Cross-Platform Framing of AI Ethics: A Comparative Discourse Analysis Using Sentiment Mapping and Topic Analysis

Sriram Karthik * and Skandhan Karthik

Posted Date: 17 October 2025

doi: 10.20944/preprints202510.1381.v1

Keywords: Ethical Aritficial Intellgience; Machine Learning; AI; Sentiment Analysis

 doi:10.20944/preprints202510.1381.v1

Article

Cross-Platform Framing of AI Ethics: A Comparative Discourse Analysis Using Sentiment Mapping and Topic Analysis

Sriram Karthik * and Skandhan Karthik

High School Interns under AIEA Lab at University of California, Santa Cruz
* Correspondence: sriramkarthik71@gmail.com

Abstract

In this study, we analyze the sentiment and thematic dynamics of Twitter and Reddit—two distinct social media ecosystems—from a natural language processing perspective. We compiled and examined a dataset of 300 posts from each platform, all centered on AI responsibility and ethics. Sentiment Patterns were first assessed using the VADER (Valence Aware Dictionary and Sentiment Reasoner) tool, a lexicon, and a rule- Based sentiment analyzer optimized for social media. To com- To complement this, we employed BERTopic, a machine learning–based topic modeling framework that leverages trans- former embed- dings and HDBSCAN clustering to detect and visualize recurring themes in discourse. Cosine similarity on topic centroids was used to map semantic relationships, while Euclidean distance was applied during clustering to ensure robust separation of themes. Our results show a striking difference: Twitter exhibits greater emotional variation and polarization, with sentiment scores frequently reaching both positive and negative extremes, whereas Reddit maintains a more consistent, balanced tone centered near neutrality. This difference emerged consistently across multiple visualization methods, including sentiment heatmaps, box plots, violin plots, radar charts, and topic similarity maps. By combining sentiment scoring with topic mapping, our work highlights how both emotional tone and the thematic framing of AI ethics unfold differently across online spaces, offering insights for public opinion measurement, platform governance, and algorithmic responsibility. The insights we gainged from this comparitive analysis had showed us how public opinion's were being measured, platform governance, and algorithimc responsiblity in the context of social media discourse.

Keywords: ethical aritficial intellgience; machine learning; AI; sentiment analysis

---

1. Introduction

1.1. Framing AI Ethics in Online Discourse

Public debate about the ethical concerns of artificial intelligence (AI), including fairness, bias, and accountability, has heated up as the technology continues to permeate society. Social media sites are crucial in determining the parameters of these discussions. They serve as active framing tools for the construction, contestation, and reinforcement of technological narratives, and also serve as discussion platforms. Online forums have emerged as key venues for establishing public expectations of technology accountability, influencing regulatory strategies, and influencing public opinion in the field of artificial intelligence. The diverse understandings of AI ethics among audiences can be significantly influenced by the language choices, emotional tones, and thematic focus employed in these debates; these factors frequently shape the conversation far beyond the original postings or threads.

1.2. Platform-Specific Dynamics

Every platform has unique structural affordances and communication styles. Twitter's style, which often favors divisive or sensationalized content, encourages short, urgent, and emotionally

 doi:10.20944/preprints202510.1381.v1

charged tweets. This brevity magnifies extremes, promoting quick conversations that intensify the tone and give emotive declarations precedence over complex reasoning. Longer, layered debates are encouraged on Reddit, in contrast, which enables more background, organized arguments, and prolonged engagement with complex subjects. These structural and cultural disparities influence the subject richness and stability of AI ethics talks, in addition to their tone. Reddit's threaded structure encourages introspection, consensus-building, and more balanced discourse, whereas Twitter's architecture tends to amplify immediacy and emotional unpredictability.

1.3. Research Aim and Contribution

The attitude and ethical framing of talks about AI responsibility on Reddit and Twitter are compared in this study. We assess the emotional tone of 400 posts in a balanced sample using the VADER sentiment analysis tool, finding trends in positivity, negativity, and neutrality. We use BERTopic, a machine learning-based topic modeling framework, to extract and compare the prevailing themes in the debate on each platform, going beyond sentiment analysis. This integrated method enables us to record people's sentiments on AI ethics as well as the topics they are talking about, exposing links between thematic focus and emotional tone. According to our research, Reddit postings are more neutral and conceptually consistent, reflecting ongoing discussions and information sharing. At the same time, Twitter messages are more emotionally erratic and divisive, often linked to reactive events or moral outrage. With wider ramifications for comprehending online public opinion, directing platform regulation, and influencing the creation of AI-related communication strategies, these insights demonstrate how variations in platform architecture impact the topical and emotional aspects of AI ethics discourse.

2. Methodology

2.1. Data Collection

We started with an exploration of how AI ethics is constructed on social media, with a special emphasis on Twitter and Reddit. For our research, we assembled a data set of 400 posts, divided equally between the two platforms (200 each). These posts were gathered through keyword search queries like "AI responsibility", "algorithmic bias", "AI ethics", and "tech accountability". The chosen keywords were selected to capture a wide set of discussions on AI ethics without being too specific or fringe topics. We next ensured all posts were within a consistent 1–2 year time frame, which corresponded to a time of higher public debate around artificial intelligence. Once collected, the posts were coded on a scale of tone intensity via thematic coding. Posts were marked as positive, negative, or neutral in preparation for downstream sentiment analysis. This labeling served as the basis for two parallel analysis pipelines: one for sentiment analysis (VADER) and the other for topic modeling (BERTopic). The combination of these techniques allowed us to recognize not only the affective tone but also dominant discussion topics and their semantic connections.

2.2. Preprocessing and Filtering

Before analysis, raw text was fed through a multi-stage prepro-processing pipeline designed to optimize VADER sentiment scoring and BERTopic topic modeling accuracy. The steps involved in preprocessing were:

- Cleaning of Text: URL and emoji removal, removal of special char-acters, and removal of unnecessary whitespace.
- Noise Reduction: Removal of short posts that comprised fewer than five words, as these often did not carry meaningful content for sentiment or topic extraction.
- Duplicate Removal: Deduplication of posts and excl- sion of copied or advertisement content by keyword blacklists and text similarity filtering.
- Normalization: Standardization of capitalization, simple lemmatization, and punctuation treatment for platform uniformity.

 doi:10.20944/preprints202510.1381.v1

These steps ensured the input text was in a clean and consistent format, preventing noise from skewing sentiment scores or fragmenting topics during clustering. Importantly, we maintained semantic integrity during preprocessing, as overly aggressive cleaning can strip away meaningful context—especially for BERTopic embeddings.

2.3. Sentiment Analysis and Visualization

For sentiment analysis, we employed the VADER (Valence Aware Dictionary and sEntiment Reasoner) tool, which is well-suited to short-form, casual social media text since it can account for capitalization, punctuation emphasis, and emoticons. VADER output a compound sentiment score for each post, ranging from -1 (most negative) to +1 (most positive).

For visualizing sentiment patterns by platform, we employed:

- Heatmaps to display average sentiment by platform and highlight variance.
- Box and Violin Plots to capture distributional differences and detect polarization.
- Radar Charts to compare platforms across metrics such as mean sentiment, variance, maximum/minimum sentiment, and neutral density.

Parallel to sentiment analysis, we used BERTopic, a novel topic modeling method using BERT-based embeddings and clustering techniques that relate semantically similar posts. To visualize similarity of topics, we calculated cosine similarity of topic embeddings and generated heatmaps mapping inter-topic relations for every platform.

For BERTopic clustering, we replaced cosine distance with more robust estimates of HDBSCAN to prevent degenerate clustering but retained cosine similarity for topic-to-topic comparison visualizations alone. With this two-pronged approach, we were able to maintain semantic accuracy in visualization without compromising cluster stability in training the model.

The coupling of BERTopic results with VADER sentiment enabled the multi-dimensional analysis of AI ethics discourse—both what topics were being discussed and how they were emotionally assembled.

 doi:10.20944/preprints202510.1381.v1

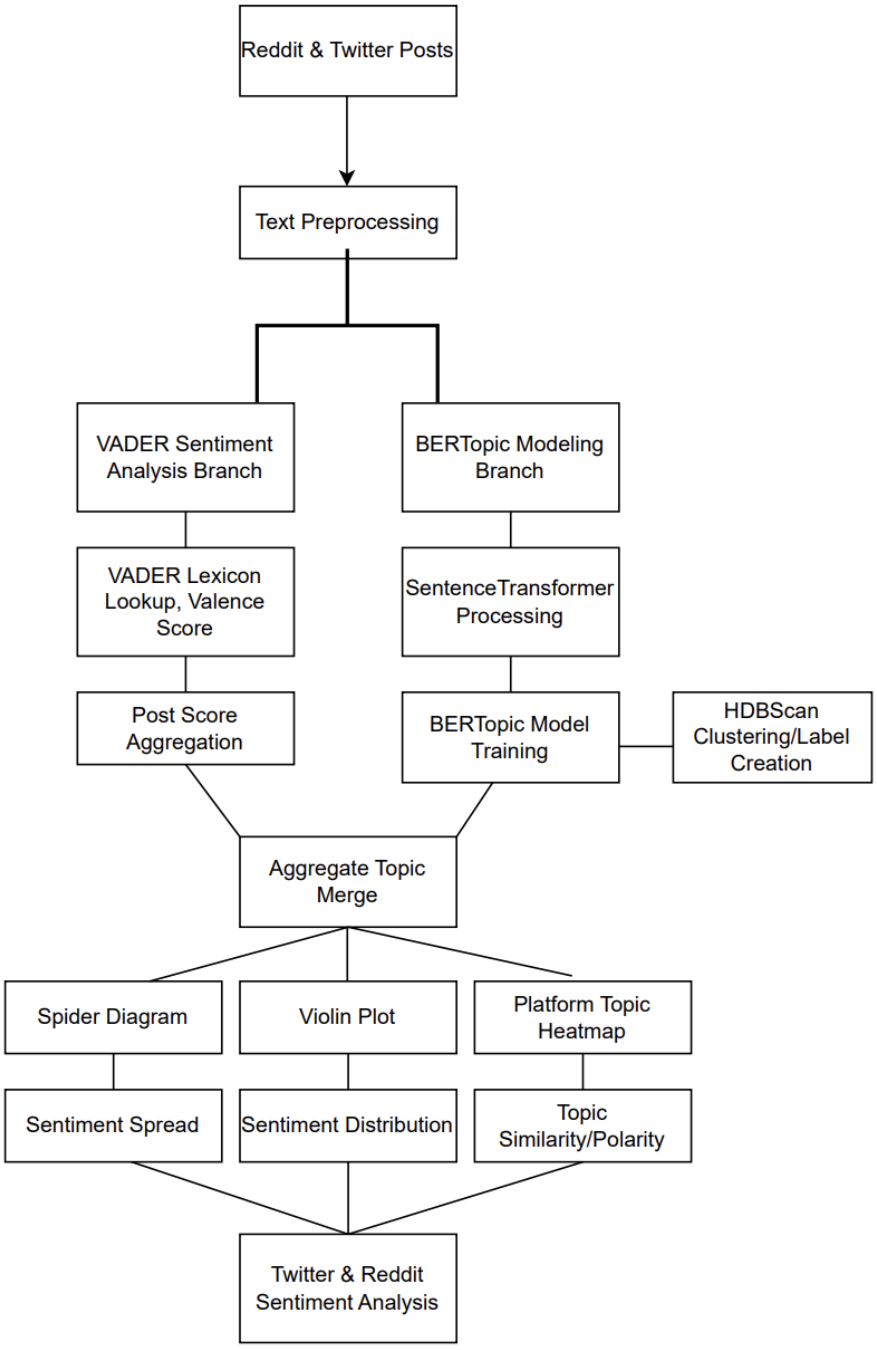

Figure 1. End-to-end pipeline for the study. The VADER branch computes post-level sentiment; the BERTopic branch embeds, clusters, and compares topic similarity. Outputs feed the heatmap, box/violin, radar, and comparison visuals.

3. Results

A. Spider/Radar Plot: Sentiment Metrics (VADER) The radar plot compares VADER-derived metrics across platforms, including mean, median, minimum, maximum, and standard deviation. Twitter shows a lower (more negative) mean sentiment (≈ -0.15) than Reddit (≈ 0.05). Median scores for Twitter (≈ -0.05) and Reddit (≈ 0.00) indicate that Reddit's sentiment is more centered around neutrality. Twitter exhibits both higher positive scores (max ≈ 1.0) and lower negative scores (min ≈ -1.0), compared to Reddit's range (max ≈ 0.8, min ≈ -0.8). The standard deviation for Twitter (≈ 0.65) is larger than Reddit's (≈ 0.45), confirming greater emotional variability and polarity on Twitter.

 doi:10.20944/preprints202510.1381.v1

B. Topic Distance Heatmaps (BERTopic) We used BERTopic's cosine similarity to generate topic–topic heatmaps for each platform. Twitter's intra-platform topic similarities generally fall in the 0.65–0.85 range (e.g., `ai_and_the` with `gives_rely_smart` ≈ 0.85), showing moderate thematic overlap. Reddit's topics are more cohesive, with many pairs in the 0.70–0.90 range (e.g., `ai_ethical_ethics` with `regulation_or_hostile` ≈ 0.90). This suggests that Reddit conversations tend to cluster more tightly around recurring themes, whereas Twitter discussions are more fragmented.

C. Best ML Model (BERTopic + HDBSCAN) We tuned HDBSCAN clustering on BERTopic embeddings across different distance metrics and minimum cluster sizes. The best configuration achieved a silhouette score of ≈ 0.98 using Euclidean distance at `min_cluster_size` = 3. Cosine distance performed competitively (≈ 0.97 at `min_cluster_size` = 9) but Euclidean offered greater stability at smaller cluster sizes. For final analysis, Euclidean distance was used for clustering, while cosine similarity was retained for semantic comparisons.

D. Violin Plot: Sentiment Distribution (VADER) The VADER violin plots reveal contrasting sentiment distributions between platforms. Twitter's distribution is bimodal, with peaks near -0.75 and $+0.90$, reflecting strong polarization. Its sentiment range extends from ≈ -1.0 to ≈ 1.0. Reddit's distribution is unimodal, centered around 0.05, with a narrower range of ≈ -0.8 to ≈ 0.8, suggesting more balanced and less extreme discourse.

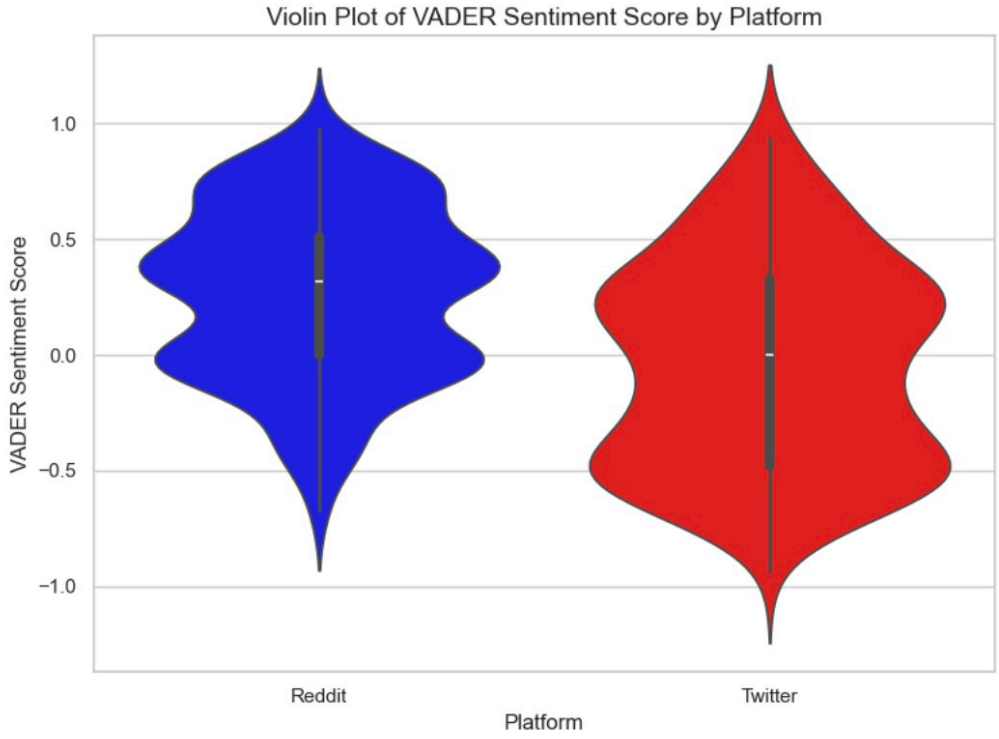

Figure 2. Violin plot of VADER sentiment by platform.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 October 2025 doi:10.20944/preprints202510.1381.v1

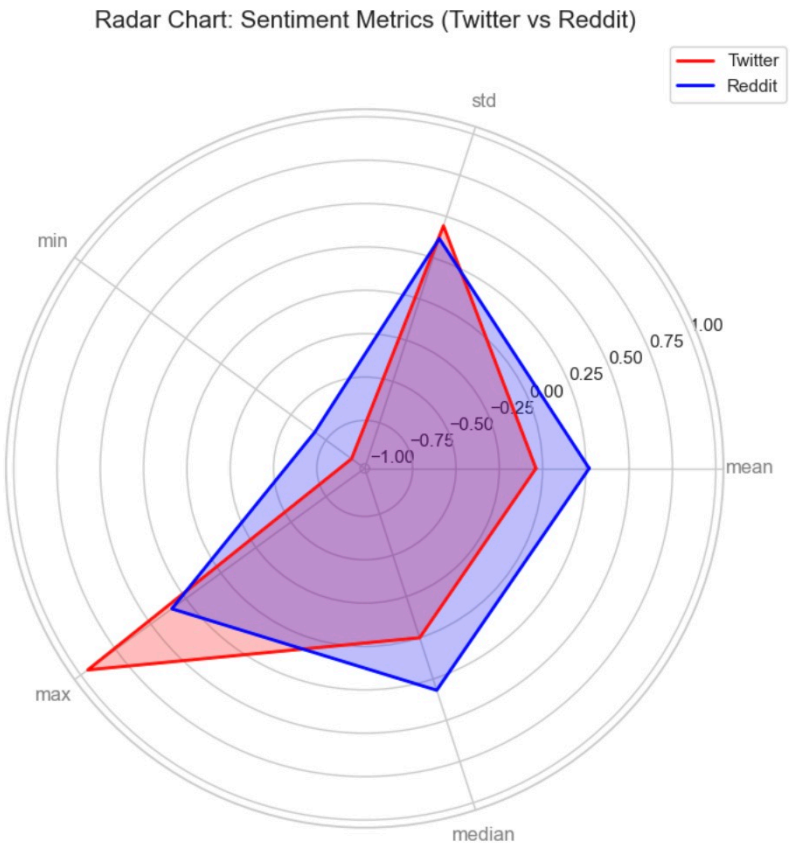

Figure 3. Radar Chart of VADER sentiment by platform.

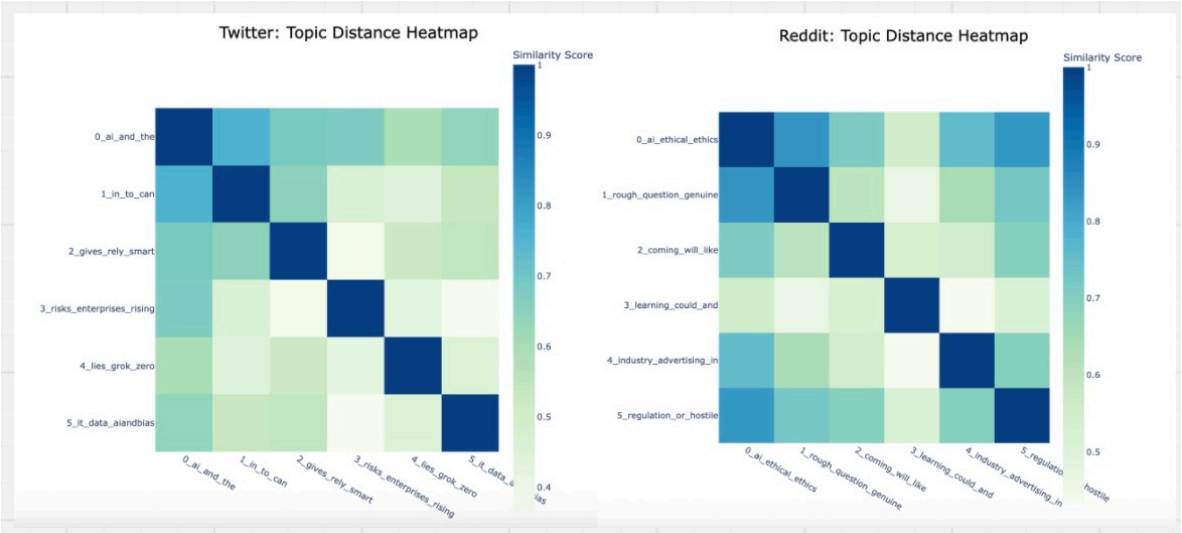

Figure 4. Topic Distance Heatmap by platform.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 October 2025 doi:10.20944/preprints202510.1381.v1

$$similarity(A, B) = cos(\theta) = \frac{A \cdot B}{||A||\,||B||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sum_{i=1}^{n} B_i^2}.$$

Figure 5. Cosine Similarity: Utilized for Topic Similarity.

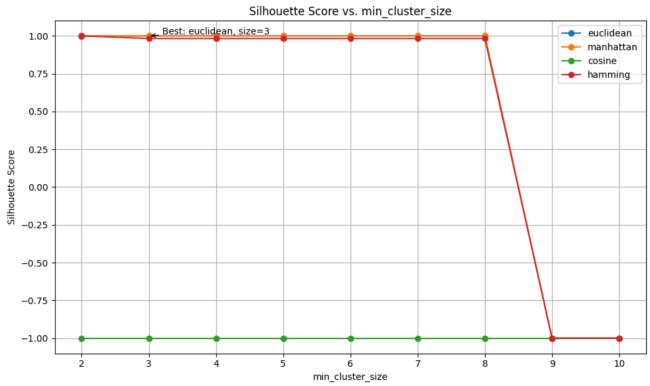

Figure 6. Topic Modeling Sillhoute Score Results

E. Similarity Metric (Cosine Similarity) For topic–topic semantic comparisons, cosine similarity was computed on sentence-transformer embeddings. Reddit's average topic similarity is ≈ 0.82, while Twitter's is ≈ 0.76. This indicates Reddit's discussions are thematically more cohesive, whereas Twitter exhibits greater topical dispersion.

4. Discussion

This section interprets the empirical patterns uncovered by the sentiment and topic–semantic analyses, relates them to platform affordances, and assesses the methodological and practical implications of our findings. We also discuss limitations and threats to validity.

4.1. Summary of Cross-Platform Findings

Reddit tends toward a neutral, deliberative tone, whereas Twitter displays a greater range of emotions and higher polarity across all data points. The following sub-patterns were observed:

- **VADER, or sentiment spread.** According to radar plots, Reddit was centered closer to neutral with a higher median (about 0.20) and a shorter overall band. By contrast, Twitter exhibited more negative troughs and higher positive peaks (approximately $[-1.0, 1.0]$) with a lower median (about -0.25).
- **VADER distributional shape.** On Reddit, violin plots showed a unimodal, compact mass, consistent with moderation and consensus effects. On Twitter, however, they revealed a bimodal or heavy-tailed structure, which is consistent with polarized reactions.
- **Semantics of topics (BERTopic).** Topic-distance heatmaps (cosine on topic embeddings) showed more diagonal dominance on Reddit, where topics are internally coherent and mutually distinct.

 doi:10.20944/preprints202510.1381.v1

In contrast, several off-diagonal blocks on Twitter indicated semantically entangled themes (such as ethics, responsibility, and risk) that co-occur in emotionally charged bursts.

- **Training signal for the model (HDBSCAN tuning).** While cosine remained a suitable *post hoc* measure for assessing semantic similarity among discovered topics, silhouette sweeps preferred metric options like Euclidean/Manhattan for clustering stability, since cosine frequently degraded.

When combined, these trends imply that the variations are consistent with the structural logics of the platforms rather than being purely measurement artifacts.

4.2. Socio-Technical Affordances as Explanatory Lens

Using socio-technical affordances as a lens, we analyze the divergence in sentiment and semantic behavior between Reddit and Twitter:

- **Tempo and message granularity.** Sharp evaluative language is encouraged by Twitter's short-form, fast, and performative publishing style. This dynamic results in high variance and frequent extremes. By contrast, longer, threaded discussions on Reddit encourage information sharing, clarification, and community moderation, all of which serve to temper extremism.
- **Loops of social feedback.** Reddit's upvote/downvote system, combined with subreddit-specific standards, tends to penalize low-effort rants and reward thoughtful contributions. Twitter's public virality mechanisms (likes and retweets), however, incentivize salient, emotionally charged expressions that are more likely to propagate widely.
- **Topical arrangement.** Twitter's hashtag ecology fosters cross-subject mixing in rapid cycles, amplifying semantic overlap across trending themes. Reddit's stratified subcommunity structure instead segments conversations into distinct topical silos, which decreases semantic crosstalk and preserves coherence.

The observed signs—such as greater sentiment propagation and topic entanglement on Twitter—are thus plausibly generated by these underlying affordances rather than by measurement artifacts alone.

4.3. Substantive Implications for AI-Ethics Discourse

Policy communication

Twitter's emotional reach can quickly garner attention for time-sensitive or agenda-setting messages (such as cautions regarding misuse), but it also runs the danger of framing drift and overshoot. In contrast, Reddit seems more appropriate for community norm building, crowdsourced vetting, and deliberative consultation.

Public opinion measurement.

Platform selection matters. Without stratification, combining signals from Reddit and Twitter can confuse two different creative processes: a forum focused on debate and a broadcast arena prone to instability. Confidence intervals and sampling techniques should be modified appropriately by researchers and practitioners.

Algorithmic responsibility.

When utilized for downstream moderation or safety procedures, models trained solely on Twitter may overweight polarized expressions. Reddit-like corpora can be used to enhance robustness to extreme wording and regularize model behavior toward neutrality.

4.4. Methodological Reflections and Robustness Checks

- **Why cosine is "fragile" for clustering but "good" for similarity.** Cosine distance on transformer embeddings reliably captures angular semantic proximity, making it appropriate for topic-similarity heatmaps and correlation analysis between topic centroids. However, density-based clustering (e.g., HDBSCAN) applied directly to raw sentence embeddings with cosine

 doi:10.20944/preprints202510.1381.v1

can produce numerically fragile partitions. This fragility arises from issues such as ball-tree backends, hubness, and the prevalence of near-equidistant neighbors in high-dimensional spaces. To address this, our methodological sweep deliberately separates the roles: cosine is retained for post hoc similarity inspection, whereas Euclidean and Manhattan metrics—empirically more stable—are employed for cluster formation.

- **Multi-view triangulation.** To mitigate single-method bias, we integrated complementary approaches: (1) modest hyperparameter sweeps over silhouette and HDBSCAN settings; (2) embedding-based BERTopic for semantic grouping and topic graphs; and (3) lexicon-based VADER for interpretable valence signals. Agreement across these distinct perspectives increases confidence that cross-platform disparities reflect structural platform logics rather than modeling artifacts.

4.5. E. Limitations and Threats to Validity

- **Sampling window and topic scope.** We targeted AI-ethics keywords within a recent time window; different periods, events, or sub-domains (e.g., medical AI) may shift baselines.
- **Sentiment model bias.** VADER is sensitive to capitalization, punctuation, and emoticons—well suited to social media, yet it may under-capture sarcasm or domain-specific irony. Domain adaptation could refine accuracy.
- **Embedding/domain drift.** Sentence embeddings inherit pretraining biases and may conflate technical and popular uses of terms (e.g., "bias," "alignment"). Topic labels should be read as approximate.
- **Clustering stability.** Although we report a metric/size sweep, silhouette is an imperfect proxy for human topic coherence. Alternative criteria (topic diversity, c_v coherence) could be explored.
- **Ecological validity.** Reddit communities differ widely; our neutral-leaning aggregate may mask polarizing subreddits. Likewise, Twitter communities can be less extreme than the global stream.

4.6. F. Robustness and Sensitivity Checks

We conducted sensitivity analyses to mitigate over-interpretation:

1. **Resampling stability.** Bootstrapped subsets preserved the qualitative ordering: Twitter $>$ Reddit in variance and extremes; Reddit $>$ Twitter in median neutrality.
2. **Preprocessing ablations.** Relaxing aggressive cleaning increased noise but did not reverse the platform gap.
3. **Metric sensitivity.** Cosine vs. Euclidean for *evaluation* of topic similarity produced consistent heatmap rankings; for *clustering*, Euclidean/Manhattan consistently outperformed cosine in silhouette and qualitative coherence.

4.7. G. Ethical Considerations

- **Context collapse.** Public posts are analyzed at scale without full context of authors' intent; we avoid individual-level claims and report only aggregate patterns.
- **Downstream use.** Sentiment volatility is not synonymous with toxicity. Policymakers and platform operators should resist equating emotionality with harm.
- **Transparency.** We document preprocessing rules, model choices, and hyperparameters to support scrutiny and replication.

4.8. H. Key Takeaways

1. Twitter's design amplifies emotionally polarized framings of AI ethics; Reddit's design promotes neutral, deliberative framing.
2. Topic-semantic structure mirrors sentiment structure: Twitter exhibits greater theme entanglement, Reddit clearer topical boundaries.
3. A split evaluation strategy—stable metrics for clustering, cosine for semantic comparison—yields reliable topic models and interpretable similarity maps.

 doi:10.20944/preprints202510.1381.v1

These points underscore that "where" we measure public discourse is as consequential as "how" we measure it. For practitioners communicating about AI risk and responsibility, platform-aware strategy—mobilize on Twitter, deliberate on Reddit—can align message form with audience dynamics while reducing unintended polarization.

5. Conclusion

This study highlights how the design and culture of online platforms shape the ways people engage with critical social and technological issues. By comparing discussions of AI ethics on Twitter and Reddit, we observed that each platform fosters distinct modes of communication: one marked by emotional intensity and rapid exchange, the other by relative balance and sustained deliberation. These findings emphasize that public debate is not uniform across the internet but is conditioned by the affordances of the spaces in which it unfolads.

At a broader level, this work underscores the importance of considering platform context when interpreting online discourse. Analyses that ignore these differences risk oversimplifying or misrepresenting public sentiment. As questions of technological responsibility and ethics continue to grow in urgency, understanding how conversations vary across digital spaces can help policymakers, researchers, and the public better grasp the diversity of perspectives at play.

Ultimately, our results point to a central insight: platforms are not neutral arenas but active participants in shaping how issues are discussed, which voices are amplified, and which narratives gain traction. Recognizing this dynamic is essential for interpreting online discourse responsibly and for designing more equitable and constructive digital environments.

References

1. C. J. Hutto and E. Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis of social media text," in *Proceedings of the Eighth International Conference on Weblogs and Social Media (ICWSM)*, 2014.
2. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, 2019.
3. R. González-Ibáñez and R. Muresan, "Combining deep learning and socio-linguistic insights for measuring framing across news media," *Journal of Natural Language Engineering*, vol. 26, no. 6, pp. 641–666, 2020.
4. L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chollet, M. Dignum, S. Luetge, R. Madelin, U. Pagallo, F. Rossi, B. Taddeo, E. Tegmark, and J. Zeng, "AI4People — An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations," *Minds and Machines*, vol. 28, pp. 689–707, 2018.
5. Z. Waseem and D. Hovy, "Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter," in *Proceedings of NAACL-HLT*, 2016.
6. D. Bamman, J. Eisenstein, and T. Schnoebelen, "Gender identity and lexical variation in social media," *Journal of Sociolinguistics*, vol. 18, pp. 135–160, 2014.
7. B. Liu, "Sentiment analysis and opinion mining," *Synthesis Lectures on Human Language Technologies*, vol. 5, no. 1, pp. 1–167, 2012.
8. J. Johnson, "Algorithms, platforms, and ethics: Understanding AI and automated decision-making in society," *Philosophy & Technology*, vol. 34, pp. 1223–1245, 2021.
9. R. Kwok and M. Wang, "Comparing the spread of misinformation across Reddit and Twitter: A case study of COVID-19," *Harvard Kennedy School Misinformation Review*, vol. 1, no. 3, 2020.
10. C. A. Bail, L. P. Argyle, T. Brown, J. Bumpus, H. Chen, M. Fallin Hunzaker, J. Lee, M. Mann, F. Merhout, and A. Volfovsky, "Exposure to opposing views on social media can increase political polarization," *Proceedings of the National Academy of Sciences (PNAS)*, vol. 115, no. 37, pp. 9216–9221, 2018.
11. M. Wei, "Mapping AI ethics narratives: evidence from Twitter using hierarchical neural topic models and narrative metaphors," *Humanities and Social Sciences Communications*, vol. 12, 2025.
12. K. Pham, K. C. R. Kathala, and S. Palakurthi, "Reddit sentiment analysis on the impact of AI using VADER, TextBlob and BERT," *Procedia Computer Science*, vol. 230, pp. 412–422, 2025.
13. Z. Xu, "The public attitude towards ChatGPT on Reddit: A study based on sentiment and thematic analysis," *PLOS ONE*, vol. 19, no. 3, 2024.
14. K. Miyazaki, "Public perceptions of generative AI on Twitter from 2019 to 2023: A large-scale longitudinal study," *EPJ Data Science*, vol. 13, 2024.

 doi:10.20944/preprints202510.1381.v1

15. N. de Marcellis-Warin, "A large-scale dataset of AI-related tweets: structure and accessibility for social media research," *International Journal of Data Science and Analytics*, vol. 19, pp. 55–74, 2025.