Study 1 Supplementary Material

Table of Content

Descriptive Statistics of the Main Tweet Dataset

The following tables provide descriptive statistics for the abortion tweet dataset used in Study 1a. Table 1 summarizes the yearly distribution of tweets across stances. Table 2 presents the distribution of moral versus non-moral tweets by stance. Table 3 reports the frequency of each moral foundation expressed in tweets, grouped by stance.

The total number of tweets included in year-level analyses is slightly lower than the overall dataset size. This discrepancy arises because a small number of tweets (n = 803; 0.01%) lacked valid timestamp metadata and could not be assigned to a calendar year. These tweets were excluded from year-based analyses but retained for analyses that did not require temporal information.

Table S1: Counts of Each Stance by Year

Year	Pro-choice	Pro-life	Neutral	Throw Out
2015	143,163	334,970	118,597	18,025
2016	92,125	216,881	73,157	8,959
2017	104,068	257,394	43,404	11,446
2018	104,744	246,298	32,813	15,223
2019	165,101	338,119	40,702	17,484
2020	84,274	182,622	20,315	22,322
2022	3,584,087	568,083	1,159,897	2,803,861
2023	30,795	11,140	7,687	117,672

Table S2: Counts of Moral vs. Non-moral Tweets by Stance

Stance	Moral	Non-moral
Pro-choice	2,705,136	1,603,374
Pro-life	1,165,214	990,541
Neutral	165,675	1,331,220
Throw Out	832,016	2,183,055

Table S3: Counts of Each Moral Foundation by Stance

Stance	**Care**	**Purity**	**Loyalty**	**Authority**	**Fairness**
Pro-choice	2,080,495	435,688	778,661	329,426	2,191,213
Pro-life	1,008,099	476,478	177,186	179,090	488,974
Neutral	98,411	25,505	72,127	9,626	117,409
Throw Out	742,063	180,658	109,707	235,287	396,064

Inter-Annotator Reliability

The following table presents the inter-annotator reliability scores for each pair of annotators involved in the manual coding of tweets for abortion stance classification. We calculated two reliability metrics: Gwet's AC1 and Brennan-Prediger (PABAK). These metrics were selected for their robustness in handling category imbalance and their ability to adjust for chance agreement, making them more appropriate for our annotation task than traditional measures such as Cohen's kappa.

The results demonstrate consistently high reliability across all annotator pairs. Gwet's AC1 scores ranged from 0.796 to 0.970, with a mean value of 0.896, while Brennan-Prediger scores ranged from 0.769 to 0.956, with a mean of 0.871. The 95% confidence intervals indicate statistical significance for all calculated reliability estimates. The particularly strong agreement between annotators B and D (AC1 = 0.970, PABAK = 0.956) suggests exceptional consistency in their application of the annotation guidelines. Overall, these metrics indicate strong inter-annotator reliability, supporting the validity of our manually annotated dataset as a foundation for training our stance classification models.

Table S4: *Inter-rater Reliability Scores for Each Pair of Annotators*

	Gwet's AC1		Brennan-Prediger (PABAK)	
Annotators	Estimate	95% CI	Estimate	95% CI
A & B	.870	[.720, 1.000]	.859	[.702, 1.000]
A & C	.916	[.818, 1.000]	.882	[.749, 1.000]
A & D	.796	[.641, .956]	.769	[.594, .945]
C & B	.934	[.841, 1.000]	.892	[.784, 1.000]
C & D	.892	[.784, 1.000]	.867	[.738, .995]
B & D	.970	[.910, 1.000]	.956	[.867, 1.000]

Abortion Stance Annotations and Model Development

Manual annotations and label distribution

We manually annotated 5,625 tweets drawn from two sources to train the stance classifier: 2,248 tweets from Dataset 2 (2015–2020) and 3,377 tweets from Dataset 1 (the Chang et al. (2023) corpus). Label counts by source appear in Table S5.

Table S5: Label counts for the manually annotated abortion-stance dataset by source.

Dataset	Pro-choice	Pro-life	Neutral	Throw out	Total
Dataset 1 (Chang et al., 2023) subset	1,748	621	521	487	3,377
Dataset 2 (2015–2020) subset	852	1,007	278	111	2,248
Combined	**2,600**	**1,628**	**799**	**598**	**5,625**

Model development details and early results

As annotation for abortion stance labels progressed, we fine-tuned three pre-trained transformer models on our stance labels. We began with RoBERTa (Liu et al., 2019), a strong general-purpose baseline for short text. In this early round (n = 2,388), to balance classes we merged *Neutral* and *Throw out* into a single non-stance label, split the data 60%/20%/20% (train/validation/test), and obtained a macro F1 of 0.67 (±0.022). We next fine-tuned

SsciBERT (Shen et al., 2023), a BERT (Devlin et al., 2019) model variant further pre-trained on social-science literature, on 3,192 annotated tweets using the same 3-label setup and an 80%/10%/10% split; because the data were imbalanced, we report both macro F1 = 0.662 (±0.021) and weighted F1 = 0.781 (±0.017), which did not improve over RoBERTa. Finally, we used BERTweet (Nguyen et al., 2020), a model pre-trained on large-scale Twitter data and tailored to short, informal posts, trained on the full annotated set (n = 5,625) with all four labels and an 80%/10%/10% split. BERTweet achieved the best performance (macro F1 = 0.758, ±0.009 and weighted F1 = 0.77, ±0.011), and we used it to label the full 10,976,231-tweet corpus.

Metric notes. Macro F1 gives equal weight to each class; weighted F1 accounts for class frequencies. We report means ± standard deviations across random seeds where applicable.

Comparing Fine-Tuned Models with Large Language Models

We compared our best fine-tuned classifiers with two recent large language models (LLMs)- the open-source Llama 3.1 (Grattafiori et al., 2024) and the closed-source reasoning model GPT o4-mini (OpenAI, 2025). For this task, we manually annotated 500 tweets and evaluated three tasks: (i) binary morality detection, (ii) Moral Foundations Theory (MFT) multilabel classification, and (iii) four-class abortion-stance classification. For the morality pipeline, *all models* first performed the binary classification; tweets predicted as *moral* were then passed to the multilabel MFT classifier for foundation tagging. To minimize invalid outputs, LLM prompts specified a JSON schema and included short examples; the full prompts appear in Figure S1, Figure S2, and Figure S3. Results are summarized in Table S6, Table S7, and Table S8.

Table S6: Binary morality classification performance (500 tweets).

Model	**Accuracy**	**Precision**	**Recall**	**F1**
Fine-Tuned RoBERTa	.712	.943	.624	.751
GPT-o4-mini	**.724**	.895	**.684**	**.775**
Llama-3.1-8B-Instruct	.566	**.912**	.417	.572

Table S7: Multiclass abortion-stance classification performance (500 tweets).

Model	**Accuracy**	**Macro F1**	**Weighted F1**
Fine-Tuned bertweet-base	.802	.559	.784
GPT-o4-mini	**.864**	**.732**	**.873**
Llama-3.1-8B-Instruct	.826	.683	.835

What the results show. Across tasks, our fine-tuned models are competitive with strong LLM baselines. In the *binary morality* stage (Table S6), GPT-o4-mini attains the highest F1 (0.775), but the margin over our RoBERTa (0.751) is modest (ΔF1 $\approx$ 0.024). In the downstream *multilabel MFT* tagging (Table S8), our RoBERTa fine-tune leads most metrics (e.g., $F_{\text{micro}} = 0.662$). For *abortion stance* classification (Table S7), GPT-o4-mini ranks best (macro F1 0.732), with Llama 3.1 second; here the gap to our RoBERTa is larger, suggesting instruction-tuned LLMs currently have an edge on nuanced stance judgments. Overall, where our model is not the top performer (especially the binary morality task), the differences from the best model are small; the notable exception is stance classification, where GPT-o4-mini shows a clearer advantage.

Table S8: Multilabel Moral Foundations (MFT) tagging performance.

Model	Subset Acc.	$\mathbf{P}_{\text{micro}}$	$\mathbf{R}_{\text{micro}}$	$\mathbf{F}_{\text{micro}}$	$\mathbf{P}_{\text{macro}}$	$\mathbf{R}_{\text{macro}}$	$\mathbf{F}_{\text{macro}}$
Fine-Tuned RoBERTa	**.518**	.630	**.697**	**.662**	.504	**.557**	**.520**
GPT-o4-mini	.426	**.671**	.497	.571	**.601**	.471	.508
Llama-3.1-8B-Instruct	.326	.454	.532	.490	.440	.541	.445

Prompt for Binary Morality Classification

```
You are an expert in moral psychology and discourse analysis. Your task is
to analyze a tweet about abortion and determine whether the tweet expresses
a moral concern (i.e., appeals to right/wrong, good/bad, justice, duty,
loyalty, harm, purity, etc.) or not.
Output format: a JSON object with exactly one key "label" whose value is
either "moral" or "non-moral".
Label guide
- moral: Appeals to values such as harm, fairness, justice, loyalty,
obedience, sanctity, or other moral principles.
- non-moral: Informational, emotional, or opinion-based without referencing
a moral concern.
Examples
1) My opinion on abortion and when it can be done → "non-moral"
2) We who reject #abortion do not reject those who have had abortions.
Rather, we embrace them with mercy. I serve as Pastoral Director of the
world's largest ministry for healing after abortion. Help others find these
resources at our site, https://t.co/AuQitfVBU1 → "moral"
3) Life is a gift from God, not something to be ended at will. When we
destroy the unborn, we stray from the path He set for us. → "moral"
```

Figure S1: LLM prompt used to classify tweets as *moral* vs. *non-moral*.

Prompt for Moral Foundations (Multilabel)

```
You are an expert in moral psychology and discourse analysis. Your task is
to analyze a tweet previously classified as "moral" and assign one or more
Moral Foundations based on MFT. A tweet can reflect support for or violation
of a foundation|both count.
Output format: a JSON object with key "foundations" whose value is an array
listing every foundation present.
Foundations
- care: Harm, suffering, compassion, empathy, cruelty; protecting the
vulnerable, preventing suffering, calling something heartless.
- fairness: Justice, rights, equality, cheating; freedom, discrimination,
fairness, hypocrisy, bias.
- authority: Tradition, leadership, law, social order|or defiance of them;
obedience, discipline, or rebellion against unjust authority.
- loyalty: Group belonging, patriotism, betrayal, solidarity; in-group vs.
out-group, \us vs. them", loyalty to a cause or movement.
- purity: Moral disgust, sanctity, sin, degradation, contamination; bodily
purity, spiritual corruption, abortion as sacred or sinful.
If the tweet is not moral, output an empty JSON object.
Examples
1) We who reject #abortion do not reject those who have had abortions... →
["care"]
2) @elisa1121 Why does anyone think they have the right to tell a woman what
to do?? ... #prochoice → ["fairness"]
3) We should stand by the women in our communities not turn our backs on
them when they need us most. → ["loyalty"]
4) Life is a gift from God, not something to be ended at will... →
["purity"]
5) @kdmport @TexasTribune This [pointing up] is how we do it. What do we
want? Power ... #prochoice → ["authority"]
6) Please sign this petition to protect special babies! ... #prolife →
["care","purity","loyalty"]
7) Easy being #antichoice when you're not the one pregnant! #prochoice =
#prolife ... → ["fairness","care"]
```

Figure S2: LLM prompt used to assign Moral Foundations Theory (multilabel) categories.

Prompt for Abortion Stance (4-Class)

```
Your task is to classify tweets about abortion into one of four categories:
"life", "choice", "neutral", or "throw_out".
Output format:  a JSON object with exactly one key "label" whose value is
one of the four categories.
Guidelines
- life:  Supports anti-abortion views or policies; defends the rights of the
unborn; criticizes abortion; promotes \life" as the moral choice.
- choice:  Supports the legal right to abortion, bodily autonomy,
reproductive freedom; criticizes anti-abortion policies.
- neutral:  Discusses abortion-related topics (laws, elections, news,
statistics) without a clear stance; may include sarcasm, questions, or
hard-to-categorize observations.
- throw_out:  Not about abortion or too vague/unclear to label reliably.
Examples
1) The Barstool Bros' Split Over Abortion Could Determine the Future of the
GOP → "neutral"
2) @tdmalone1016 @LifeNewsHQ Urgent time for all the Churches...  should
repent now!  ...  Amen → "life"
3) Abortion is not the exclusive need of any political group...  There is a
legal Pt.  → "choice"
4) Is your complaint that 36 minutes means that they gave less than 11
seconds of air time...  → "throw_out"
```

Figure S3: LLM prompt used for four-way abortion stance classification.

Comparing VADER with GPT-3.5 Turbo

To assess the robustness of sentiment classification, we compared VADER sentiment labels with sentiment labels generated by `GPT-3.5-turbo` on a validation subset of 500 abortion-related tweets. Agreement between the two methods was high (87.2% agreement), and quadratic-weighted Cohen's κ indicated near-perfect agreement ($\kappa = .87$). Disagreements were primarily confined to adjacent categories (e.g., neutral vs. weakly valenced sentiment), with very few polarity reversals (positive vs. negative). Given this high level of convergence, we rely on VADER sentiment for all primary analyses due to its scalability and transparency for large-scale text data.

Table S9: Confusion matrix comparing VADER and GPT-3.5-turbo sentiment labels

VADER	**GPT-3.5-turbo**		
	Negative	Neutral	Positive
Negative	146	9	4
Neutral	17	116	14
Positive	6	14	174

Prompt for Sentiment Classification (3-Class)

```
Your task is to classify the sentiment (emotional tone) expressed in the
following tweet into one of three categories: "negative", "neutral", or
"positive".
Sentiment refers to emotional valence, not ideological position. A tweet
may express positive or negative sentiment toward either pro-life or
pro-choice views.
Output format: a JSON object with exactly one key "label" whose value is
one of the three categories.
Guidelines
- negative: The tweet expresses negative emotional tone, such as anger,
criticism, hostility, fear, disgust, or frustration, toward specific
stances, policies, individuals, or groups.
- neutral: The tweet discusses abortion-related topics (e.g., laws, court
decisions, elections, news, statistics) without clear emotional valence, or
expresses mixed or ambiguous emotion.
- positive: The tweet expresses positive emotional tone, such as approval,
pride, hope, gratitude, praise, or empowerment, toward a position, policy,
individual, or group, regardless of whether the stance is pro-life or
pro-choice.
Examples
1) The Supreme Court will hear arguments on the abortion ban next month. →
"neutral"
2) This ruling is a devastating attack on women's rights. → "negative"
3) Every child deserves protection, and I'm proud to stand for life. →
"positive"
4) Proud to see so many people standing up for reproductive freedom. →
"positive"
```

Figure S4: LLM prompt used for three-class sentiment classification of abortion-related tweets. Sentiment is defined independently of ideological stance.

Supplementary Tables for Five-Class Abortion Stance Analysis

The tables presented here offer detailed breakdowns of our dataset, including the distribution of tweets across the five stance categories and the prevalence of different moral foundations within each category. Additionally, we include the exact prompt used to instruct GPT-3.5 Turbo in classifying tweets according to their stance on abortion.

Table S10 presents the distribution of tweets across our five classification categories, showing that tweets where pro-choice advocates discuss pro-life views were most common in our dataset, followed by tweets exclusively expressing pro-life views. Table S11 provides a detailed breakdown of moral foundation usage across the different stance categories, demonstrating how care and fairness foundations dominate in pro-choice discourse, while care and purity foundations are more prevalent in pro-life discourse.

Table S10: Tweet Counts by Label Category

Label Category	**Number of Tweets**
Pro-choice talking about Pro-life	62,719
Only Pro-life	41,116
Only Pro-choice	33,468
Pro-life talking about Pro-choice	32,940
Neither	29,590

Table S11: Counts of Moral Foundations by Tweet Label Category

Label Category	**Care**	**Fairness**	**Authority**	**Loyalty**	**Purity**
Pro-choice talking about Pro-life	48,387	48,751	6,636	20,949	15,100
Only Pro-life	35,383	15,399	7,643	5,940	16,163
Only Pro-choice	26,877	26,266	4,461	6,048	3,236
Pro-life talking about Pro-choice	29,737	14,250	3,249	4,249	13,988
Neither	22,877	18,153	5,503	6,818	8,035

GPT-3.5 Turbo Classification Prompt for Abortion Stance Analysis

The following prompt was used to instruct GPT-3.5 Turbo in classifying tweets according to their stance on abortion and their reference patterns. This prompt was designed to categorize tweets into a five-category framework that captures not only whether a tweet expresses a pro-life or pro-choice stance, but also whether it references or critiques opposing viewpoints. The prompt includes clear category definitions and illustrative examples to guide the model in making accurate classifications. This approach allowed us to analyze both the primary stance of tweets and their engagement with opposing perspectives, providing a more nuanced understanding of abortion discourse on Twitter.

Prompt for Abortion Stance Classification (5-Class)

Your task is to classify tweets about abortion based on their *ideological framing*. Each tweet should be assigned to exactly one of the following five categories.
Output format: a JSON object with exactly one key "label" whose value is one of the five categories listed below.
Categories
- Pro-life talking about Pro-choice: The tweet expresses a pro-life position and references, critiques, or characterizes pro-choice views, arguments, or individuals.
- Pro-choice talking about Pro-life: The tweet expresses a pro-choice position and references, critiques, or characterizes pro-life views, arguments, or individuals.
- Only Pro-life: The tweet expresses support for pro-life views but does not reference or discuss pro-choice views or individuals.
- Only Pro-choice: The tweet expresses support for pro-choice views but does not reference or discuss pro-life views or individuals.
- Neither: The tweet does not clearly express or reference either pro-life or pro-choice ideological positions, or is too general, descriptive, or ambiguous to classify.
Examples
1) Pro-life is the only way to save innocent lives. How can pro-choice advocates support this murder? → "Pro-life talking about Pro-choice"
2) Women should have control over their own bodies. Pro-lifers need to stop meddling. → "Pro-choice talking about Pro-life"
3) Every life is sacred, and abortion is a crime. → "Only Pro-life"
4) Reproductive rights are human rights. Everyone should have access to abortion care. → "Only Pro-choice"
5) The debate on abortion continues to divide the country. → "Neither"

Figure S5: LLM prompt used for five-class abortion stance classification.

References

Chang, R.-C., Rao, A., Zhong, Q., Wojcieszak, M., & Lerman, K. (2023). #Roeoverturned: Twitter dataset on the abortion rights controversy. *Proceedings of the International AAAI Conference on Web and Social Media*, *17*(1), 997–1005. https://doi.org/10.1609/icwsm.v17i1.22207

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *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)* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., & et al. (2024). The llama 3 herd of models. https://doi.org/10.48550/arXiv.2407.21783

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. https://arxiv.org/abs/1907.11692

Nguyen, D. Q., Vu, T., & Nguyen, A. T. (2020). Bertweet: A pre-trained language model for english tweets. https://doi.org/10.48550/arXiv.2005.10200

OpenAI. (2025, April). Introducing openai o3 and o4-mini [Accessed: 2025-08-18].

Shen, S., Liu, J., Lin, L., Huang, Y., Zhang, L., Liu, C., & Wang, D. (2023). Sscibert: A pre-trained language model for social science texts. *Scientometrics*, *128*(2), 1241–1263. https://doi.org/10.1007/s11192-023-04693-4