Measuring Social Media Polarization Using Large Language Models and Heuristic Rules

Jawad Chowdhury, Rezaur Rashid, and Gabriel Terejanu

Department of Computer Science, University of North Carolina at Charlotte, Charlotte, NC 28223, USA
{mchowdh5, mrashid1, gabriel.terejanu}@charlotte.edu

Abstract. Understanding affective polarization in online discourse is crucial for evaluating the societal impact of social media interactions. This study presents a novel framework that leverages large language models (LLMs) and domain-informed heuristics to systematically analyze and quantify affective polarization in discussions on divisive topics such as climate change and gun control. Unlike most prior approaches that relied on sentiment analysis or predefined classifiers, our method integrates LLMs to extract stance, affective tone, and agreement patterns from large-scale social media discussions. We then apply a rule-based scoring system capable of quantifying affective polarization even in small conversations consisting of single interactions, based on stance alignment, emotional content, and interaction dynamics. Our analysis reveals distinct polarization patterns that are event dependent: (i) anticipation-driven polarization, where extreme polarization escalates before well-publicized events, and (ii) reactive polarization, where intense affective polarization spikes immediately after sudden, high-impact events. By combining AI-driven content annotation with domain-informed scoring, our framework offers a scalable and interpretable approach to measuring affective polarization.

Keywords: Affective Polarization, Social Media Discourse, Large Language Models, Stance Detection, AI for Social Impact

1 Introduction

The rise of social media platforms in recent years has transformed political discourse by enabling real-time information exchange and broader audience engagement [9, 25]. This transformation, driven by the evolving media landscape, continues to shape how information is produced, distributed, and consumed while simultaneously redefining how individuals interact and maintain connections in digital spaces [4–6, 16]. While these platforms facilitate engagement, they have also intensified ideological divisions, as algorithmic content curation reinforces preexisting beliefs by prioritizing content aligned with users' prior views, limiting exposure to diverse perspectives. This selective exposure contributes to *affective polarization*, where individuals develop strong positive emotions toward their

in-group members while exhibiting hostility toward those from opposing groups or with opposing views [8, 19, 26]. Studies suggest that such polarization is not only shaped by political ideology but also by the emotional tone, discourse structure, and interaction patterns within online discussions. Affective polarization has been linked to increased political radicalization, reduced bipartisan cooperation, and the spread of misinformation [31, 33]. Understanding its dynamics is crucial for evaluating the broader societal implications of online discourse.

Social media platforms, such as Twitter (now X), can further amplify these divisions through algorithmic content curation, which prioritizes engagement-driven interactions and often promotes sensationalized, polarizing content [5, 21, 24]. Research suggests that online echo chambers reinforce polarization by predominantly exposing users to like-minded perspectives while restricting interaction with opposing viewpoints [10,12]. However, while exposure to counter-ideological content has the potential to correct misperceptions and reduce polarization in some cases [7], it can also provoke defensive responses, particularly in contentious social movements, where ideological conflict is often accompanied by toxic interactions and digital aggression [27]. These dynamics highlight the complex role of social media in shaping ideological divides and underscore the need for robust methodologies to quantify and analyze affective polarization in online discourse.

Existing approaches to measuring affective polarization in social media largely rely on sentiment analysis, stance detection, and/or network-based polarization indices [11, 17, 23, 29]. Sentiment analysis techniques classify text as positive, negative, or neutral, providing a general sense of emotional tone but often failing to capture the complexity of political discourse, such as sarcasm or implicit bias [22]. Stance detection methods aim to determine whether a user supports, opposes, or remains neutral on an issue, yet they frequently struggle with linguistic nuances, especially in highly polarized debates where positions are subtly framed [1, 18]. Recent studies have combined multimodal signals [28] or social network structures [14, 32] to improve measurement, but these still face limitations in capturing the nuanced emotional dimensions driving polarization.

Recently, large language models (LLMs) have emerged as powerful tools for text classification and content analysis, offering deeper contextual understanding than traditional sentiment classifiers [2, 3, 34]. LLMs leverage vast amounts of data to better capture subtle variations in ideological framing, rhetorical strategies, and the emotional undercurrents of online discourse [13, 30]. Their ability to process language in a more context-aware manner makes them particularly well-suited for analyzing sentiment shifts, emotional intensity, and therefore affective polarization. Studies have successfully applied LLMs for detecting political bias, misinformation, and ideological framing, demonstrating their potential in large-scale social media analysis [15, 20, 35]. However, despite these advancements, the application of LLMs in systematically quantifying affective polarization, which involves measuring both stance alignment and emotional intensity, remains underexplored. Addressing this gap requires a synergetic approach that can combine LLM-driven annotation with domain expertise to ensure a more

interpretable and scalable approach to measuring affective polarization across social media platforms.

Building on this line of work, our study introduces a significant departure from existing methodologies. Rashid et al. [23] developed a counterfactual framework that leverages the network-based polarization index [29] to examine the role of influential users in shaping polarization on Twitter (now X). Their study assessed how removing influential conversations altered polarization scores but relied on pre-defined sentiment-based classifiers. In contrast, our work employs LLM-based annotation to extract stance, affect, and agreement patterns from large-scale discussions, enabling a more nuanced understanding of affective polarization.

Furthermore, we integrate heuristic rules to score polarization, addressing limitations in existing sentiment-based approaches and ensuring interpretability in measuring discourse intensity. Specifically, our LLM-driven framework classifies tweets based on their stance (support, opposition, or neutrality on an issue), affective content (presence of emotionally charged language indicative of polarization), and agreement patterns (the extent to which replies align or conflict with the stance of the original post). These extracted attributes are then used within a structured scoring system that quantifies polarization by evaluating stance alignment, emotional intensity, and disagreement dynamics.

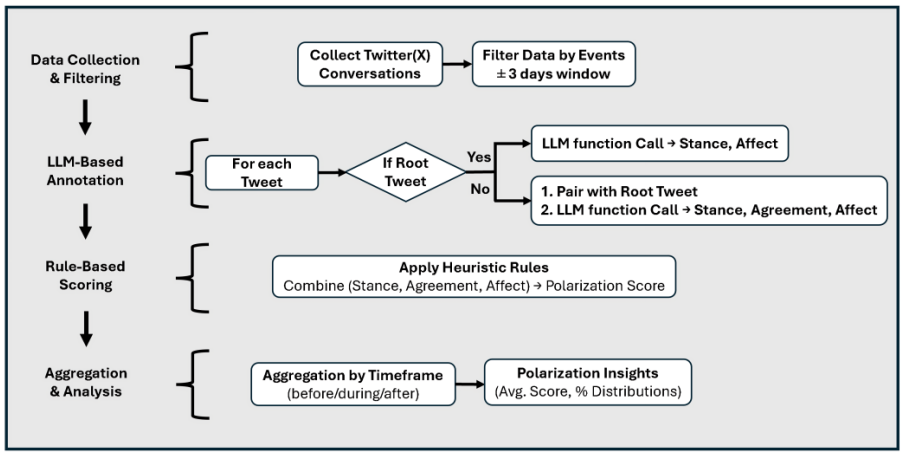

Fig. 1. Detailed workflow pipeline of our methodology, illustrating data collection and filtering, LLM-based annotation, application of heuristic rules, and aggregation for polarization insight generation.

By leveraging LLMs and domain-informed heuristics, our method provides a scalable, interpretable, and more context-aware approach to polarization measurement than traditional sentiment-based techniques. The major contributions of our study are stated as follows:

- We propose a novel LLM-based framework for measuring affective polarization, enhancing traditional sentiment analysis-based methods by incorporating large-scale language understanding.
- We introduce a scoring system that systematically quantifies affective polarization, capturing nuanced discourse dynamics such as stance shifts, emotional intensity, and disagreement patterns. Unlike network-based polarization measures, our scoring system is explainable and can effectively quantify polarization even in very small conversations involving only a single interaction.
- We conduct a large-scale empirical analysis of affective polarization in highly contentious discussions on climate change and gun control, uncovering key event-driven polarization trends and distinguishing anticipatory vs. reactive polarization.

The remainder of this paper is structured as follows: Section 2 details our proposed framework, covering data collection, LLM-based annotation, and affective polarization scoring using heuristic rules. Section 3 presents our findings, analyzing how affective polarization evolves over time in response to major events. Finally, Section 4 offers concluding remarks and outlines directions for future research.

2 Proposed Approach & Implementation

This study employs large language models (LLMs) to analyze affective polarization in online discussions surrounding contentious issues. Unlike previous studies that relied on sentiment classifiers or rule-based stance detection, we introduce a hybrid approach that integrates opensourced pretrained large language model LLaMA 3.1 70B for automatic text analysis with predefined scoring system to systematically quantify affective polarization. Our approach leverages LLMs to extract critical attributes such as stance, affective content, and agreement levels between users and their posts, while predefined rules assign affective polarization scores based on interaction patterns.

The overall workflow pipeline followed in our proposed work is illustrated in Figure 1, and consists of four main stages. **(1) Data collection:** retrieving Twitter (X) conversation threads related to highly debated sociopolitical topics e.g. climate change and gun control. **(2) LLM-based annotation:** extracting stance, affect, and agreement between tweets using LLaMA 3.1 70B. **(2) Predefined rule application:** assigning affective polarization scores using domain heuristics based on stance alignment, affect, and agreement information extracted by LLM in the previous step. **(4) Aggregation & analysis:** computing polarization score at the conversation level and evaluating polarization trends over time.

2.1 Data Collection and Structure

Our study employs a comprehensive dataset from Twitter (now X), focusing on two contentious political issues - climate change and gun control using a keyword-

based approach. An initial set of tweets were retrieved via the Twitter API prior to its 2023 restrictions; based on curated keywords and hashtags relevant to each topic, including terms like "#ClimateCrisis," "#GunReformNow," and event-specific phrases tied to major incidents (e.g., IPCC reports, mass shootings). To capture full user interactions, conversation cascades were expanded recursively to include all referenced tweets (replies, quotes, parents), ensuring inclusion of relevant discussions even without explicit keyword matches.

The climate change dataset spans June 1, 2021, to May 31, 2022, comprising 46M tweets from 4.8M unique users across 726,378 conversation threads of at least three tweets. The gun control dataset covers January 1, 2022, to December 31, 2022, with 14.4M tweets from 2.66M unique users across 335,000 conversation threads. To focus on threads with substantial engagement and to mitigate the influence of trivial or low-activity conversations, threads were restricted to those with ≥ 20 tweets and ≥ 10 unique users. Finally, each dataset consists of structured conversation threads, which we define as:

- **Parent Tweet:** The original post initiating the discussion.
- **Child Tweet:** Replies engaging with the parent tweet.

Each tweet-reply pair was analyzed independently to extract stance, affect, and agreement, forming the basis for our affective polarization scoring. While the dataset contains a broad collection of conversations, our analysis specifically focuses on eight key events: four related to climate change and four to gun control. For each event, conversations were segmented into three distinct timeframes:

- **Before:** Conversations occurring from 3 days before the event up to the day before the event starts.
- **During:** Conversations occurring between the official start and end dates of the event.
- **After:** Conversations occurring from the day after the event ends to 3 days after.

This segmentation allows us to analyze how affective polarization evolves over time, particularly whether polarization intensifies in anticipation of an event (before), peaks during the event (during), or escalates in response to its aftermath (after). The selection of these events is detailed in Section 3. By structuring our dataset with these temporal segments, we provide a granular view of online discourse dynamics, enabling comparisons across different event types and their corresponding shifts in polarization levels.

2.2 LLM-Based Classification

We employed LLaMA 3.1 70B, an open-source pretrained model specialized in understanding and generating human-like text with structured prompt engineering to classify:

- **Tweet Stance:** *Belief, Disbelief, Do Not Know* (for climate change); *Pro, Anti, Do Not Know* (for gun control).

- **Tweet Affect:** Whether the tweet contains emotionally charged language indicative of affective polarization.
- **Agreement Level:** Whether the child tweet agrees or disagrees with the parent tweet.

To ensure deterministic outputs, we configured the model to minimize randomness in classification by setting the temperature of the LLM to zero.

2.3 Affective Polarization Scoring with Heuristic Rules

Building upon the LLM-extracted attributes from the previous section, we define the affective polarization score using heuristic rules. These rules account for critical aspects of interaction dynamics between the parent tweet (original post) and the child tweet (reply), considering three key factors: stance alignment, which determines whether the reply tweet aligns or opposes the stance of the parent tweet; affective expression, assessing whether any of the tweets exhibit strong emotional language indicative of polarization; and agreement level, evaluating whether the reply tweet explicitly agrees or disagrees with the parent tweet. By combining LLM-based classification with structured heuristic rules, we enable a systematic, scalable, and interpretable quantification of affective polarization across social media discourse.

In our scoring system in Table 1, low scores (0 to 4) reflect constructive interactions and respectful disagreements. Specifically, a score of 0 represents the ideal, highlighting civil exchanges that actively seek mutual understanding and collaboration despite differing views, fostering openness and healthy dialogue. Conversely, high scores (8 and 10) represent interactions dominated by affective polarization, including heated conflict, incivility, and emotional hostility. A score of 10 is especially concerning, indicating interactions entirely within ideological echo chambers that reinforce negative emotions without exposure to opposing perspectives, thus intensifying polarization and potentially driving further hostility. High polarization scores, particularly in the upper half of the spectrum in Table 1, represent a shift away from productive dialogue toward emotional reactivity and out-group animosity, undermining effective communication and reinforcing intolerance. The stark contrast between scores 0 and 10 emphasizes the importance of fostering respectful, diverse discourse.

Importantly, our proposed heuristic scoring approach provides a significant advantage over traditional statistical methods by effectively quantifying polarization even in small conversations consisting of a single interaction. The overall polarization score for longer conversations is then computed by averaging the polarization scores across all individual interactions, ensuring scalability and adaptability to varied conversation lengths and structures.

2.4 An Illustrative Example: Prompting, Annotation, and Scoring

To demonstrate the combined effect of LLM-based annotation and predefined scoring, we present an illustrative example using the LLaMA 3.1 70B model.

Table 1. Heuristic Rules for Polarization Scoring. The table defines interaction categories based on stance similarity, affect presence, and agreement between tweets.

Reply vs. Parent Stance	Affect in Reply or Parent	Agreement	Score	Discourse Quality Category and Description
Opposite Stance	No (in both)	Yes	0	**Constructive Dialogue** - Constructive discussion with no negative affect and mutual agreement.
Same Stance	No (in both)	No	2	**Cordial Disagreement** - Healthy debate within the same stance group.
Opposite Stance	No (in both)	No	4	**Respectful Disagreement** - Polite disagreement across opposing stance groups without negative emotions.
Same Stance	No (in both)	Yes	6	**Echoic Agreement** - Echo chamber effect but without emotional hostility.
Opposite Stance	Yes (in either)	Yes	6	**Hard-Fought Agreement** - Agreement across opposing stances but with negative emotions.
Opposite Stance	Yes (in either)	No	8	**Heated Conflict** - Hostile disagreement between opposing stance groups.
Same Stance	Yes (in either)	No	8	**Discordant Allies** - Intra-group disagreement but with strong negative emotions.
Same Stance	Yes (in either)	Yes	10	**Polarizing Echo Chamber** - Echo chamber with emotional reinforcement, no interaction with opposing stance.

Table 2. Illustrative example of LLM-based annotation. This table presents simulated tweets resembling real discussions to ensure anonymity, along with LLM-extracted attributes such as stance, affect, agreement information, and their explanations by the language model employed.

Attribute	Parent Tweet	Reply Tweet
Tweet Text	'The science is clear. Climate change is real, and urgent action is needed before it's too late!'	'Climate change is a hoax pushed by the corrupt elites to brainwash the masses. Wake up, fool!'
Stance	**Belief**	**Disbelief**
Explanation	The tweet clearly states that climate change is real, indicating a belief in the topic.	This tweet expresses disbelief in climate change, calling it a hoax and implying that those who believe in it are brainwashed. The tone is confrontational and dismissive.
Affect	**No**	**Yes**
Explanation	Although the tone of the tweet is urgent and emphasizes the need for action, it does not contain deeply negative emotions or attitudes toward opposing views.	The tweet contains deeply negative emotions towards people who hold opposing views on climate change, calling them 'fool' and implying they are brainwashed. This language is emotionally charged and divisive.
Agreement	-	**No**
Explanation	-	The parent tweet and the reply have opposing views on climate change. The reply denies its existence, while the parent tweet affirms it and calls for urgent action. They clearly disagree on the topic.

Table 2 showcases the extracted stance, affect, and agreement attributes for a simulated conversation thread with just two tweets, while Table 1 illustrates the heuristic rules. The annotations in Table 2 were generated using a prompt template where the model was provided with paired tweets in the following format:

```
Tweet1: [Reply Tweet]
Tweet2: [Parent Tweet]
```

In cases where Tweet1 was the parent tweet (i.e., the first tweet in a thread), Tweet2 was provided as an empty string (""), and agreement-related fields were not applicable. Along with this input, the model received the following function-call instructions:

```
Analyze the content of the provided tweets to assess their stance and
emotional tone with respect to the topic: climate change.
Provide detailed classifications for stance, agreement between tweets,
and affective polarization (emotional negativity towards opposing views).
```

The function call requested the model to generate six fields, each with specific instructions:

- `tweet1_stance_explanation`: Provide a brief explanation of tweet1's stance on the topic climate change. If tweet1 expresses belief that climate change is real or expresses disbelief, explain the reasoning. If the stance is unclear, label it as don't know.
- `tweet1_stance`: Classify tweet1's stance on the topic: climate change. Possible values: belief, disbelief, don't know. This classification should be based on the explanation provided in 'tweet1_stance_explanation'.
- `tweets_agreement_explanation`: Provide an explanation of whether tweet1 and tweet2 agree or disagree on the topic: climate change. Agreement indicates similar views; disagreement means opposing views. If tweet2 is not available, state 'not applicable'. If the agreement is unclear, provide reasoning.
- `tweets_agreement`: Classify the agreement between tweet1 and tweet2 with respect to the topic: climate change. Possible values: yes (agreement), no (disagreement), don't know (unclear).
- `tweet1_affect_explanation`: Explain whether tweet1 contains deeply negative emotions or attitudes specifically towards people who hold opposing views on the topic: climate change. The focus is on emotional negativity beyond the stance itself.
- `tweet1_affect`: Classify tweet1's affective polarization, i.e., emotional negativity specifically towards opposing views on the topic: climate change. Possible values: yes (contains affective polarization), no (doesn't contain), don't know (uncertain).

This prompt structure was applied consistently to generate the annotations presented in Table 2 and throughout our experiments. Based on the extracted

attributes by the LLM-based annotations and applying the heuristic rules to the extracted features, the reply tweet is classified as Heated Conflict (Score = 8), as it: (1) holds the opposite stance, (2) contains emotionally charged language, and (3) expresses disagreement, indicating hostility between opposite stance groups. This example highlights a simple illustration of our approach and the ability of the framework to systematically evaluate and categorize affective polarization.

2.5 Implementation Details

The classification pipeline was implemented using LLaMA 3.1 70B via Ollama for stance and affect classification, combined with Langchain for structured prompt engineering and deterministic response generation. Our initial dataset encompasses a total of $2,551$ conversations across the climate change-related events and $3,201$ conversations across the gun control-related events, considering different timeframes for each event. This hybrid approach provides an efficient and interpretable method for large-scale affective polarization quantification, bridging computational advancements in LLMs with human-guided analysis.

3 Case Studies & Results

To systematically analyze affective polarization on social media, we examined its evolution across two case studies as previously mentioned: climate change and gun control. Our framework quantifies polarization before, during, and after key events, allowing us to observe distinct temporal patterns in online discourse. The following subsections present findings for each case study, highlighting how different types of events influence the dynamics of affective polarization.

3.1 Affective Polarization in Climate Change Discourse

In this section, we analyzed discussions surrounding major climate change-related events as one of the case studies for the current research. Specifically, we examined how affective polarization levels fluctuate before, during, and after four key climate-related events: the Intergovernmental Panel on Climate Change (IPCC) assessment report release, the United Nations (UN) Environment Assembly, Hurricane Ida and a major heatwave. To provide a structured overview, Table 3 presents the starting and ending timeframes of the considered events. For each event, we computed the mean affective polarization scores and the standard errors across all conversations before, during, and after the event. The results are illustrated in Figure 2. As shown in the figure, affective polarization scores increase noticeably during and after these events compared to the preceding period. This aligns with expectations that major climate-related developments amplify emotional intensity in public discourse. Additionally, the volume of conversations (denoted by 'n' in the figure) also significantly increases during and after these events, reflecting heightened social media interactions.

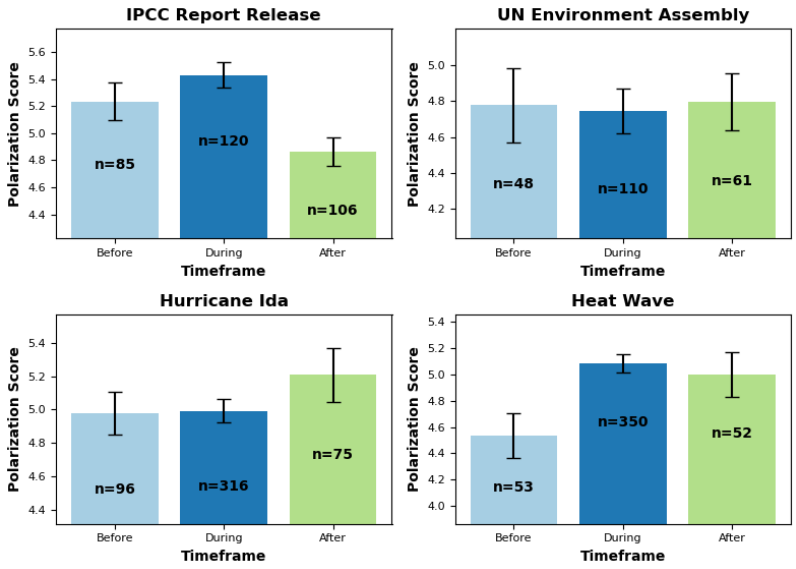

Fig. 2. Affective polarization scores across different timeframes (before, during, and after) for four climate change-related events. The bar plots illustrate the average polarization score, with error bars representing the standard error of the mean. The variable n in each bar denotes the number of conversations analyzed during the corresponding timeframe for each event.

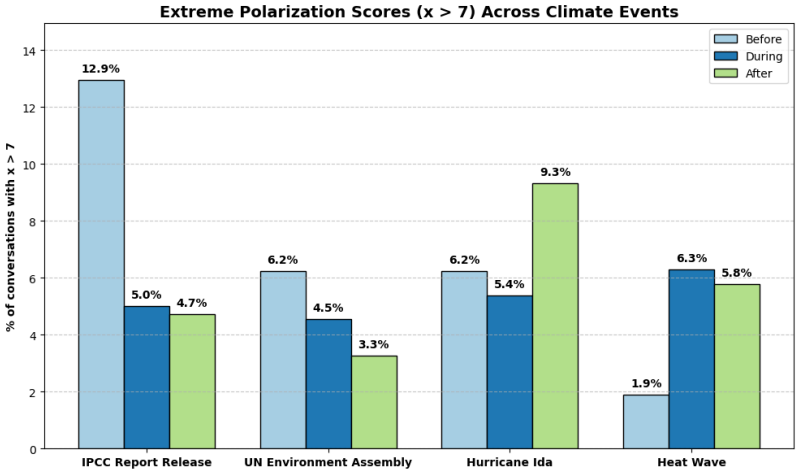

Fig. 3. Percentage of conversations exhibiting extreme polarization scores ($x > 7$) across different timeframes (before, during, and after) for climate change-related events.

To better understand the intensity of polarization, we focused exclusively on conversations exhibiting extreme polarization (scores greater than 7). Figure 3 shows the percentage of highly polarized conversations during different timeframes (before, during, and after) for each climate-related event. The analysis reveals distinct patterns: well-publicized events such as the IPCC assessment report release and the UN Environment Assembly displayed heightened extreme polarization primarily before these events, reflecting anticipation and ideological engagement. Conversely, unpredictable climate events like Hurricane Ida and the heatwave experienced spikes in extreme polarization predominantly during and after the events, highlighting the role of spontaneous emotional responses and reactive discourse in shaping polarization patterns.

Table 3. Climate change-related events analyzed in this study, including their start and end dates.

Event	Start Date	End Date
IPCC Assessment Report Release	2021-08-09	2021-08-09
Hurricane Ida	2021-08-26	2021-09-04
Major Heatwave	2021-06-25	2021-07-07
UN Environment Assembly	2022-02-28	2022-03-02

These findings suggest that the temporal characteristics of affective polarization in climate-related discussions are closely linked to the nature of the triggering event. Anticipatory polarization is more prominent for scheduled, policy-driven events, whereas reactive polarization is dominant for sudden climate-related disasters.

3.2 Affective Polarization in Gun Control Discourse

This section presents an analysis of discussions surrounding major gun control-related events to examine affective polarization trends in the aftermath of mass shootings. Specifically, similar to the previous case study, we analyzed how affective polarization fluctuated before, during, and after four major gun-related incidents: the Texas Robb Elementary School Shooting, the Illinois Highland Park Parade Shooting, Multiple Shootings in Maryland, Illinois, and Virginia, and the Colorado Spring Nightclub Shooting. Table 4 presents the starting and ending timeframes for the considered events. For each event, we computed the mean affective polarization scores and the standard errors across all social media conversations before, during, and after the event. As seen in Figure 4, affective polarization scores increase significantly during and after these events compared to the period before. This pattern suggests that mass shootings amplify ideological divides in gun control discussions, with emotional intensity peaking n response to the event. The number of conversations (denoted by 'n' in the figure) also substantially increases during and after these events, reflecting heightened online engagement.

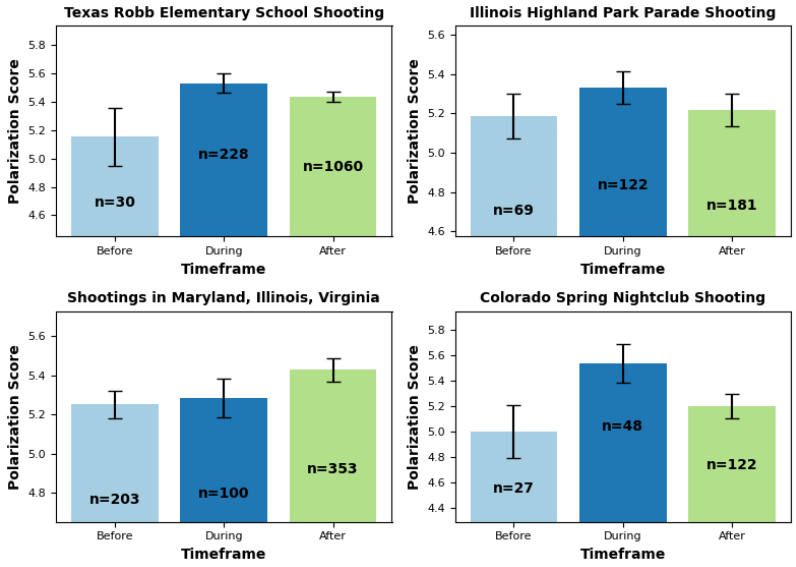

Fig. 4. Affective polarization scores across different timeframes (before, during, and after) for four gun control-related events. The bar plots display the average polarization score, with error bars indicating the standard error of the mean. The variable n in each bar represents the number of conversations analyzed during the corresponding timeframe for each event.

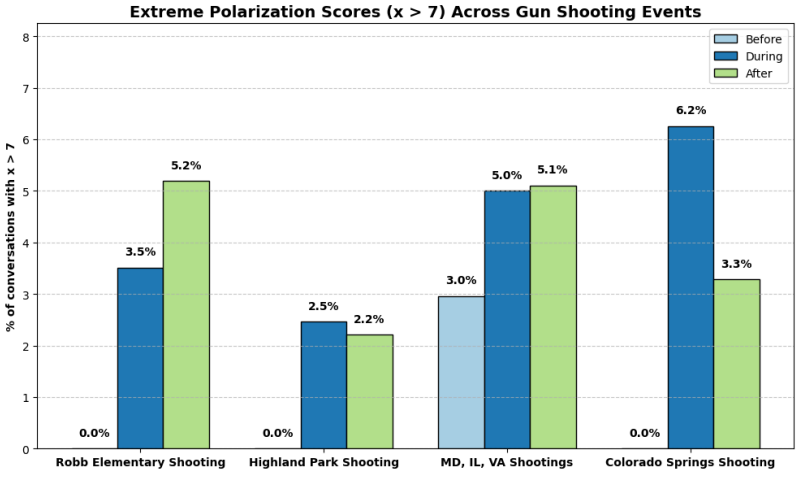

Fig. 5. Percentage of conversations exhibiting extreme polarization scores ($x > 7$) across different timeframes (before, during, and after) for gun control-related events.

Table 4. Gun control-related events analyzed in this study, including the starting and ending dates.

Event	Start Date	End Date
Texas Robb Elementary School Shooting	2022-05-24	2022-05-24
Illinois Highland Park Parade Shooting	2022-07-04	2022-07-04
Multiple Shooting in Maryland, Illinois, Virginia	2022-06-07	2022-06-07
Colorado Spring Nightclub Shooting	2022-11-19	2022-11-20

Similar to our previous analysis, we exclusively examined conversations with extreme polarization scores (greater than 7). Figure 5 illustrates the proportion of highly polarized conversations occurring before, during, and after four gun-related events. The analysis clearly indicates that extreme polarization is notably elevated during and after mass shooting incidents. Unlike climate change events, which often saw spikes in extreme polarization before anticipated events, gun control-related discussions primarily demonstrate reactive polarization, spiking directly in response to shootings.

4 Discussion

In this study, we introduced a novel framework combining large language models (LLMs) and predefined heuristics to quantify affective polarization in online discussions on charged topics like climate change and gun control. Our hybrid approach uses LLMs to efficiently extract stance, affect, and agreement dynamics at scale, while domain experts guide the polarization scoring process through intuitive rules. The primary goal is not benchmarking LLM performance, but presenting a structured, human-in-the-loop framework that accelerates affective polarization analysis.

Our results show that climate change and gun control events significantly influence affective polarization. Climate discussions exhibited anticipation-driven polarization, with extreme opinions emerging before events like the IPCC report release or the UN Environment Assembly. Gun control debates showed a reactive pattern, with polarization intensifying primarily after mass shootings. The volume of conversations also surged following these events, highlighting real-world triggers in online discourse. Although most conversations remained moderately polarized, spikes in extreme polarization aligned closely with major events, providing deeper insights into the temporal dynamics of affective polarization.

Future work could extend this analysis by including a broader range of events and topics, leveraging the generalizability of our proposed framework. Validating LLM-based polarization scoring across diverse sociopolitical contexts and exploring interventions to mitigate extreme polarization could also offer valuable insights for policymaking and social media governance. In summary, our study highlights the benefits of combining AI-driven analysis with domain informed frameworks, enabling interpretable measurement of online polarization, even in brief interactions.

Acknowledgment

This research was supported by the Army Research Office under Grant W911NF-22-1-0035. The views expressed are those of the authors and do not necessarily reflect the official policies of the Army Research Office or the U.S. Government. The U.S. Government retains the right to reproduce and distribute reprints for governmental purposes.

References

1. AlDayel, A., Magdy, W.: Stance detection on social media: State of the art and trends. Information Processing & Management 58(4), 102597 (2021)
2. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in Neural Information Processing Systems 33, 1877–1901 (2020)
3. Chae, Y., Davidson, T.: Large language models for text classification: From zero-shot learning to fine-tuning. Open Science Foundation 10 (2023)
4. Chambers, D.: Social media and personal relationships: Online intimacies and networked friendship. Springer (2013)
5. Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrociocchi, W., Starnini, M.: The echo chamber effect on social media. Proc. National Academy of Sciences 118(9), e2023301118 (2021)
6. Dahlgren, P.: Media and political engagement: Citizens, communication and democracy. Cambridge University Press (2009)
7. Druckman, J.N., Klar, S., Krupnikov, Y., Levendusky, M., Ryan, J.B.: (Mis) estimating affective polarization. The Journal of Politics 84(2), 1106–1117 (2022)
8. Feldman, D., Rao, A., He, Z., Lerman, K.: Affective polarization in social networks. arXiv e-prints pp. arXiv–2310 (2023)
9. Garimella, K., De Francisci Morales, G., Gionis, A., Mathioudakis, M.: Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship. In: Proc. 2018 World Wide Web Conference. pp. 913–922 (2018)
10. Gillani, N., Yuan, A., Saveski, M., Vosoughi, S., Roy, D.: Me, my echo chamber, and I: Introspection on social media polarization. In: Proc. 2018 World Wide Web Conference. pp. 823–831 (2018)
11. Guerra, P., Meira Jr, W., Cardie, C., Kleinberg, R.: A measure of polarization on social media networks based on community boundaries. In: Proc. International AAAI Conference on Web and Social Media. vol. 7, pp. 215–224 (2013)
12. Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., Westwood, S.J.: The origins and consequences of affective polarization in the United States. Annual Review of Political Science 22(1), 129–146 (2019)
13. Khan, N.D., Khan, J.A., Li, J., Ullah, T., Zhao, Q.: Leveraging Large Language Model ChatGPT for enhanced understanding of end-user emotions in social media feedbacks. Expert Systems with Applications 261, 125524 (2025)
14. Lerman, K., Feldman, D., He, Z., Rao, A.: Affective polarization and dynamics of information spread in online networks. npj Complexity 1(1), 8 (2024)
15. Linegar, M., Kocielnik, R., Alvarez, R.M.: Large language models and political science. Frontiers in Political Science 5, 1257092 (2023)
16. Marlowe, J.M., Bartley, A., Collins, F.: Digital belongings: The intersections of social cohesion, connectivity and digital media. Ethnicities 17(1), 85–102 (2017)

17. Martínez-España, R., Fernández-Pedauye, J., De Lucía, J.G.P., Rojo-Martínez, J.M., Bakdid-Albane, K., García-Escribano, J.J.: Methodology for measuring individual affective polarization using sentiment analysis in social networks. IEEE Access (2024)
18. Mets, M., Karjus, A., Ibrus, I., Schich, M.: Automated stance detection in complex topics and small languages: The challenging case of immigration in polarizing news media. PLoS ONE 19(4), e0302380 (2024)
19. Overgaard, C.S.B.: Perceiving affective polarization in the United States: how social media shape meta-perceptions and affective polarization. Social Media+ Society 10(1) (2024)
20. Pendyala, V.S., Hall, C.E.: Explaining Misinformation Detection Using Large Language Models. Electronics 13(9), 1673 (2024)
21. Piccardi, T., Saveski, M., Jia, C., Hancock, J.T., Tsai, J.L., Bernstein, M.: Social Media Algorithms Can Shape Affective Polarization via Exposure to Antidemocratic Attitudes and Partisan Animosity. arXiv preprint arXiv:2411.14652 (2024)
22. Rani, N., Walia, R., et al.: A Comprehensive Review of Sentiment Analysis: Techniques, Datasets, Limitations, and Future Scope. In: 2024 Sixth International Conference on Computational Intelligence and Communication Technologies (CCICT). pp. 403–409. IEEE (2024)
23. Rashid, R., Melton, J., Ghorbani, O., Krishnan, S., Reid, S., Terejanu, G.: Quantifying Influencer Impact on Affective Polarization. In: 2024 International Conference on Machine Learning and Applications (ICMLA). pp. 1135–1140 (2024)
24. Rodilosso, E.: Filter bubbles and the unfeeling: How AI for social media can foster extremism and polarization. Philosophy & Technology 37(2), 71 (2024)
25. Saaida, M.: The role of social media in shaping political discourse and propaganda. Science for All Publication 3(2), 1–8 (2023)
26. Serrano-Puche, J.: Digital disinformation and emotions: exploring the social risks of affective polarization. International Review of Sociology 31(2), 231–245 (2021)
27. Suárez Estrada, M., Juarez, Y., Piña-García, C.: Toxic social media: Affective polarization after feminist protests. Social Media+ Society 8(2) (2022)
28. Thareja, R.: Multimodal sentiment analysis of social media content and its impact on mental wellbeing: An investigation of extreme sentiments. In: Proc. 7th Joint Int. Conf. on Data Science & Management of Data. pp. 469–473 (2024)
29. Tyagi, A., Uyheng, J., Carley, K.M.: Affective polarization in online climate change discourse on Twitter. In: 2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). pp. 443–447. IEEE (2020)
30. Wang, X., Li, X., Yin, Z., Wu, Y., Liu, J.: Emotional intelligence of large language models. Journal of Pacific Rim Psychology 17 (2023)
31. Yair, O.: A note on the affective polarization literature. SSRN 3771264 (2020)
32. Yarchi, M., Baden, C., Kligler-Vilenchik, N.: Political polarization on the digital sphere: A cross-platform, over-time analysis of interactional, positional, and affective polarization on social media. In: Dissonant Public Spheres, pp. 185–226. Routledge (2024)
33. Yu, X., Wojcieszak, M., Casas, A.: Partisanship on Social Media: In-Party Love Among American Politicians, Greater Engagement with Out-Party Hate Among Ordinary Users. Political Behavior 46(2), 799–824 (2024)
34. Zhang, W., Deng, Y., Liu, B., Pan, S.J., Bing, L.: Sentiment analysis in the era of large language models: A reality check. arXiv preprint arXiv:2305.15005 (2023)
35. Zhang, Y., Sharma, K., Du, L., Liu, Y.: Toward mitigating misinformation and social media manipulation in LLM era. In: Proc. ACM Web Conference 2024. pp. 1302–1305 (2024)