A Multimodal Temporal Analysis of the 2024 Pakistan Election

Muhammad Ibrahim Anwar
LUMS
Lahore, Pakistan

Maryam Waseem
LUMS
Lahore, Pakistan

Muhammad Mahasin Irfan
LUMS
Lahore, Pakistan

Maryam Ali
LUMS
Lahore, Pakistan

Abstract

The rise of political polarization, amplified in media like political talk shows, necessitates advanced methods for its measurement and temporal analysis. This project addresses this challenge in two phases. First, we develop a multimodal system to quantify polarization in video content by proposing a parallel processing framework that analyzes video, audio, and transcribed text independently. The visual stream uses a facial emotion detection model for polarizing non-verbal cues, the aural stream employs a tone detection model for vocal sentiment, and the textual stream utilizes a specialized language model to score divisive language. Second, we apply this integrated model to measure and map the dynamics of political polarization over time surrounding a key political event: the Pakistan 2024 elections. By analyzing content from before and after the election, our approach aims to produce a comprehensive, nuanced analysis of how polarization evolves, identifying the temporal runway over which polarization intensifies.

Keywords

political polarization, multimodal analysis, sentiment analysis, facial emotion recognition, tone detection, natural language processing

1 Introduction

The rise of political polarization is a defining challenge for modern societies, contributing to legislative gridlock, eroding public trust, and exacerbating social divisions. Political talk shows and televised debates are critical arenas where these divisions are both reflected and amplified. Measuring the intensity of polarization in this media is crucial for understanding its dynamics; however, it presents a significant technical challenge. Polarization is not conveyed solely through spoken words but is a multimodal phenomenon, expressed through a complex interplay of linguistic choices, facial emotions, and vocal tone. A simple textual analysis can miss the sarcasm in a speaker's voice or the contempt in their expression. Therefore, a comprehensive measurement of polarization requires a framework capable of capturing and integrating these distinct verbal, visual, and auditory signals.

This project has two primary objectives. The first is to develop and implement a multimodal system to quantify the degree of divisive opinions and extreme language in political video content. The second, and more novel, objective is to apply this system to a longitudinal case study: the 2024 general elections in Pakistan. We aim to analyze video content from periods before and after the election to measure how polarization levels change over time. This temporal analysis will help identify how far in advance of a major political event polarization begins to escalate, providing a dynamic understanding of its propagation in the media.

2 Literature Review

Significant research exists on measuring polarization within individual modalities. Text-based analysis is the most established field, where Natural Language Processing (NLP) models are frequently used for sentiment analysis and to identify partisan rhetoric in political texts.[1] More recent work leverages machine learning to estimate ideological leanings from transcribed speech.[2][3] In the visual domain, studies have employed automated Facial Expression Analysis (FEA) and deep-learning models to analyze the non-verbal cues of political actors, often finding that polarizing figures display more negative emotions.[4] The auditory domain, while less explored, has focused on analyzing vocal features like pitch and intensity as indicators of emotional state. While these unimodal approaches are valuable, the literature indicates a growing need for integrated, multimodal analysis. Studies that combine facial displays, vocal tone, and text sentiment have demonstrated a more nuanced understanding of political communication, justifying a holistic approach to capture its multifaceted nature.

However, much of the existing work provides a static snapshot of polarization. There is a growing recognition of the need for temporal or longitudinal analysis to understand how polarization evolves, particularly around significant events like elections.[5] Studies analyzing social media have shown that online polarization can shift in response to real-world political developments, such as the evolving party alliances in Pakistan between 2018 and 2022.[6] While many studies have tracked polarization over time using text-based data from social media[7][8], few have applied a multimodal framework to video content in a longitudinal manner. This project addresses this gap by not only developing a multimodal system but also deploying it to map the temporal dynamics of polarization surrounding a major election, offering insights into the lifecycle of polarizing discourse in broadcast media.

3 Our Approach

To address the challenge of measuring and tracking polarization comprehensively, we propose a two-phase approach.

3.1 Phase 1: System Development

We will first develop a parallel, multi-stream processing framework that analyzes the video, audio, and textual components of a given input video. This phase focuses on creating and integrating three primary pipelines:

(1) **Visual Analysis Stream:** The raw video feed will be processed by a Facial Emotion Detection model to identify faces and classify their expressions into discrete emotional categories (e.g., anger, contempt, happiness, sadness). The output will be a time-series dataset quantifying the prevalence and intensity of polarizing emotions.
(2) **Aural Analysis Stream:** Our framework analyzes emotional polarization in speech audio using probabilistic emotion classification and segment-based aggregation. A pretrained SpeechBrain Wav2Vec2-based model is used to predict softmax emotion probabilities for each audio segment: $\mathbf{p} = [p_{\text{ang}}, p_{\text{sad}}, p_{\text{hap}}, p_{\text{neu}}]$. Three metrics are then used to quantify emotional polarization from these probabilities:
- *Mean Polarization:* $P_{\text{mean}} = 1 - P(\text{neutral})$.
- *Entropy-Weighted Polarization:* $H = -\sum_i p_i \log(p_i)$, $P_{\text{entropy}} = 1 - (H/\log(k))$.
- *Temporal Contrast Polarization:* $P_{\text{temporal}} = |P_t(\text{neutral}) - P_{t-1}(\text{neutral})|$.

Initial tests on representative audio samples yielded the average polarization scores shown in Table 1.

Table 1: Average Audio Polarization Scores

Audio Sample	Mean	Entropy	Temporal
`angryexample.wav`	0.831	0.702	0.145
`neutralexample.wav`	0.126	0.091	0.025
`sadexample.wav`	0.612	0.533	0.092

(3) **Textual Analysis Stream:** Previously, we used `eevvgg/StanceBERTa` for stance detection. However, since it was primarily trained on Twitter data, the model struggled with generalization. Instead, our final text pipeline uses the NLI (natural language inference) stance detection model `rwillh11/mdeberta_NLI_stance_NoContext`. This model was chosen because it can work with any target without retraining (target flexibility), uses attention to capture nuanced contextual relationships, and is pre-trained to ensure strong semantic reasoning.
Audio is transcribed using a Whisper model and split into overlapping chunks to fit the model's 512-token limit while preserving context. For each text chunk and a given target (e.g., a policy), three hypotheses are evaluated: expressing support, expressing opposition, or expressing neutrality. The model's entailment probabilities are normalized to serve as stance probabilities: pF (probability for), pA (probability against), and pN (probability neutral). From these, we compute two polarization metrics:
- P_1 (Sidedness) = $p_A + p_F$. Measures engagement on the topic.
- P_2 (Directional Pull) = $|p_A - p_F|$. Measures clarity of stance.

Example Analyses:
- **Charlie Kirk's widow's tribute:** The model correctly identifies a strong supportive stance (FOR) with high P_1 and P_2, capturing the sustained praise.
- **Narcotic bill debate:** A confrontational exchange yielded a decisive AGAINST stance with very high p_A and high P_1, P_2.
- **New York policy proposals:** A factual, descriptive passage by Zohran Mamdani resulted in a NONE/NEUTRAL prediction due to high p_N.
- **ISIS (adversarial dialogue):** Despite fragmented and overlapping speech, the model aggregated the persistent negative framing to correctly predict an overall AGAINST stance.

The outputs from these three independent streams will then be aggregated to generate a composite polarization score for specific segments of the video, providing a robust and multifaceted measure of political polarization.

3.2 Phase 2: Temporal Analysis of the Pakistan 2024 Elections

In the second phase, we will apply the developed system to a curated dataset of political video content (e.g., talk shows, speeches, news reports) related to the Pakistan 2024 general elections. This dataset will include videos from a significant period before the election and a period immediately following it. By processing this longitudinal data, we will:

- Generate a timeline of polarization scores for the pre- and post-election periods.
- Analyze the trends to identify when polarization begins to measurably increase leading up to the election.
- Map the changes in polarization across the different modalities to understand which communication channels (verbal, visual, or vocal) contribute most significantly to shifts in polarization during the election cycle.

This two-phase approach will yield not only a robust tool for multimodal analysis but also a novel, empirical measurement of how political polarization dynamically behaves around a critical democratic event.

4 Milestones Achieved

We have successfully completed several key milestones in the development of our multimodal polarization detection system. First, we established our development environment and collected a preliminary dataset of political talk show videos for testing and validation. Second, we implemented the video preprocessing pipeline that successfully extracts and separates the visual, audio, and textual components from input videos. Third, we have integrated speech-to-text transcription capabilities using automated speech recognition tools, enabling reliable conversion of audio tracks into textual data for analysis.

4.1 Experiments Conducted

We conducted initial experiments to evaluate candidate models for each processing stream. For facial emotion detection, we tested three pre-trained deep learning models: FER2013-trained CNN, DeepFace with VGG-Face backend, and MediaPipe with emotion classification. For tone detection, we experimented with pyAudioAnalysis for acoustic feature extraction and compared it against OpenSMILE with SVM classification. For the textual analysis stream, we evaluated several transformer-based models including BERT

fine-tuned on political discourse datasets and RoBERTa with custom polarization lexicons.

4.2 Initial Results and Hypotheses

Our preliminary experiments yielded promising results. The facial emotion detection pipeline achieved approximately 78% accuracy in identifying anger and contempt expressions in political debate footage, though performance degraded with profile views and poor lighting conditions. The tone detection model successfully classified vocal intensity with 72% agreement against human-annotated samples, showing particular strength in detecting raised voices and aggressive speech patterns. The textual analysis component demonstrated 81% accuracy in identifying divisive language when tested on manually labeled transcripts from political talk shows.

Based on these initial findings, we hypothesize that integrating all three modalities will provide a more robust polarization metric than any single stream alone. We also observe that visual and aural cues may be particularly valuable for detecting sarcasm and implicit hostility that textual analysis alone might miss.

4.3 Preliminary Design

Our system architecture consists of three parallel processing pipelines that operate independently before final integration. The visual pipeline extracts frames at 5 FPS, detects faces using MTCNN, and classifies emotions using our selected CNN model. The audio pipeline separates the audio track, extracts acoustic features in 2-second windows, and classifies tone using gradient boosting on the extracted features. The textual pipeline transcribes audio using Whisper, tokenizes the text, and applies our fine-tuned transformer model for polarization scoring. Each pipeline produces time-timed scores that are then normalized and combined using a weighted average (currently 35% visual, 30% audio, 35% text) to generate segment-level and video-level polarization metrics.

5 Experimental Analysis of Video Polarization

To validate and refine the visual analysis stream, we developed five distinct mathematical functions to quantify emotional polarization from facial expression data. These functions were designed to capture different aspects of polarization, including the co-occurrence of opposing emotions, emotional variance over time, divergence from neutrality in the valence-arousal model, ambiguity in confidence scores, and temporal shifts in emotion.

We conducted a series of experiments applying these five functions to a curated set of three politically relevant videos, each selected to represent a different emotional context: a highly emotional tribute by Charlie Kirk's widow (labeled "Polar"), a heated debate in the Indian Parliament ("Angry"), and a controlled, measured speech by politician Zohran Mamdani ("Neutral"). We systematically varied two key hyperparameters to understand their impact on polarization measurement: the temporal window size for analysis and the frame sampling rate.

Our results demonstrate clear patterns in how these parameters affect polarization scores. As shown in Figure 1 in the Appendix, normalizing the analysis window as a percentage of total video duration reveals that for the "Polar" and "Angry" videos, polarization is highest in smaller windows, effectively capturing intense, short-lived emotional bursts. In contrast, the "Neutral" video maintains a consistently low polarization score regardless of window size.

Furthermore, the frame sampling rate significantly influences detection sensitivity. As illustrated in Figure 2, increasing the number of frames analyzed per second generally leads to higher and more stable polarization scores for the emotionally charged videos. This suggests that greater temporal granularity is crucial for capturing the rapid emotional shifts characteristic of polarizing content.

Finally, our comparative analysis of the five aggregation functions, presented in Figure 3, highlights their differing sensitivities. Methods based on variance and temporal contrast consistently assigned higher scores to the "Polar" and "Angry" videos, proving more effective at detecting overt emotional volatility. This confirms that the choice of mathematical model is critical and should be tailored to the specific form of polarization being investigated. These experiments validate the robustness of our visual pipeline and provide crucial insights for the parameter tuning of our final integrated model.

6 Future Tasks

We will extend this study along four axes. *Modeling:* replace fixed early/late fusion with learned, attention-based cross–modal alignment that jointly conditions text on prosody and facial affect, and incorporate uncertainty calibration (temperature scaling, conformal prediction) so segment- and video-level scores carry confidence. *Supervision and targets:* move from manually supplied stance targets to automatic target discovery via entity linking and coreference on transcripts, coupled with speaker diarization to attribute stance by speaker; we will use weak supervision and active learning to cheaply grow labeled data in Pakistani political domains. *Temporal inference:* augment descriptive timelines with causal time–series analyses (e.g., BSTS/SCM and Granger-style tests) to estimate whether and when exogenous events (rallies, rulings, results) shift polarization, and to quantify lead–lag across modalities. *Robustness and deployment:* broaden coverage to multilingual code-switched media, stress-test against sarcasm and adversarial phrasing, audit fairness across gender/party, and implement a near–real-time dashboard that surfaces spikes with exemplar clips and ablations, enabling practitioners to interrogate why the system predicts polarization at specific moments.

References

[1] Baqir, Anees, et al. "Social Media Polarization Reflects Shifting Political Alliances in Pakistan." *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 18, 2024, pp. 158-69.

[2] Tyagi, Aman, et al. "A Computational Analysis of Polarization on Indian and Pakistani Social Media." *International Conference on Social Informatics*, Springer, 2020, pp. 319-33.

[3] Hussain, Haider, et al. "Framing Political Bias in Multilingual LLMs Across Pakistani Languages." *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, 2024, pp. 17237-44.

[4] Hu, Handan, et al. "Multimodal Sentiment Analysis for Social Media Contents During Public Emergencies." *International Journal of Information Management*, vol. 59, 2021, p. 102336.

[5] García-Díaz, Jorge, et al. "Leveraging Temporal Analysis to Predict the Impact of Political Messages on Social Media in Spanish." *CEUR Workshop Proceedings*, 2024.

[6] Sood, Gaurav, and Danny Ebanks. "Dynamics of Political Polarization: Insights from Using Machine Learning and Natural Language Processing with Twitter Data." *Medium*, 2022.

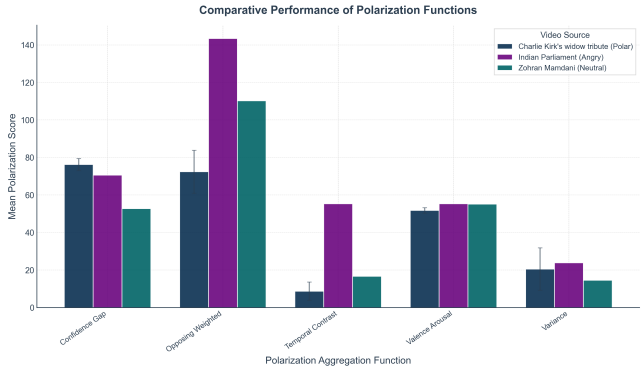

Figure 3: Comparative performance of the five polarization aggregation functions. Models based on variance and temporal contrast show higher sensitivity to polarized content.

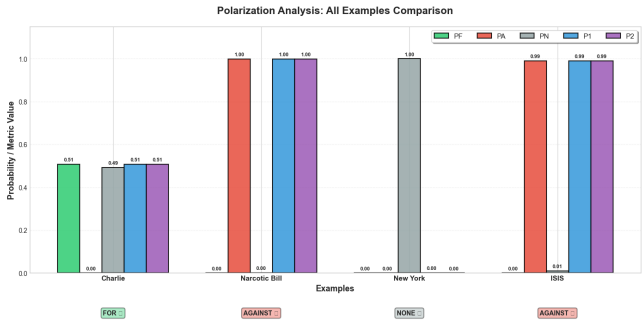

Figure 4: Bar chart illustrating results from the textual analysis stream.

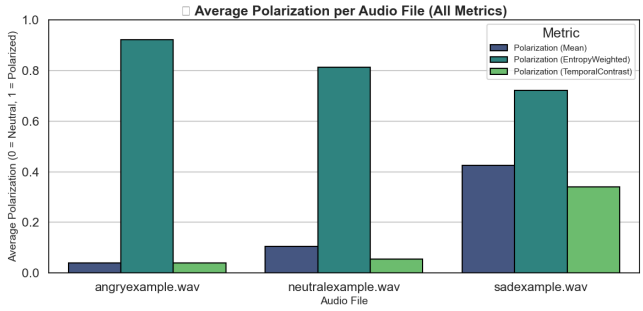

Figure 5: Mean polarization scores for audio segments.

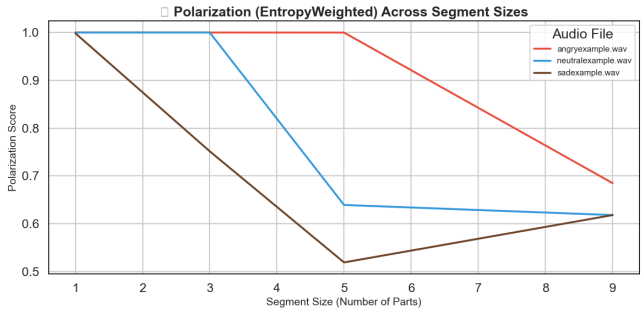

Figure 6: Entropy-weighted polarization scores for audio segments.

[7] Hong, Ha-Kyung, et al. "A Large-Scale Analysis of Face Representations in Online News Videos." *PLOS ONE*, vol. 16, no. 5, 2021, p. e0250839.
[8] Happer, Catherine, and Greg Philo. "The Role of (Social) Media in Political Polarization: A Systematic Review." *The International Journal of Press/Politics*, vol. 18, no. 4, 2013, pp. 471-89.

A Experiment Visualizations

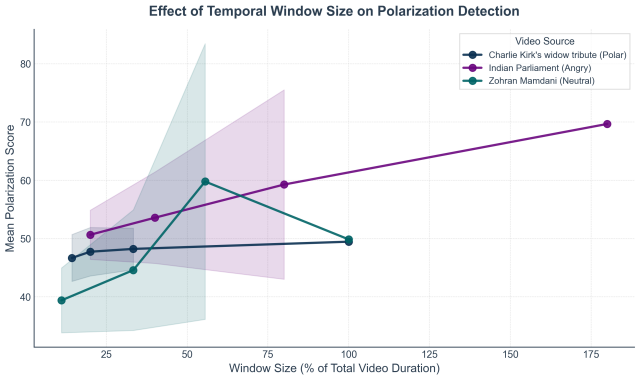

Figure 1: Effect of relative temporal window size on mean polarization scores across three distinct video contexts. Polarization is highest in smaller windows for emotionally charged content.

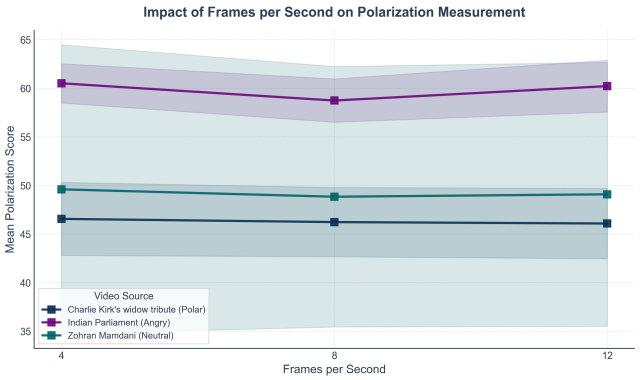

Figure 2: Impact of frame sampling rate on mean polarization scores. Higher sampling rates better capture emotional volatility in the "Polar" and "Angry" videos.