*: that have an increasing (or decreasing) proportion of semantically similar attitudes over time.

Hoyle et al. (2023) demonstrates that the decompositions can be reliably compared to one another with sentence embeddings (Reimers and Gurevych, 2019). For a given target attitude, we take its embedding $\boldsymbol{e}_i$ and calculate the cosine similarity to the embeddings of all other attitudes generated by the method in each time period, $\boldsymbol{s}_{i,y} = cos(\boldsymbol{e}^T, E_y) \in \mathbb{R}^n$. To operationalize attitude prevalence, we compute the proportion of attitudes with similarity above a threshold h (here 0.75).[12]

Figure 4 presents the changes in four target expressions within `r/Conservative`. These expressions were selected by considering those connected to known external narratives, as well as those with strong linear trends over the period.

In the case of immigration (fig. 4b), changes in prevalence appear to align with the U.S. election cycle. "A wall is a viable solution for border control" sees an increasing similarity to other attitudes starting in mid-2015 with the announcement

[12]We use the `all-mpnet-base-v2` and set h by reviewing the qualitative semantic similarity of random pairs at various thresholds; similar trends occur with alternative thresholds (± 0.05) and encoders (`gte-large-en-v1.5`, Li et al. 2023).



(a) Abortion Topic

(b) Immigration Topic

Figure 4: Changes in the prevalence of target attitudes over time. Prevalence is defined as the proportion of other attitude embeddings with a similar (≥ 0.75) cosine similarity to the target. (Numbers in parentheses are PAIRSCALE scores. Bands represent the range of values over five comment samples; smaller bands in earlier periods are due to a smaller pool of comments).

of Donald Trump's candidacy (with the unofficial slogan "build the wall"), building to a peak in late 2016 during the election, falling, then rising again during the next election in 2020. Interestingly, this discussion is almost immediately supplanted by a critique of then-president Joe Biden's border policy in late 2021.

With respect to abortion (fig. 4a), we discover an increase in attitudes favoring state control of abortion law. This increase in prevalence coincides with challenges to abortion rights before the U.S. Supreme Court starting in 2020 and culminating with the overruling of Roe v. Wade in 2022—the attitude has a negative PAIRSTANCE score (-0.25), representing a moderating stance that downplays the impact of the decision.

Interestingly, there is also a strong *decrease* in prevalence among a collection of related attitudes expressions about one specific topic: the (false) claim that "Planned Parenthood profits from selling fetal tissue". These are the clear results of a doctored video and misinformation campaign from 2015 designed to discredit Planned Parenthood (Damann, 2018; Coker, 2023), which appear to have successfully impacted the attitudes of `r/Conservative` members in 2015. However, its effect diminishes over time (or it was absorbed as common "knowledge" that goes unexpressed). As further validation, a dynamic topic model (Blei and Lafferty, 2006) also uncovers terms related to this campaign (see appendix A.7 for details).

8 Quantifying Attitude Shifts

Comparing topical attitudes represented in natural language enables us to focus on ideological shifts across time. We construct a collection of abortion-related attitude expressions from the 300 comments sampled from each time period for both `r/Conservative` and `r/democrats`, eliciting pairwise comparisons from an LLM. To comment on some particularly salient attitude shifts, we consider time slices four years apart, to account for the natural variation from the election cycle.

Abortion. Our experiments suggest a shift towards more extreme viewpoints towards the second half of 2019 and a reduction in scores from 2018 to 2022. Figure 5 shows a visual representation of this shift through movement of the median score. For abortion, "moderate" attitudes center around the belief that abortion is an individual's right; and that there should be limits to abortion rather than a total ban. We notice that the number of moderate attitudes about abortion reduced in both number and intensity in 2019.

More surprising is the reduction of the mean score in 2022. In the wake of the overturning of Roe v. Wade, rather than an *increase* in extreme views, we see opposition to an outright federal ban, as evidenced by the overlapping interquartile scores with `r/democrats` attitudes in Fig. 5. The main themes of opposition tend to have a libertarian bent: they are centered around government overreach,

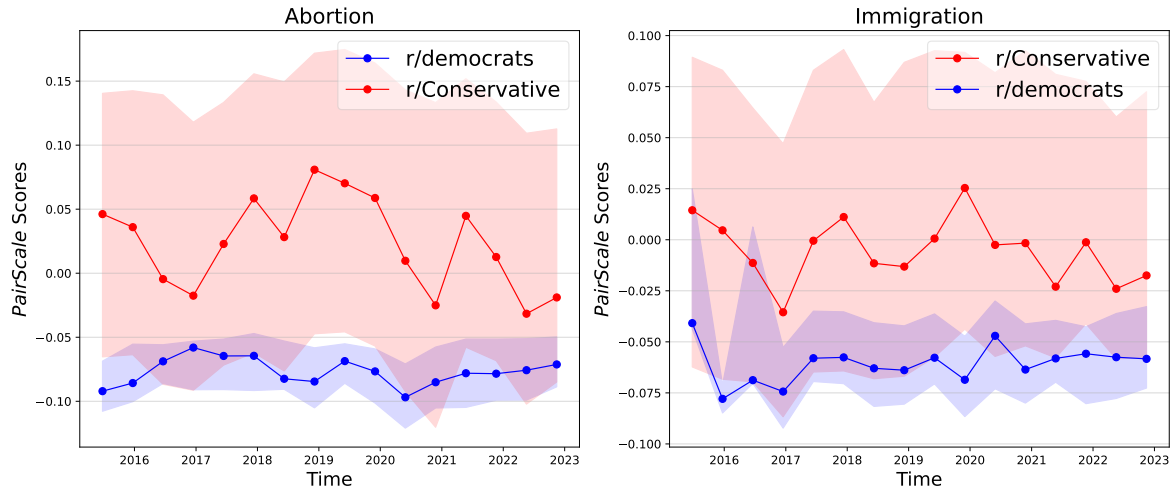


Figure 5: Vizualising the shift in median attitude score over time across two topics - abortion and immigration in two communities, `r/democrats` and `r/Conservative`. The bands mark the interquartile range of PairStance scores obtained by attitudes in that time period. There is a higher overlap in immigration.

and the fact that setting abortion laws should be under the purview of individual states.

More generally, our method enables us to study attitudes about the topic in a fine-grained manner. Scores help to contextualize attitudes by quantifying their ideological position on a scale that is defined by the most extreme views in that period. Attitudes on the extreme right of the scale take moral issue with abortion, opposing the practice at all periods, e.g., "Pro-abortion individuals will face divine punishment". However, there are some points of overlap between the two subreddits: "Abortion bans can cause immense harm and suffering", an attitude expressed within `r/democrats`, falls within the 25th percentile of PAIRSCALE scores from `r/Conservative` in some periods.

Immigration. In immigration, we see relative overall stability except for an increase in 2020. Attitudes contributing to this increase centered around dissatisfaction with the Biden administration's management of immigration and border issues. Furthermore, there was strong disapproval regarding reports that illegal immigrants were receiving COVID-19 stimulus checks and other benefits, which were perceived to be facilitated by Democratic policies.

Our results demonstrate that the two issues we study are nuanced and multi-faceted, even within a seemingly like-minded online community, and cannot be reduced to a binary/ternary stance.

9 Conclusion

We have introduced a framework for measuring attitude expressions about an issue in an online community on a continuous scale, validating key components using human judgments. Using this framework we characterize dynamics and changes in political subreddits over a key time period for U.S. politics. Our framework allows practitioners such as political scientists to keep a pulse on shifting political discourse, and to discover emerging extreme views on a range of issues. The approach, validation, and substantive illustration of its utility lay the foundation for further attitude studies using "retroactive surveys", and the flexibility of the LLM pairwise-comparison method creates potential for wider applications that go beyond for-against scales.

Although we have first demonstrated and validated PAIRSCALE in a study of community attitudes, the method was conceived with broader applications in mind. Currently, we are working on its applications in mental health, where, in addition to making it possible to quantify and compare relevant constructs across communities, e.g. stress in communities that have lesser and greater access to mental health providers (Hoffmann et al., 2023), the method also offers a new way to make more effective use of limited resources, for example, ranking a population of individuals in treatment for schizophrenia by the degree of evidence for psychotic symptomatology to inform intervention strategies (Kelly et al., 2021).

Limitations

While the approach outlined in our paper offers valuable insights into the movement of conservative attitudes on Reddit along with comments from `r/democrats` as an anchoring point, the study is limited to sampled comments from a limited time period—2015 to 2022. Despite our validation (and the similarities of our findings with Waller and Anderson 2021), it is possible that different samples or modeling choices could produce different results. We outline some potential issues below.

LLM Bias. LLM biases (due to either pretraining data or alignment) could potentially affect multiple points of the process. When generating attitudes, LLMs may sanitize or ignore extreme positions—e.g., on the abortion topic, we had expected to see more attitudes taking issue with womens' sexual freedom based on the raw data, but it is possible that models are unlikely to generate attitudes relating to that topic. The automated pairwise comparisons could also face similar problems if some topics are deemed systematically more extreme than others—one could argue that attitudes relating to "Planned parenthood profits from selling fetal tissue" are more extreme than "Abortion takes the life of a human being", as the former connects to a fringe (and false) belief about malignant actors perpetrating a broad conspiracy, as opposed to the latter, which merely expresses a commonly-held moral stance. We try to mitigate this risk by conducting human validation, but it cannot be eliminated. We additionally find that certain LLMs (such as LLAMA-3.1-8B-INSTRUCT) were reluctant to compare extreme or offensive attitudes about contentious issues.

Exogenous Factors. Importantly, we do not know the extent to which `r/Conservative` is representative of conservatives in the U.S. (or globally). In fact, it is likely that it differs in important ways: while some studies have indicated its American users are broadly representative of the U.S. population, albeit more liberal (Shatz, 2017), it is not clear whether that is the case for `r/Conservative` in particular, nor whether it remains the case today. The underlying population of `r/Conservative` is also rapidly evolving, as we show in Figure 7 (in the Appendix), which adds a layer of complexity in interpreting the shift of PAIRSCALE scores. In Figure 6 in the Appendix, we try to quantify the influence of new users.

As such, our study should be understood as an investigation of this particular conservative community, rather than conservatives in general—although we hope our findings can serve as a jumping-off point for the analysis of other groups.

Acknowledgments

We thank our anonymous reviewers for their very helpful comments, and for engaging with us throughout the review process. We also thank Neha Srikanth for her feedback on an earlier version of the draft. This work was supported in part by the U.S. National Science Foundation award 2124270. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

References

Robert Adcock and David Collier. 2001. Measurement validity: A shared standard for qualitative and quantitative research. *American Political Science Review*, 95(3):529–546.

Emily Allaway and Kathleen McKeown. 2020. Zero-Shot Stance Detection: A Dataset and Model using Generalized Topic Representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8913–8931, Online. Association for Computational Linguistics.

Karyn Amira. 2022. Donald trump's effect on who is considered "conservative". *American Politics Research*, 50(5):682–693.

Elliott Ash, Germain Gauthier, and Philine Widmer. 2024. Relatio: Text semantics capture political and economic narratives. *Political Analysis*, 32(1):115–132.

L.R. Atkeson and R.M. Alvarez. 2018. *The Oxford Handbook of Polling and Survey Methods*. Oxford Handbooks. Oxford University Press.

David Bamman and Noah A. Smith. 2015. Open extraction of fine-grained political statements. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 76–85, Lisbon, Portugal. Association for Computational Linguistics.

Kenneth Benoit, Kevin Munger, and Arthur Spirling. 2019. Measuring and explaining political sophistication through textual complexity. *American Journal of Political Science*, 63(2):491–508.

George F Bishop, Robert W Oldendick, and Alfred J Tuchfarber. 1978. Effects of question wording and format on political attitude consistency. *Public Opinion Quarterly*, 42(1):81–92.

David M. Blei and John D. Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 113–120, New York, NY, USA. Association for Computing Machinery.

Emma Borg and Sarah A. Fisher. 2021. Semantic content and utterance context: a spectrum of approaches. In Piotr Stalmaszczyk, editor, *The Cambridge Handbook of the Philosophy of Language*, Cambridge Handbooks in Language and Linguistics. Cambridge University Press.

Ralph Allan Bradley and Milton E. Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

David Carlson and Jacob M. Montgomery. 2017. A pairwise comparison framework for fast, flexible, and reliable human coding of political texts. *American Political Science Review*, 111(4):835–843.

Santiago Castro. 2017. Fast Krippendorff: Fast computation of Krippendorff's alpha agreement measure. https://github.com/pln-fing-udelar/fast-krippendorff.

Xi Chen, Paul N. Bennett, Kevyn Collins-Thompson, and Eric Horvitz. 2013. Pairwise ranking aggregation in a crowdsourced setting. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, page 193–202, New York, NY, USA. Association for Computing Machinery.

Martijn Schoonvelde Christian Baden, Christian Pipal and Mariken A. C. G van der Velden. 2022. Three gaps in computational text analysis methods for social sciences: A research agenda. *Communication Methods and Measures*, 16(1):1–18.

Calvin R. Coker. 2023. "do you think this is not happening?": Rhetorical laundering and the federal hearings over planned parenthood. *Women & Language*, 46(1):1–18. University of Louisville, Department of Communication.

Iain J. Cruickshank and Lynnette Hui Xian Ng. 2024. Prompting and fine-tuning open-sourced large language models for stance classification.

Taylor Damann. 2018. Project veritas and the changing face of fake news. *Gateway Journalism Review*, 47(351).

Michael Dimock, Carroll Doherty, and Jocelyn Kiley. 2013. Growing support for gay marriage: Changed minds and changing demographics. *Gen*, 10:1965–1980.

Sean Gerrish and David Blei. 2012. How they vote: Issue-adjusted models of legislative behavior. *Advances in neural information processing systems*, 25.

Lukas Gienapp, Benno Stein, Matthias Hagen, and Martin Potthast. 2020. Efficient pairwise annotation of argument quality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5772–5781, Online. Association for Computational Linguistics.

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):e2305016120.

Kyle Glandt, Sarthak Khanal, Yingjie Li, Doina Caragea, and Cornelia Caragea. 2021. Stance detection in COVID-19 tweets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1596–1611, Online. Association for Computational Linguistics.

G. Goertz. 2020. *Social Science Concepts and Measurement: New and Completely Revised Edition*. Princeton University Press.

Thomas L. Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl_1):5228–5235.

Jennifer A Hoffmann, Megan M Attridge, Michael S Carroll, Norma-Jean E Simon, Andrew F Beck, and Elizabeth R Alpern. 2023. Association of youth suicides and county-level mental health professional shortage areas in the us. *JAMA pediatrics*, 177(1):71–80.

Thomas Holtgraves and Katie Bray. 2023. Us liberals and conservatives live in different (linguistic) worlds: Ideological differences when interpreting business conversations. *Journal of Applied Social Psychology*, 53(7):674–683.

Alexander Hoyle, Pranav Goel, Denis Peskov, Andrew Hian-Cheong, Jordan Boyd-Graber, and Philip Resnik. 2024. Is automated topic model evaluation broken? the incoherence of coherence. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA. Curran Associates Inc.

Alexander Hoyle, Rupak Sarkar, Pranav Goel, and Philip Resnik. 2023. Natural language decompositions of implicit content enable better text representations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13188–13214, Singapore. Association for Computational Linguistics.

Alexander Miserlis Hoyle, Rupak Sarkar, Pranav Goel, and Philip Resnik. 2022. Are neural topic models broken? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5321–5344, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Deanna L Kelly, Max Spaderna, Vedrana Hodzic, Glen Coppersmith, Shuo Chen, and Philip Resnik. 2021. Can language use in social media help in the treatment of severe mental illness? *Current research in psychiatry*, 1(1):1.

Ashiqur R. KhudaBukhsh, Rupak Sarkar, Mark S. Kamlet, and Tom M. Mitchell. 2020. We don't speak the same language: Interpreting polarization through machine translation. *Preprint*, arXiv:2010.02339.

Natalia Knoblock. 2020. Silent majority or vocal minority: A corpus-assisted discourse study of trump supporters' facebook communication. *Open Library of Humanities*, 6(2):1–37.

Klaus Krippendorff. 2004. *Content Analysis: An Introduction to Its Methodology*, 2nd edition. Sage Publications, Thousand Oaks, CA.

Jon A Krosnick and Matthew K Berent. 1993. Comparisons of party identification and policy preferences: The impact of survey question format. *American Journal of Political Science*, pages 941–964.

Yanna Krupnikov and Blake Findley. 2018. 483Survey Experiments: Managing the Methodological Costs and Benefits. In *The Oxford Handbook of Polling and Survey Methods*. Oxford University Press.

Benjamin E Lauderdale and Tom S Clark. 2014. Scaling politically meaningful dimensions using texts and votes. *American Journal of Political Science*, 58(3):754–771.

Verlan Lewis. 2021. The problem of donald trump and the static spectrum fallacy. *Party Politics*, 27(4):605–618.

Yingjie Li, Chenye Zhao, and Cornelia Caragea. 2021. Improving stance detection with multi-dataset learning and knowledge distillation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6332–6345, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

Adam M Mastroianni and Jason Dana. 2022. Widespread misperceptions of long-term attitude change. *Proceedings of the National Academy of Sciences*, 119(11):e2107260119.

Keith T. Poole and Howard Rosenthal. 1985. A spatial model for legislative roll call analysis. *American Journal of Political Science*, 29(2):357–384.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Itamar Shatz. 2017. Fast, free, and targeted: Reddit as a source for recruiting participants online. *Social Science Computer Review*, 35(4):537–549.

Sooahn Shin. 2024. *Measuring Issue Specific Ideal Points from Roll Call Votes*. Ph.D. thesis, Harvard University. Ph.D. Candidate, Department of Government and Institute for Quantitative Social Science.

Umme Aymun Siddiqua, Abu Nowshed Chy, and Masaki Aono. 2019. Tweet stance detection using an attention based neural ensemble model. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1868–1873, Minneapolis, Minnesota. Association for Computational Linguistics.

Isaac Waller and Ashton Anderson. 2021. Quantifying social organization and political polarization in online platforms. *Nature*, 600(7888):264–268.

Patrick Y. Wu, Jonathan Nagler, Joshua A. Tucker, and Solomon Messing. 2023. Large language models can be used to estimate the latent positions of politicians. *Preprint*, arXiv:2303.12057.

Patrick Y. Wu, Jonathan Nagler, Joshua A. Tucker, and Solomon Messing. 2024. Concept-guided chain-of-thought prompting for pairwise comparison scoring of texts with large language models. In *2024 IEEE International Conference on Big Data (BigData)*, pages 7232–7241.

A Appendix

A.1 Impact of New Users

We discuss the impact of exogenous factors, including new users in Section 9. The flow of new users can be found in Figure 7. In Figure 6, we show how PAIRSCALE scores across time have been affected by attitude expressions from new and existing users. While there's a clear distinction between new and existing users in the case of abortion (new users being less opposed to it), no such pattern exists for immigration.

A.2 Adapting Inferred Decompositions from Hoyle et al. (2023)

In our paper, attitude expressions, following Hoyle et al. (2023), are natural language statements inferred from the surface content of the text item, and capture both implicit and explicit propositions connected with the author's communicative intent. To adopt their example, if somebody expresses "Federal lands and waters should be protected from fossil fuel extraction", then an inferred attitude might be "Preserving natural resources for future generations is important". However, our approach to

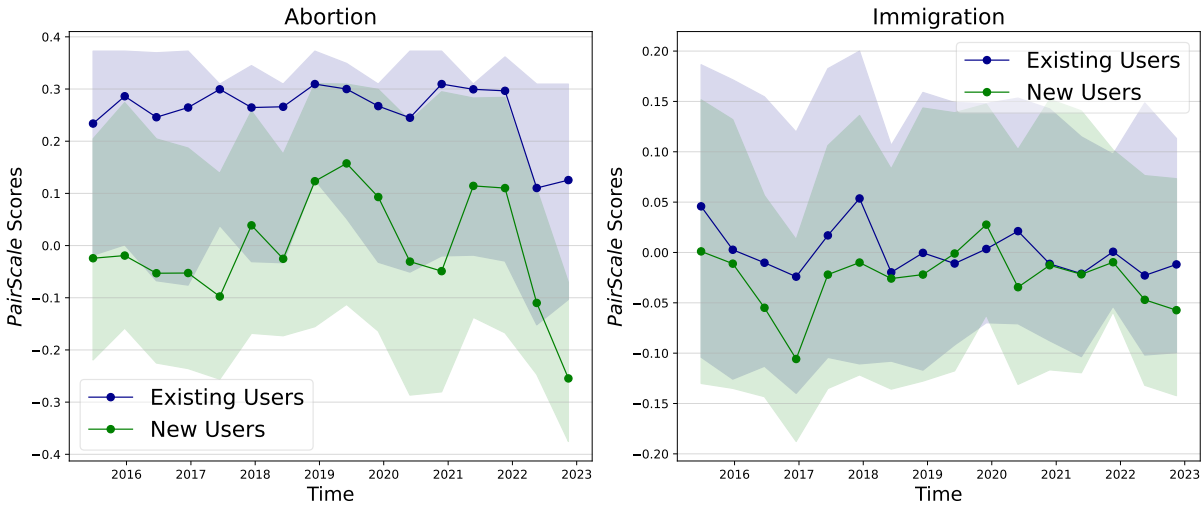


Figure 6: Figure showing how PAIRSCALE scores for two issues across r/Conservative varied for new and existing users. Existing users seem to have stronger attitudes against abortion than new users, while there's no such clear pattern in the case of immigration.

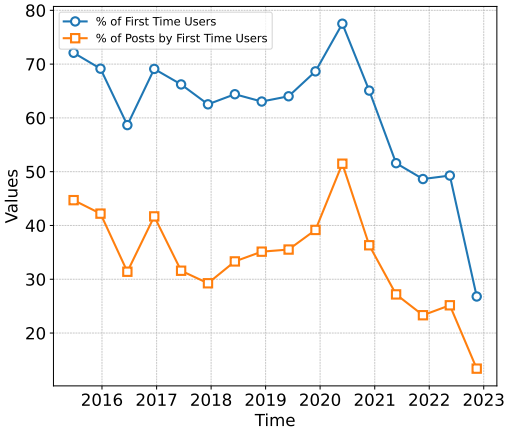


Figure 7: Figure showing the percentage of users in a six-month slice who have never engaged in r/Conservative before (in blue). The orange line shows the percentage of posts created by new users at each of these periods.

extracting attitude expressions differs from Hoyle et al. (2023) in the following key ways-

1. Unlike their approach of capturing general inferences from the text, we only extract decompositions that are salient with respect to the issue being investigated.

2. While they extract decompositions from a single piece of text, we undertake the more challenging task of doing a contextual decomposition - extracting decompositions from $text_1$ (the target text) based on $text_2$ (preceding comments in the thread) as context.

3. We allow the model to output <Insufficient Context> if the context is not enough to tease out the attitudes of the user, and Not Topical if the target text is not about the topic.

A.3 Prompts

A.3.1 Prompt for Decomposition

Prompt For Decomposition: Immigration

```
Human utterances communicate attitudes that may or may not be explicit in the literal meaning of the utterance. Given a **comment** made in Reddit in reply to a post or another comment (together called: **context**), state all the salient subjective attitudes expressed in that comment **about immigration or border control**. Implicit attitudes might be implied from the comment depending on the **context** or may be embedded as assumptions or implications made by the user through their comment. All generated attitudes should be short, and written in direct speech and simple sentences.

NOTE: If there's a reference to a person or an object in the comment, use only the given **context** to resolve that reference rather than making a guess. If there is insufficient context to output an attitude, only output <Insufficient context>. If the comment or context is not at all about immigration, border control or related topics (even indirectly or implicitly), just output <Not Topical>.

These comments were taken from the r/democrats subreddit. Output only the subjective attitudes or beliefs about immigration in a list of independent propositions from the perspective of the commenter. Use the format below:

###############

Context: {context}
Comment: {comment}

Attitudes:
```

A.3.2 Prompt Used for Pairwise Comparisons

Prompt For Pairwise Comparison: Immigration

```
Given two **attitudes** about immigration, select the **attitude** that is more against immigration. Opinions against immigrations focus on the negative impact of immigration on society, culture, or economy. If both attitudes are against immigration equally, return the attitude that holds a more negative view of immigration. If the two **attitudes** are semantically same or one is as anti-immigration as the other, output TIE. If the two attitudes are not comparable, or at least one of them is not a subjective opinion about immigration, output NONCOMP. Choose only from ATTITUDE1, ATTITUDE2, NONCOMP or TIE.

###############

ATTITUDE1: {attitude1}
ATTITUDE2: {attitude2}
ANS:
```

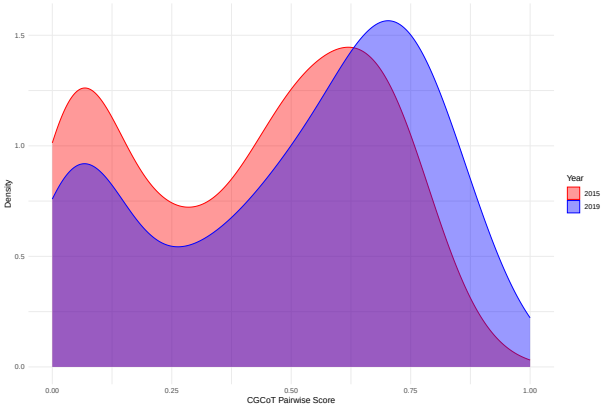


Figure 8: We recreate the ideological shift on abortion from 2015 to 20195 by adapting Wu et al.'s (2024) approach with a similar topical construct focusing on abortion. While we see our substantive finding rediscovered, comparisons over comments using this method are unable to track individual attitudes or perform retroactive surveys.

A.4 Further comparison with Wu et al. (2024)

A.5 Topic Model Details

The optimum hyperparameters for running a topic model in our dataset were taken from (Hoyle et al., 2024). In our preprocessing step, we used the following hyperparameters from the publicly available repo accompanying Hoyle et al. (2024):

1. We lowercased all the text items.
2. The max vocab was chosen to be 100k
3. The minimum size of a document was chosen to be 5 tokens
4. Each token needed to have 2 characters minimum
5. We joined commonly occurring entities into a single term

A.6 Consistency of Trends

We sample 300 comments from each time period and report our trends and comparisons with `r/democrats` in Figure 5 based on a single sample. As a robustness check, we report the variation in PAIRSCALE scores for attitudes generated from separate comment samples in Figure 9; the scores are highly stable and the trends remain consistent.

A.7 Comparison with Dynamic Topic Models

Dynamic Topic Models (DTM, Blei and Lafferty, 2006) is another way of capturing the shift of

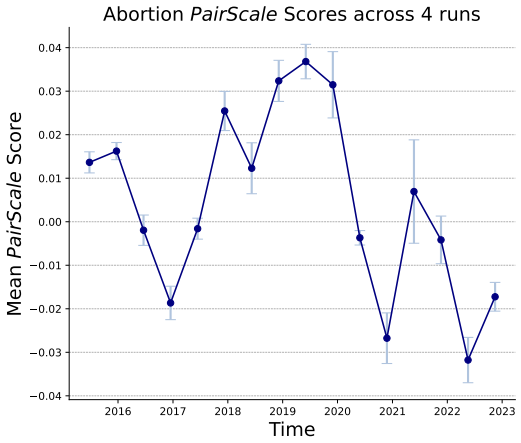


Figure 9: Consistency of attitude scores per time period across 4 samples. The line represents the mean score and the error bars represent the standard deviation.

a topic over time. We train a Dynamic Topic Model (implemented in `tomotopy`[13]) on comments about abortion through all time slices and note that it also discovers a "Planned Parenthood" topic, which in 2015 consists of words that hint at the misinformation-fueled narrative about Planned Parenthood selling baby parts that we uncovered in Figure 4(a). The top words were - "`planned`", "`parenthood`", "`baby`", "`parts`", "`aborted`", "`fetal`", "`video`", "`selling`", "`videos`". However, while the topic words convey the emergence of new aspects, they don't inform us how comments on these new attitudes are received in a particular period. Our PAIRSCALE scores associated with these attitudes tell us that it was a very well-accepted (even a moderate view) among Conservatives at the time.

A.8 Annotation Details

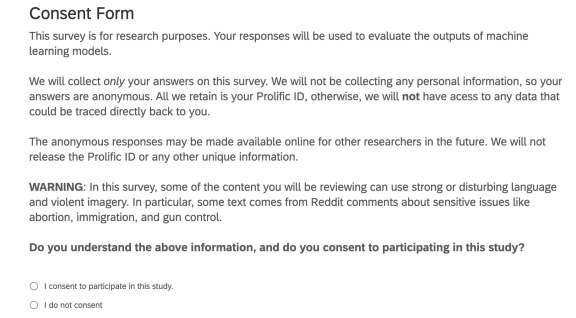
A.8.1 Consent for Data Collection

The consent form shown to all annotators is shown in Figure 10.

A.8.2 Judging Plausibility of Extracted Attitudes

For this task, we used the instructions outlined in Figure 11. An example of the kind of question posed to the annotator can be found in Figure 12, and we used a Likert rating scale shown in Figure 13.

[13] `https://bab2min.github.io/tomotopy`


Consent Form

This survey is for research purposes. Your responses will be used to evaluate the outputs of machine learning models.

We will collect *only* your answers on this survey. We will not be collecting any personal information, so your answers are anonymous. All we retain is your Prolific ID, otherwise, we will **not** have acess to any data that could be traced directly back to you.

The anonymous responses may be made available online for other researchers in the future. We will not release the Prolific ID or any other unique information.

WARNING: In this survey, some of the content you will be reviewing can use strong or disturbing language and violent imagery. In particular, some text comes from Reddit comments about sensitive issues like abortion, immigration, and gun control.

Do you understand the above information, and do you consent to participating in this study?

○ I consent to participate in this study.
○ I do not consent

Figure 10: Consent form shown to annotators before each annotation task.

A.8.3 Details of Validating Pairwise Comparisons

For validating pairwise comparisons, a screenshot of our annotation instructions can be found in Figure 14, with a screenshot of the survey containing an example pair in Figure 15

A.9 Details of Other Packages Used

1. We use the Python package choix[14] to run Bradley-Terry. We use the regularization parameter $\alpha = 0.1$ recommended by the author of the package.

2. We use the fast-krippendorff (Castro, 2017) package[15] to compute Krippendorff's alpha.

[14] `https://pypi.org/project/choix/`
[15] https://github.com/pln-fing-udelar/fast-krippendorff

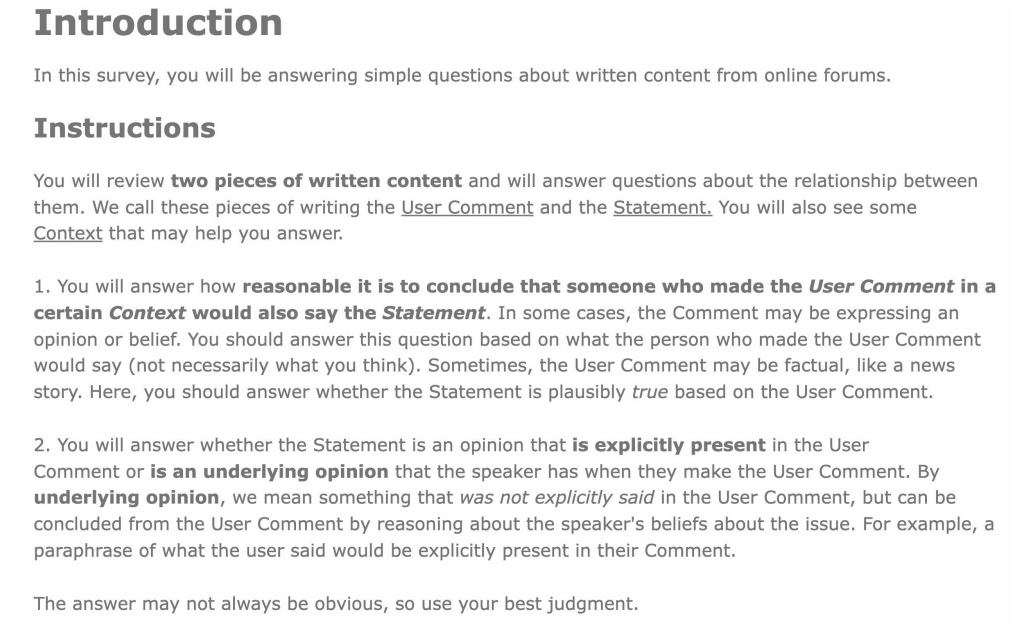


Figure 11: Instructions provided to the annotators for judging the validity of a decomposition from a given comment generated using LLAMA-3.1-70B-INSTRUCT

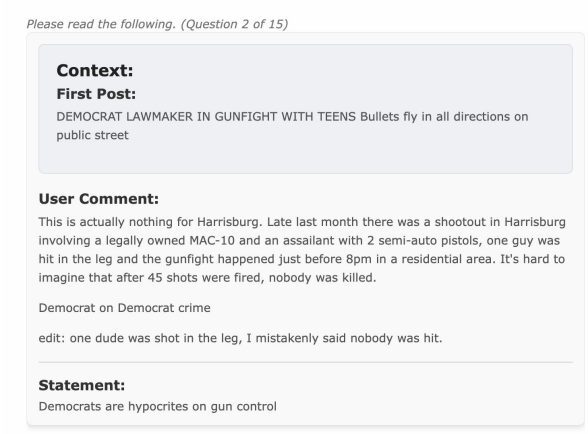


Figure 12: Example of the comment-decomposition pair shown to crowdworkers.

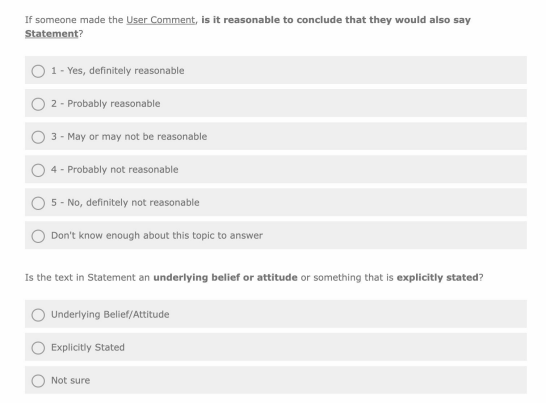


Figure 13: Rating scale displayed to crowdworkers to judge validity of comment-decomposition pairs.

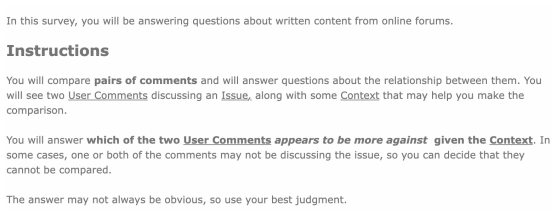


Figure 14: Instruction given to annotators to judge the outcome of a pairwise comparison (between two Reddit comments).

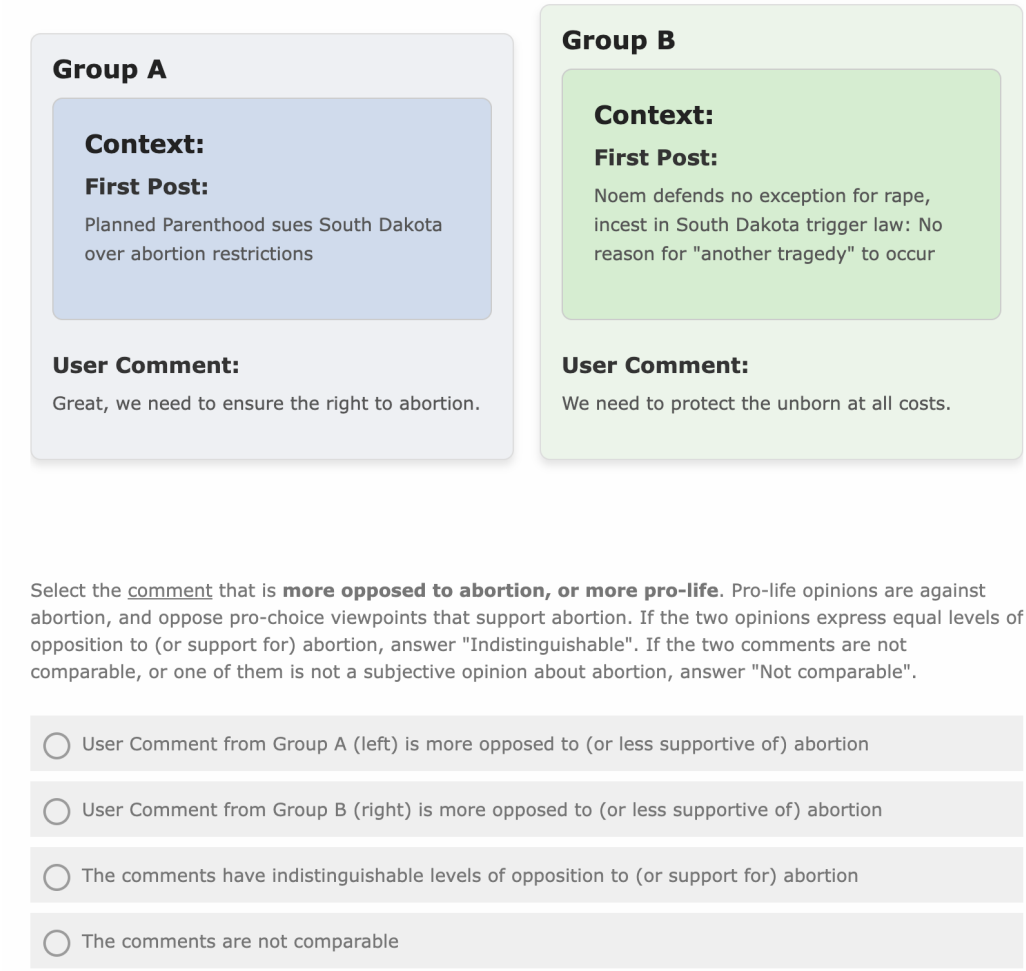


Figure 15: Example of a pairwise comparison task given to a human annotator: Given two Reddit comments with some context, the task is to choose which comment is more *against* abortion.