Proceedings of the Twentieth International AAAI Conference on Web and Social Media (ICWSM 2026)

Large-Scale Multimodal Content Analysis and Annotation with Vision-Language Models

Harsha Nemani[1], **Kiran Garimella**[2]

[1]International Institute of Information Technology
[2]Rutgers University
nemani.v@research.iiit.ac.in, kiran.garimella@rutgers.edu

Abstract

Most social media research still treats content as text, even though images and videos carry a large share of what people actually see and share. This gap is especially serious for WhatsApp in India, where political communication often travels as posters, memes, screenshots, short clips, and forwarded videos in Hindi and other low-resource languages. Text-only analysis can therefore miss a large fraction of political content and can distort conclusions about what topics spread and what becomes viral.

We present a large-scale multimodal analysis of WhatsApp data collected via data donation across roughly 100 locations in India during the 2024 Indian General Elections. Using recent vision-language models (VLMs) and large language models (LLMs), we build a multimodal processing toolkit that represents text, images, and videos in a shared framework. We use this toolkit to produce three results. First, we map topic prevalence and diffusion across modalities and show that topic distributions differ sharply by modality: key domains such as politics are disproportionately non-textual, so text-only pipelines systematically undercount them. Second, we evaluate whether LLM/VLM-based classifiers can replace human annotation for socially sensitive labels. Models perform well for broad categories such as political and news content, but they are substantially less reliable for misinformation, hate speech, and caste-coded hostility, with failures driven by class imbalance, implicit meaning, and cultural context. Third, we connect WhatsApp content to mainstream narratives by building a reference set from YouTube uploads of prime-time television news over the same period and measuring multimodal narrative overlap under format transformation (e.g., headlines as screenshots and segments as clips). The narratives that align with mainstream coverage are carried largely by non-text WhatsApp messages and are far more likely to be viral than baseline content.

Together, these findings show that multimodal methods are necessary for valid measurement of political communication on encrypted platforms, and they provide an auditable toolkit for studying diffusion, harms, and cross-channel narrative dynamics in low-resource settings.

1 Introduction

WhatsApp is one of the most important communication infrastructures in India. It is used for everyday coordination, but it is also a primary route through which political messages circulate, often with high credibility because they arrive through trusted social ties (Trauthig and Woolley 2025). A large research literature has documented WhatsApp's relevance for political communication, including the spread of misinformation and hate speech (Narayanan et al. 2019; Varanasi, Pal, and Vashistha 2022). At the same time, a basic fact about WhatsApp in India is still easy to underestimate: much of the content that moves at scale is not textual (Garimella and Eckles 2020). It is multimodal media such as posters, memes, screenshots, short clips, and forwarded videos, frequently mixing Hindi and other regional languages with visuals, audio, and symbolic cues. This creates a measurement problem. If we want to understand what people are exposed to and what spreads on WhatsApp, we need methods that can represent and analyze content across modalities and across low-resource languages, rather than treating WhatsApp as a text archive.

The Indian case also makes this measurement problem politically consequential. The ruling Bharatiya Janata Party (BJP) has invested heavily in digital campaigning and has built a professionalized social media operation, often described as an IT cell, with a substantial presence on WhatsApp (Mahapatra and Plagemann 2019). In parallel, multiple reports describe increasing political and corporate influence over mainstream media through ownership concentration, economic pressure, and constraints on editorial independence (International Journalists' Network 2024). Together, these conditions motivate a widely discussed hypothesis in journalism and political communication: political narratives may be produced and amplified across channels in a tightly coupled way, with mainstream media and WhatsApp reinforcing each other. The empirical difficulty is that cross-channel tracing is easiest when the objects are purely textual. In practice, the most portable objects between TV, YouTube, and WhatsApp are often screenshots and clips. A narrative from a television segment might be shared as a textual message, a screenshot (image) or a short form clip (video). If we cannot measure multimodal content, we cannot credibly study cross-channel narrative coupling in the form it actually takes.

This paper introduces a multimodal content processing toolkit for WhatsApp and applies it to a large dataset collected via data donation across roughly 100 locations in

northern India during the 2024 general election period. Using state-of-the-art large language models and vision-language models, the toolkit represents text, images, and videos in a shared framework, which lets us study topic prevalence and diffusion, evaluate how reliable model-based labels are for harms such as misinformation and hate speech, and measure narrative overlap between mainstream media coverage and WhatsApp circulation.

We organize the paper around three main results:

Result 1: Multimodal Topic Prevalence. We use LLM and VLM representations to estimate the prevalence of high-level topics across text, images, and videos, and we examine how topic distributions shift with virality using WhatsApp's "forwarded many times" signal. We show that key topics, especially politics, are heavily non-textual, and that text-only pipelines materially misrepresent both what is discussed and what spreads.

Result 2: LLM's for Multimodal Annotation. We build an annotation pipeline for sensitive categories such as misinformation, hate speech, and caste-coded hostility across modalities. While models perform reasonably on coarse categories like news and political content, they are not reliable substitutes for humans on socially grounded harm labels. We document systematic failure modes and show how performance varies by modality and by category, supported by qualitative case studies.

Result 3: Multimodal Narrative Alignment between Mainstream media and WhatsApp. We construct a mainstream narrative reference set from YouTube videos of prime-time television coverage during the same period as the WhatsApp data and use embedding-based matching to measure narrative overlap between television and WhatsApp. Using this, we show that detectable cross-channel narratives are disproportionately carried by non-textual WhatsApp messages and that matched narratives are far more likely to be viral than baseline WhatsApp content, providing a concrete demonstration of why multimodal tools are essential for studying narrative coupling in practice.

Together, these results provide both substantive findings about political communication on WhatsApp during an election period and a reusable methodological toolkit for the broader computational social science community. By demonstrating how modern generative models can be repurposed as rigorous measurement instruments, we offer a scalable framework for analyzing multimodal, multilingual discourse that extends beyond WhatsApp to any platform where meaning is constructed through the complex interplay of text, image, and video.

2 Related Work

WhatsApp as Political Infrastructure in India. WhatsApp has been widely studied as a key channel for political communication in India, in part because it supports trusted, high-reach diffusion through groups and forwarding under end-to-end encryption. Work on party strategy and online mobilization describes how political actors build organized distribution capacity and use WhatsApp as part of election communication infrastructure (Mahapatra and Plagemann 2019; Shih 2023; Woolley and Howard 2018). Methodologically, progress has been enabled by privacy-respecting data access approaches, especially data donation. Garimella and Chauchard introduce a donation-based tool and protocol that make large-scale WhatsApp research feasible without breaking encryption (Garimella and Chauchard 2024). Using WhatsApp data directly, prior work also shows that non-text content is central to political communication and misinformation. In particular, Garimella and Eckles document the importance of images in political WhatsApp groups and show that a meaningful share of circulated images match known misinformation (Garimella and Eckles 2020). Complementing quantitative work, Banaji et al. study how misinformation and rumor on WhatsApp can connect to offline harm and vigilante violence, emphasizing reception, social context, and the local meaning of messages (Banaji et al. 2019).

Multimodal Analysis and LLMs in Social Media. A large computational literature argues that multimodality is essential for studying modern political content, since meaning in memes, screenshots, and short clips often depends on the interaction between text and visuals. Benchmark work such as Hateful Memes challenge formalizes this point by constructing examples where unimodal systems fail (Kiela et al. 2020). Recent vision-language models and multimodal LLMs make it possible to analyze large mixed-media corpora without task-specific training. Contrastive pretraining (e.g., CLIP) supports retrieval and semantic grouping, and later systems such as BLIP-2 and LLaVA enable captioning and instruction-following analysis with natural language prompts (Radford et al. 2021; Li et al. 2023; Liu et al. 2023). In parallel, work in computational social science has begun to use LLMs for scalable summarization and interpretation of political video content, demonstrating how model-based pipelines can turn video corpora into analyzable text-like objects (Breuer et al. 2025). At the same time, a growing literature cautions that LLM/VLM performance on sensitive categories such as hate speech and misinformation is uneven across languages, cultures, and modalities. Empirical studies document vulnerabilities and prompt sensitivity in hate detection (Roy et al. 2023), and reviews synthesize concerns about bias, evaluation gaps, and domain mismatch (Albladi et al. 2025). Multimodal hate benchmarks that are multilingual and multicultural further emphasize that model behavior can shift substantially across contexts (Bui, Von Der Wense, and Lauscher 2025). These findings motivate the auditing focus in our annotation analysis.

Cross-Platform Narrative Coupling. A third body of work studies how attention and narratives move between mainstream media and online platforms. Classic agenda-setting theory argues that news media shape what issues become salient (McCombs and Shaw 1972). More recent intermedia agenda-setting work examines two-way coupling in hybrid systems, where social media can influence news agendas and legacy outlets can still dominate attention cycles, especially during elections (Harder, Sevenans, and Van Aelst 2017; Guo and Vargo 2020). In India, previous work provides direct evidence of tight coupling between a television debate show and online engagement, consistent with a feedback loop between broadcast agendas and coordinated online ac-

tivity (Jakesch et al. 2021; Garimella and Datta 2024). This line of work motivates our effort to measure narrative overlap between mainstream media and WhatsApp, but most existing approaches rely on text (transcripts, headlines, hashtags). Our setting requires multimodal matching because television narratives often enter WhatsApp as screenshots and short clips rather than as text.

Political organization and the Indian media environment. Finally, our framing connects to research and reporting on organized digital campaigning and the structure of India's media ecosystem. Work on computational propaganda emphasizes that influence operations are often hybrid and organizational, combining institutions, labor, and platform affordances (Woolley and Howard 2018). In India, observers have documented the growth of party-linked digital operations and the normalization of coordinated messaging (Mahapatra and Plagemann 2019; Shih 2023). Parallel reporting describes political and corporate pressure on mainstream media and concerns about declining editorial independence (International Journalists' Network 2024; Reporters Without Borders 2025). These conditions motivate the frequently discussed "central source" idea: that narratives may be pushed through aligned television coverage and WhatsApp dissemination from shared organizational infrastructure. A key barrier to testing such claims is that the most direct traces of coupling are often multimodal artifacts (clips, screenshots, posters). Our work addresses this modal gap by providing tools to link mainstream narratives to WhatsApp circulation even when the carrier is not text.

3 Data Collection

We assemble a multimodal dataset from WhatsApp and mainstream news on YouTube during the 2024 Indian General Elections, which we use to study multimodal topics and virality on WhatsApp, model-based harm annotation, and cross-channel narrative overlap.

3.1 WhatsApp data

Our primary dataset consists of WhatsApp messages collected through a large-scale data donation drive in Uttar Pradesh (UP), India's most populous state with over 220 million people. We sampled 100 locations across UP, covering urban, semi-urban, and rural settings. Locations were selected to approximate the state's demographic composition along key dimensions (age, religion, and caste), benchmarked against census data.

Participants were recruited through in-person site visits by our survey team within each sampled town. Adhering to rigorous ethical protocols approved by our institution's Institutional Review Board (IRB), we obtained informed consent from 3,127 participants who agreed to donate their WhatsApp data. This collection was facilitated by a specialized privacy-preserving tool developed by (Garimella and Chauchard 2024), which is designed to strip sensitive user information at the source. While the comprehensive protocols for data anonymization and ethical user data storage are detailed in our prior work (anonymized for peer review), the present study focuses exclusively on the technical application of LLMs and VLMs for multimodal content analysis. Consequently, we restrict our analysis to the message content itself (text, images, and video) and do not incorporate user metadata or demographic information.

The data collection period spanned five months, from April to September 2024, covering the entirety of the Indian General Election cycle. The raw corpus contained over 5.5 million messages from 4,818 unique WhatsApp groups. These groups covered a wide range of social contexts, including work/business, family and friends, religious organizations, local village councils, and news groups. Notably, groups explicitly labeled as political constituted only approximately 10% of the sample, highlighting the embedded nature of political discourse in non-political spaces.

A key feature of the dataset is the presence of viral content, identified by WhatsApp's "forwarded many times" tag, indicating content that was forwarded at least five hops from the original sender.[1] In the entire dataset, 55,936 messages (approximately 1%) were flagged as viral. To make multimodal analysis computationally tractable while preserving full coverage of viral content, we construct the analytic dataset in two parts: (i) we include *all* messages labeled "forwarded many times," and (ii) we add a uniform random sample of 5% of the remaining (non-viral) messages. This yields a final analytic dataset of 273,913 messages. In these 273k messages, there were 119,589 text messages (43.7%), 96,698 images (35.3%), and 57,626 videos (21.0%). Non-text content therefore accounts for 56.3% of the analytic dataset (images + videos), underscoring the importance of multimodal methods for studying WhatsApp communication in this setting.

3.2 Mainstream News Data (YouTube)

To enable a comparative analysis of cross-channel narrative overlap, we curated a dataset of mainstream television news coverage uploaded to YouTube during the same period: April 2024 to September 2024.

We utilized the YouTube Data API (v3) to retrieve all available video metadata (ID, title, description, tags, view/-like counts, and duration) from the official channels of five prominent news anchors: Aman Chopra, Amish Devgan, Arnab Goswami, Navika Kumar, and Sudhir Chaudhary. These figures host the five most-watched prime-time news shows in Hindi and English, serving as a proxy for dominant broadcast narratives.

To map the thematic landscape of this news coverage, we processed the video titles using the Gemini API. We employed a two-stage clustering approach: first, titles were grouped week-by-week into abstract and specific topics; second, similar topics were manually merged to consolidate videos relevant to distinct events. This process allowed us to trace specific narratives, such as the Lok Sabha elections, political controversies, and major breaking news, from television broadcasts to their potential downstream appearance on WhatsApp. The resulting corpus consists of 1,607 news headlines organized into 98 distinct topics. This structured dataset serves as our ground truth for analyzing the flow of

[1]The exact definition by WhatsApp can be found here https://faq.whatsapp.com/1053543185312573

political narratives from mainstream media into private encrypted networks.

We use this corpus as a mainstream narrative reference set for our cross-channel matching analyses. It should not be interpreted as "ground truth" about influence or exposure; rather, it provides a time-indexed representation of mainstream narrative supply that can be compared to WhatsApp content at the level of topics and narrative similarity.

4 Methods

4.1 Topic detection across modalities

Our primary pipeline for analyzing multimodal data relies on a strategy of semantic unification: we systematically convert non-textual modalities (images and videos) into dense textual descriptions, allowing us to treat the entire dataset as a single text corpus for downstream analysis.

While conceptually straightforward, this approach capitalizes on the significant recent advancements in VLMs, which can now capture granular semantic details from visual and audiovisual content. Examples of these generated descriptions are provided in Section D (Appendix), demonstrating the fidelity of this conversion. This strategy offers a unified framework that standardizes the input space, ensuring that our analytical pipeline remains consistent regardless of the source modality.

We explicitly chose this generative description approach over latent space methods. We evaluated architectures that attempt to embed all modalities into a shared vector space, such as those proposed by (Girdhar et al. 2023; Radford et al. 2021). However, we found that as of late 2025, these joint-embedding methods do not reliably align all three modalities—particularly when handling the linguistic nuances of Hindi text mixed with culturally specific visual markers. In contrast, our text-generation approach proved robust. We validated the quality of the generated descriptions through manual qualitative review, finding the performance satisfactory for capturing both the explicit and implicit meaning of the messages.

To systematically categorize the WhatsApp corpus across modalities, we developed an automated topic modeling pipeline utilizing the `Qwen2.5-7B-Instruct` model. This pipeline operates on the standardized text descriptions generated during the annotation phase (see Section D) rather than processing the raw media files directly, allowing for consistent structured labeling across all modalities.

For each message, the model was tasked with generating two outputs: (1) a primary topic classification drawn from a fixed taxonomy, and (2) a set of 5–10 secondary keywords that provide granular semantic summaries. All inferences were performed in batches with a low sampling temperature (0.001) to maximize deterministic behavior and consistency across the dataset.

We defined a taxonomy of 14 categories specifically tailored to the Indian social media context: *Religion, Politics, Entertainment, Commerce, Poetry, Employment, Violence, Health, Education, Environment, Greetings, Relationships, Sexual Content,* and *Miscellaneous*. This taxonomy was derived empirically through an iterative qualitative analysis of the dataset, designed to encompass both the high-frequency viral content and the everyday communicative themes of rural discourse.

The prompt templates and the strict output schema enforced for this task are provided in the Appendix C.

4.2 Data annotation

To evaluate the efficacy of Large Language Models (LLMs) and Vision-Language Models (VLMs) for automated content analysis, we first established a high-quality ground truth dataset through large-scale manual annotation. We recruited 40 annotators from the same geographic locations as the data collection sites to ensure high cultural and contextual validity. Because parts of the corpus contain graphic violence, hate speech, and sexually explicit material, annotators were briefed on the nature of the content before consenting to participate, given the option to skip any item without penalty. Each piece of content was reviewed by at least three independent annotators, with the final ground truth determined by majority vote. All annotators underwent rigorous training, and their work was regularly audited by the authors to maintain quality.

Inter-annotator agreement, measured using Krippendorff's alpha, ranged between 0.3 and 0.5 across most categories. These values are consistent with prior work on subjective annotation tasks such as hate speech and political content (Del Vigna12 et al. 2017; Ousidhoum et al. 2019).

We annotated each item across multiple distinct dimensions critical to understanding the circulation of problematic content in the Indian context. Key dimensions included:

1. **False or misleading claims:** We adopted a broad definition of "misleading" rather than the narrower category of "misinformation." This framing allows us to capture subtler forms of manipulation, such as cheapfakes or out-of-context media, which might strictly be factually real but deceptively framed.
2. **Hateful content:** While we annotated for all forms of hate speech, we placed specific emphasis on identifying anti-Muslim narratives (e.g., "Love Jihad," "Population Jihad," and "Forced Conversions"), as these constitute the majority of identity-based hostility.
3. **Topical categorizations:** We further classified content into specific domains, including (3) casteist content; (4) health-related advice; (5) news-related information; (6) partisan political content; and (7) religious content and iconography.

To ensure valid interpretation, annotators utilized a specialized interface that reconstructed the full conversational context. Rather than viewing content in isolation, annotators were presented with the target item embedded within its original WhatsApp group timeline. This allowed them to scroll through the surrounding chat history to examine the immediate context and conversational flow before assigning labels.

From the total corpus of 273k items, a stratified subset of 16,379 messages was manually annotated and retained for the evaluation described in Section 5.2. The distribution

of this ground truth dataset across modalities and virality is summarized in Table 1.

To scale this analysis to the full dataset, we developed an automated pipeline designed to mimic the manual annotation process. We engineered prompt templates containing precise definitions, few-shot examples, and multi-step reasoning instructions to guide model outputs. Importantly, we do not train or fine-tune any model on the annotated data; all automated labels are produced by zero-shot or few-shot prompted inference. The manually annotated set is used exclusively for evaluation: we compare the model-generated labels against the human majority-vote labels on the held-out annotated subset to compute precision, recall, and F1 per category.

We annotated the complete set of viral WhatsApp content and a random 5% subsample of the broader dataset using a suite of state-of-the-art models tailored to each modality. Text-based messages were analyzed using `Qwen2.5-7B-Instruct`, and images were processed using Google's `Gemma3-27B` multimodal model. For video content, we employed a tiered strategy: viral videos were analyzed using `Gemini 2.5 Pro`, while non-viral videos were processed using `Qwen2.5-VL-32B-Instruct`.[2]

To ensure structured and reproducible outputs, all model generations were constrained to predefined JSON schemas using the `lmformatenforcer` library. This enforcement allowed us to simultaneously extract multiple analytical dimensions—both binary and categorical—in a consistent format. The complete prompt templates and schema specifications are provided in Appendix C.[3]

Automated annotation was performed on a cluster of A100 GPUs using `bfloat16` precision. We utilized a low temperature setting to minimize variance in the model responses. For efficiency, video content processed by Qwen2.5-VL was sampled at 1 frame per second and downscaled to a maximum resolution of 360×420 pixels.

Modality	Non-viral	Viral	Total
Image	4,735	1,212	5,947
Text	7,053	601	7,654
Video	788	1,990	2,778
Total	12,576	3,803	16,379

Table 1: Distribution of manually annotated content by modality and virality.

[2]At the time of analysis (October 2025), Gemini models were the only commercial options capable of natively processing video with audio. However, due to prohibitive costs at scale, we restricted Gemini usage to viral videos. For the remaining video content, we utilized Qwen2.5-VL. This model employs a native dynamic-resolution Vision Transformer (ViT) with Window Attention to reduce computational overhead while preserving resolution. It introduces Multimodal Rotary Position Embedding (MRoPE) aligned to absolute time, allowing it to capture both spatial scales and temporal dynamics. This architecture achieves performance comparable to GPT-4o and Claude 3.5 Sonnet on multimodal benchmarks, despite processing video as a sequence of frames.

[3]Code, prompts, and additional materials are available at https://harsha20032020.github.io/whatsapp.

4.3 Matching YouTube and WhatsApp Content

Understanding how information diffuses from public broadcast media into private messaging networks remains a critical challenge in analyzing contemporary information ecosystems. To investigate this phenomenon, we developed a computational framework to link source narratives (YouTube news coverage) with downstream private discourse (WhatsApp messages), enabling us to systematically measure how specific mainstream topics permeate closed communities. Since raw audio and visual signals cannot be directly compared across platforms without significant noise, we adopted a fully semantic retrieval framework based on text representations. For WhatsApp content, we constructed a comprehensive textual representation by concatenating the raw message text (where available) with the dense LLM-generated descriptions derived in the previous step. This ensures that visual and multimodal messages are searchable within the same semantic space as text-only messages. We encoded both the WhatsApp representations and the YouTube news headlines into dense vector embeddings using LaBSE (Language-agnostic BERT Sentence Embedding) (Feng et al. 2022). LaBSE was selected specifically for its robustness in the Indian context, as it effectively captures semantic similarity across the linguistic code-mixing—spanning Hindi, English, and regional dialects—that is characteristic of our dataset.

To ensure contextual relevance, we applied a rigorous two-stage filtering process. First, we restricted the search space for each topic to WhatsApp messages sent within a window of ±7 days relative to the timestamp of the YouTube coverage. Within this temporal window, we performed semantic retrieval at the headline level, identifying the top-k most similar WhatsApp messages for every headline associated with a news topic based on cosine similarity. To mitigate spurious matches, where a message might accidentally resemble a single sensational headline without being relevant to the broader topic, we enforced a multi-headline consensus criterion. A message was only considered relevant if it exhibited high similarity scores with multiple distinct headlines belonging to the same topic. We aggregated the similarity scores across these matched headlines to produce a final confidence-ranked set of messages.

Finally, to reduce noise and focus on salient narratives, we filtered out any topics that yielded fewer than 10 relevant messages. This rigorous filtering reduced our initial set of 98 topics to a high-confidence subset of 45 topics, associated with approximately 2,884 unique WhatsApp messages. We performed manual verification on the retrieved set to assess precision, and qualitative analysis confirmed that the matches were highly relevant, successfully bridging the vocabulary gap between formal news discourse and informal messaging language. We further ablated the pipeline by varying retrieval thresholds, headline matching requirements, and keyword constraints; these experiments showed that our core findings were robust to hyperparameter variations.

A summary of our methods and datasets is shown in Figure 1.

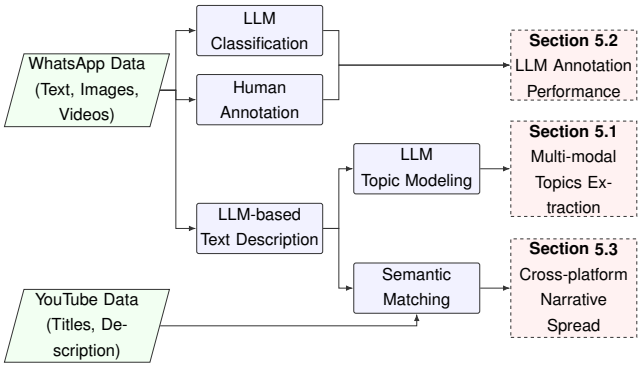


Figure 1: Overview of the methodology pipeline. WhatsApp assets are processed for classification (Sec 5.2) and description generation. Descriptions feed into topic modeling (Sec 5.1) and semantic matching with YouTube data (Sec 5.3).

5 Results

5.1 Topics Across Modalities

To characterize the information ecosystem of WhatsApp, we analyzed the corpus using our hybrid language–vision classification pipeline. This process assigned a discrete *primary topic* from a fourteen-category taxonomy and a set of descriptive *secondary keywords* to every message, regardless of its modality. Figure 2 presents the normalized frequency distribution of these primary topics, offering a macroscopic view of the dominant discourse.

The distribution reveals that WhatsApp serves a dual function in this demographic: it is simultaneously a marketplace for economic exchange and a forum for civic discourse. As shown in Figure 2, the categories of *Commerce* and *Politics* jointly account for approximately one-third of the total volume. The prevalence of *Commerce* reflects the platform's utility for local businesses and informal trade, while the high volume of *Politics* confirms its central role in the public sphere (even though the groups we sampled were not majorily explicitly political).

Notably, we observe a substantial prevalence of content related to *Violence* (3.2%). This category frequently includes graphic footage of accidents, local skirmishes, and raw news footage, suggesting that WhatsApp acts as a conduit for visceral, unfiltered content that may be absent from sanitized mainstream media streams. See Table 3 (Appendix) for detailed definitions and examples of these categories.

The remaining communicative bandwidth is distributed across sociocultural dimensions including *Religion*, *Greetings*, *Relationships*, and *Health*. Given that this analysis is derived from a randomized subsample of data donated by a demographically representative cohort, these patterns likely offer a robust approximation of general WhatsApp usage in Uttar Pradesh.

Next, to validate the granularity of our automated pipeline, we analyzed the distribution of secondary keywords (see Figure 10 in Appendix). While high-frequency tokens often reflect generic descriptors, the long tail of this distribution allows for the rapid identification of niche but socially critical themes. For instance, we successfully identified distinct clusters of content related to "mental health" (1.5%) and "calls to action" (around 2%), demonstrating that VLM-driven annotation can surface specific qualitative signals that broader taxonomies might obscure.

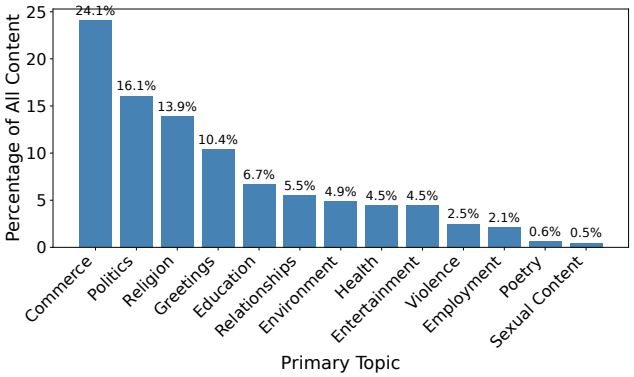


Figure 2: Primary topics as a percentage of all content. For a description of the topics, see Table 3 (Appendix). Not shown in the figure: Miscellaneous topic containing around 17% of posts.

A critical advantage of our multimodal pipeline is the ability to disaggregate topics by their semiotic composition. To quantify this, we computed a topic-modality contingency matrix, analyzing both the modality mix *within* each topic (Figure 3) and the contribution of each topic-modality pair to the overall corpus (Figure 4).

The within-topic distribution (Figure 3) uncovers striking heterogeneity in how different subjects are communicated. *Commerce*, for example, is overwhelmingly visual, with images constituting approximately 60% of the category. This aligns with the platform's use for cataloging products and sharing flyers. Conversely, utilitarian categories like *Greetings* and *Poetry* remain dominated by text.

Most significantly, *Politics* emerges as a heavily multimodal domain. Roughly 65% of political content in our sample is non-textual (images or video). This finding highlights a critical methodological blind spot in computational social science: traditional text-scraping approaches would fail to capture nearly two-thirds of the political discourse in this ecosystem.

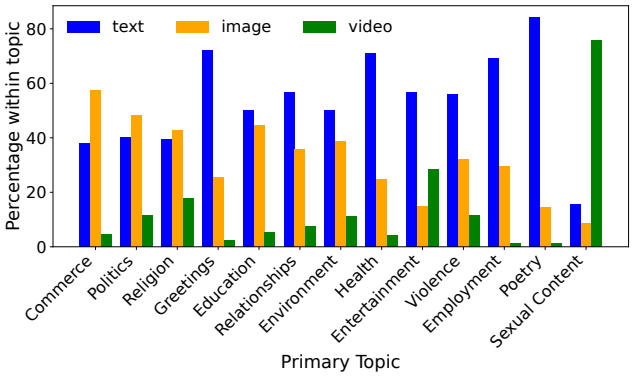


Figure 3: Modality percentage distribution within each primary topic.

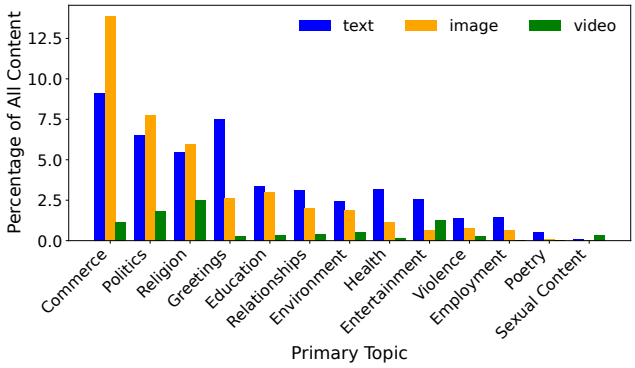


Figure 4: Modality share of total dataset per primary topic.

Finally, we examined the relationship between thematic content and diffusion mechanics. We constructed contingency matrices to analyze the "virality rate", i.e., the proportion of messages marked as *forwarded many times* within each topic.

Figure 5 illustrates a clear divergence between "high-arousal" and "utilitarian" content. Topics related to *Politics* and *Violence* exhibit markedly higher viral rates compared to the baseline. This suggests that content designed to evoke outrage, fear, or partisan identity is structurally advantaged in the WhatsApp ecosystem. In contrast, utilitarian domains such as *Education*, *Greetings*, and *Employment* remain predominantly non-viral, circulating largely within local, functional contexts rather than cascading across groups.

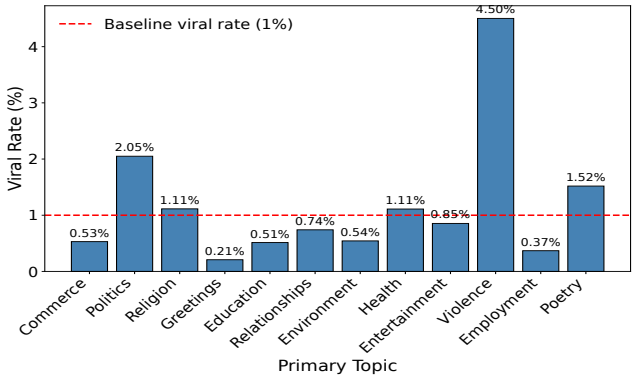


Figure 5: Virality composition (% viral vs. non-viral) for the primary topics.

An apparent outlier is *Sexual Content*, which displays a virality rate of approximately 22%. While counter-intuitive for a private messaging platform, a deeper forensic analysis revealed this was driven by a small cluster of hyper-active groups dedicated to the mass dissemination of adult content.[4]

The interaction between modality and virality is further detailed in Table 2. The data indicates that specific modality-topic combinations possess disproportionate viral potential, most notably political and sexual videos, and both videos and images in the context of violence.

Primary Topic	Text	Image	Video
Commerce	0.16	0.71	1.60
Politics	1.10	1.43	10.32
Religion	0.33	0.86	2.81
Greetings	0.00	0.67	1.32
Education	0.33	0.45	2.55
Relationships	0.16	1.08	3.21
Environment	0.19	0.56	2.01
Health	0.75	1.19	8.87
Entertainment	0.03	0.61	2.15
Violence	0.75	5.67	24.32
Employment	0.30	0.57	0.00
Sexual Content	0.85	2.82	29.04
Poetry	1.08	1.82	60.00

Table 2: Viral Rate (%) by Topic and Modality.

5.2 Do LLMs Work for Multimodal Content Annotation

Our topic analysis shows that LLM/VLM representations are strong enough to organize WhatsApp content into broad, high-level themes. A natural next question is whether the same models can support *fine-grained* social annotation at scale, especially for labels that matter for governance and harm mitigation (e.g., misinformation, hate, and caste-based hostility). These labels are harder than topic tagging for two reasons: they are rare (so the training signal is weak), and they often depend on implicit meaning, local context, and contested norms rather than explicit keywords.

Figure 6 summarizes performance across nine binary tasks.[5] Even under our best settings, performance is mixed and generally not strong enough for real-world deployment. The baseline classifiers struggle most on the rarest and most socially loaded categories. The imbalance ratios range from 2.2:1 for Political to 55:1 for Caste, with corresponding baseline F1 scores that are low for the rare categories (e.g., Caste: 0.157; Health Advice: 0.271). In contrast, categories that are both more common and more semantically explicit are substantially easier: Political achieves baseline F1 of 0.667.

To isolate how much of the failure is driven by prevalence rather than semantics, we also train imbalance-mitigated classifiers using SMOTE oversampling (Chawla et al. 2002). The gains are large for several labels, especially those where positives likely have consistent surface patterns that models can latch onto. For example, F1 improves significantly for most categories. At the same time, the ceiling remains clear: even after balancing, Caste improves only to 0.330, and several harm-oriented tasks remain far from the reliability required for automated moderation or high-stakes measurement. This pattern suggests that imbalance is a major part of the problem, but not the full story. Some labels remain difficult because the positive class is defined by impli-

[4]Specifically, we identified a set of 6 groups responsible for a high volume of sexual video content, the majority of which was forwarded. This results in a disproportionately skewed virality ratio for this category relative to the general user population.

[5]We report F1 because class imbalance is extreme and accuracy is not informative.

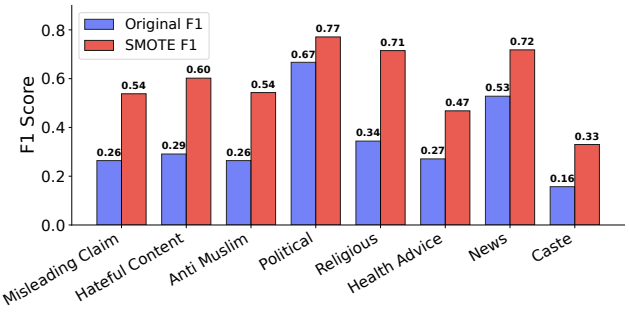


Figure 6: F1 performance of automated classifiers for nine content labels. Blue bars show baseline training under the natural class distribution; red bars show performance after training with SMOTE oversampling to mitigate class imbalance (Chawla et al. 2002).

cation and social meaning (e.g., caste-coded hostility), not by stable and explicit cues.

A useful contrast is explicit content related topics like Political, Partisan, News, etc. The implication is not that LLMs and VLMs "cannot" detect harmful content, but that they are currently much more dependable for coarse semantic identification ("this is political/news") than for socially grounded judgments about harm that require context and cultural interpretation ("this is hate/misinformation/casteist").

We also observe substantial heterogeneity by modality and by virality (Appendix Figure 12). Two points are worth highlighting. First, image and video performance is not uniformly worse than text, which is notable given the cultural specificity of visual cues. In several tasks, visual modalities yield competitive or better F1. Second, viral content is consistently easier to detect than non-viral content across many labels, which suggests that highly forwarded items tend to be more explicitly framed and more tightly aligned with recognizable category cues. This is an important measurement concern: automated systems can appear to work better when evaluated on viral content, yet still fail on the broader and subtler background conversation.

Finally, we tested whether prompt optimization can overcome these limits. We applied DSPy's MIPROv2 prompt optimizer (Khattab et al. 2024) (prompt-only, no few-shot examples) using roughly 5,000 labeled examples per task and `Qwen2.5-7B-Instruct` as both the task and optimization model. Across tasks (including political, caste-based, misinformation, and hateful content detection), we found no meaningful improvements and in some cases small degradations. This negative result suggests that the bottleneck here is not prompt wording. It is the combination of (i) severe class imbalance, (ii) limited and noisy cues for many harm categories, and (iii) genuine ambiguity and disagreement in the underlying labels.

5.3 Cross-Channel Narrative Overlap: Mainstream Media and WhatsApp

Our final analysis examines cross-channel narrative overlap in India's hybrid media system. By leveraging our semantic retrieval pipeline, we measure the extent to which narratives originating in mainstream television news (via YouTube) permeate the WhatsApp ecosystem. This analysis moves beyond simple keyword matching to capture cross-modal narrative diffusion—identifying instances where a TV news segment appears on WhatsApp not just as text, but as a video clip, a screenshot, or a meme. The goal here is not to prove that TV narrative *causes* WhatsApp content or the other way around. Instead, we show that this kind of multimodal narrative overlap can be measured at scale using modern LLM and VLM representations. We focus on two questions: (i) which mainstream television narratives show up in WhatsApp during the 2024 election period, and (ii) which modalities (text, images, videos) carry these narratives once they are on WhatsApp?

Starting from the 98 YouTube topic clusters constructed in Section 4.3, we discard topics that have too few matched WhatsApp messages to support stable estimates. Concretely, we keep topics with at least 10 matched WhatsApp messages, yielding 45 topics. All results below use the 2,884 WhatsApp messages that match these 45 topics.[6]

Even this conservative matched set shows that narratives that are salient enough in mainstream television coverage to appear repeatedly and recognizably in WhatsApp, despite translation, remix, and modality shifts. This is precisely the kind of cross-channel signal that text-only methods tend to miss when the carrier is a screenshot or a clip rather than a transcript.

Table 4 (Appendix) shows that 19 of the 45 topics (42.2%) are labeled pro-government, 6 (13.3%) anti-government, and 20 (44.4%) neutral/news. Importantly, the distribution of *messages* is more skewed than the distribution of *topics*: pro-government topics account for 58.7% of the matched WhatsApp messages. This gap matters because it suggests that mainstream narratives do not enter WhatsApp uniformly. Rather, some narrative types appear repeatedly across groups and are more likely to be repackaged into WhatsApp-friendly formats.

We emphasize two interpretations consistent with these data. First, the YouTube corpus reflects a prime-time ecosystem where government-aligned narratives are prominent; the matched WhatsApp distribution then indicates that these narratives are also present in private messaging conversations. Second, because our matched set is conditioned on what appears on television, this analysis is best read as evidence of *coupling* rather than one-directional influence. In the remainder of the section we show that the coupling is not merely textual, and that the narratives most clearly coupled across channels are disproportionately viral inside WhatsApp.

Figure 7 plots the ten topics with the largest number of matched WhatsApp messages. The top 4 topics by volume and six of these ten topics are pro-government, focusing on the promotion of the Prime Minister or government initiatives. This concentration suggests that a relatively small set of television narratives account for a large share of cross-

[6]The full topic list, stance labels (pro-government, anti-government, neutral/news), and per-topic viral rates are reported in Appendix Table 4.

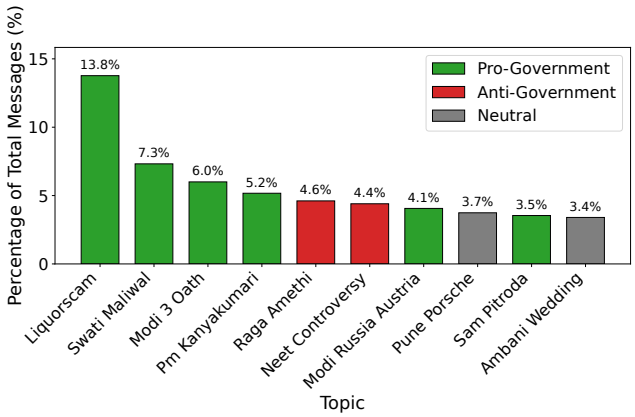


Figure 7: Volume of WhatsApp messages matching the top 10 mainstream news narratives, ranked by total message count.

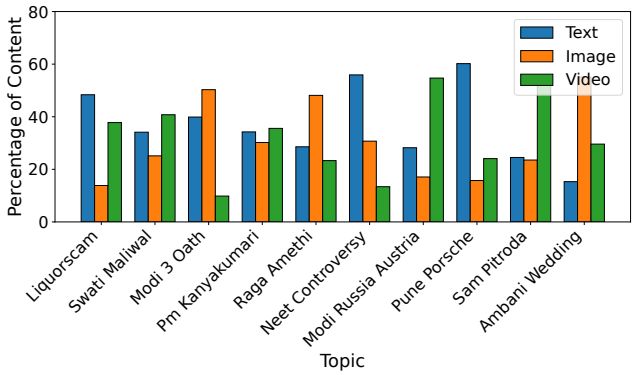


Figure 8: Modality distribution of matched WhatsApp messages for the top 10 topics. Each bar sums to 100% and shows the share of matched messages that are text, image, or video.

channel overlap, consistent with the idea that WhatsApp circulation often tracks a limited number of high-salience frames during election periods.

A key contribution of our pipeline is that it can match and analyze narratives even when they cross modalities. Figure 8 shows the modality mix of matched WhatsApp messages for the top ten topics. Two patterns stand out. First, every high-volume narrative appears in multiple modalities, indicating that narrative transport across channels is not purely textual. Second, the *dominant* carrier differs by narrative: some topics are primarily transported through videos (consistent with television segments being forwarded as short excerpts), while others appear mainly as images (consistent with screenshot and meme formats). This heterogeneity is important because it explains why text-only matching substantially underestimates cross-channel coupling: the message that best preserves the original television framing is often a screenshot or a clip.

Aggregating over the full matched set of 45 topics, only 38.3% of matched messages are text, while 61.7% are non-text (32.3% videos and 29.4% images). In other words, a text-only pipeline would miss most of the measurable overlap between mainstream television narratives and WhatsApp circulation in our data.

We next ask whether cross-channel narratives are merely present on WhatsApp or whether they are disproportionately represented among messages that spread widely. The answer is clear. Across the 45 retained topics, the average fraction of messages labeled viral is 28.35% (median 26.35%), far above the baseline viral prevalence in the overall dataset (approximately 1% as seen in Figure 5). For the top ten topics, the viral fraction is even higher (36.71%). This gap is substantively meaningful even under conservative interpretation: narratives that are salient enough to be strongly represented on television and identifiable on WhatsApp are more likely to be the kind of content that gets forwarded at scale.

Moreover, virality differs by stance. Pro-government topics exhibit higher viral rates on average (mean 41.44%) than neutral/news topics (mean 19.70%) and anti-government topics (mean 30.54%). While the number of anti-government topics is small and we do not interpret

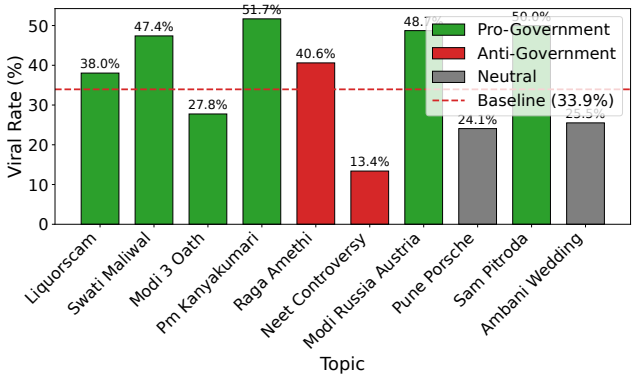


Figure 9: Viral versus non-viral composition of matched WhatsApp messages for the top 10 topics. Bars show counts, partitioned by WhatsApp's "forwarded many times" label.

these differences causally, the pattern is consistent with existing qualitative accounts of organized partisan amplification on WhatsApp and suggests that cross-channel coupling and WhatsApp virality may interact in politically asymmetric ways during elections.

In summary, these results show that cross-channel narrative coupling is both measurable and strongly multimodal. Most of the identifiable overlap between mainstream television narratives and WhatsApp circulation in our data would be missed by text-only analysis, and the narratives that do overlap are disproportionately represented among viral WhatsApp messages. This provides a concrete demonstration of why multimodal LLM/VLM tooling is not only useful for describing WhatsApp content, but also necessary for linking WhatsApp to the broader media system in which political narratives are produced and amplified.

6 Discussion

This paper set out to make progress on a practical measurement problem: how to study political communication on a social media platform when the content is multilingual and, critically, dominated by images and videos. Across three analyses, we show that treating WhatsApp as a multimodal

medium changes what becomes visible, and that modern LLM/VLM tooling makes parts of this analysis feasible at a scale that would otherwise be out of reach. At the same time, we show that the same models that are powerful for exploration and coarse categorization are not reliable substitutes for humans when the task requires culturally grounded judgments about harm (misinformation, hate speech, and especially caste-coded hostility). Finally, we demonstrate that these multimodal representations make it possible to connect private-message circulation to mainstream television narratives, including when the carrier is a screenshot or a clip rather than text.

The Necessity of Multimodal Measurement. Our results fundamentally challenge the sufficiency of text-based approaches in computational social science, particularly in the Global South (Garimella and Eckles 2020; Varanasi, Pal, and Vashistha 2022). We found that nearly 65% of political content in our dataset was non-textual. This implies that prior studies relying on scraping text or analyzing public groups may have systematically under-represented the political volume and misunderstood the nature of political mobilization.

The heavy reliance on visual modalities, such as images for memes and posters, videos for news clips and speeches, aligns with the "visual turn" in political communication (Veneti, Jackson, and Lilleker 2019). In contexts of variable literacy and high linguistic diversity, visual artifacts serve as a universal vernacular, capable of bypassing literacy barriers that restrict textual debates. The granularity provided by our VLM pipeline allows us to move beyond simply counting images to understanding their semantic role: detecting that Commerce is visual because it is a catalog, while Politics is visual because it is emotive and symbolic. This level of descriptive power, achieved without manual coding, opens new avenues for studying the aesthetics of digital campaigning at scale.

AI vs. Local Context. While our pipeline demonstrated high efficacy in categorizing general high level topics (e.g., Commerce, Politics), the same models are substantially less reliable on categories that depend on social interpretation, local context, and contested definitions, including misinformation and hate speech. This is consistent with a broader literature showing that LLMs can appear impressive on benchmarked hate detection while remaining sensitive to prompts, domain shift, and implicit context (Roy et al. 2023; Albladi et al. 2025). Our qualitative error analysis shows that the failures are clustered around specific difficulties that are central in the Indian context: coded language, humor, implicit targets, and cues that require cultural knowledge rather than surface semantics. This is also where multimodality adds difficulty: the harmful meaning may be carried by composition, insinuation, or a visual trope rather than an explicit slur.

An important nuance that emerged during our annotation process is that disagreement is not only a model problem. It is also a social problem. Our annotators were local to the region of collection, and we observed that behaviors we (as authors) viewed as clearly hateful or clearly misinformation were sometimes treated as ordinary, routine, or politically unsurprising by annotators. This does not imply that there is no ground truth, but it highlights that labels like "hate speech" and "misinformation" are not purely technical categories. They are socially negotiated and can depend on what communities have normalized, what they view as credible, and what they perceive as harmful. Even with structured guidelines, these tasks remain difficult to define in a way that is both locally meaningful and comparable across settings. The implication is not that automated labeling is futile. Rather, it strengthens the argument that model-based labeling should be treated as measurement with error: useful for aggregate patterns and hypothesis generation, but requiring auditing, sensitivity checks, and, for high-stakes claims, human validation. In other words, LLMs and VLMs can expand the frontier of feasible analysis, but they must be embedded in an evaluation culture, not used as black boxes.

The Porous Boundary Between TV and WhatsApp. Our analysis of the link between YouTube news and WhatsApp confirms the existence of a "hybrid media system" where broadcast and social channels are tightly coupled (Garimella and Datta 2024). By successfully matching multimodal artifacts—tracing a TV news segment to a WhatsApp video clip—we provide empirical evidence for the "central source" hypothesis often discussed in Indian media studies (Mahapatra and Plagemann 2019). The finding that pro-government narratives are not only more prevalent but also more viral suggests a coordinated synchronization between ruling party media strategy and private social distribution.

While our sample size of matched topics is small, the methodological contribution is valuable. We demonstrate that it is technically feasible to audit and measure the gap between public broadcast and private social media platforms, going beyond keyword or video matching. This capability is crucial for tracking how narratives are seeded by elites and then amplified through grassroots forwarding chains. It moves the conversation from anecdotal evidence of "IT Cell" operations to measurable, traceable data trails (Jakesch et al. 2021).

Limitations. This paper also has limits that matter for interpretation. The WhatsApp dataset is collected through data donation and covers a specific region and period (UP, April–September 2024), so it is not a national census and may reflect selection effects; in particular, voluntary data donation may over-represent users who are more digitally literate or more willing to share, and the resulting sample may not capture the full diversity of WhatsApp usage patterns. Our viral signal is operationalized using WhatsApp's "forwarded many times" tag, which captures high forwarding but not full diffusion paths. Our model-based labels and matches can encode bias and can fail for culture-specific content; although we audit failures qualitatively, we do not claim that our estimates are ground truth. For cross-channel analysis, YouTube uploads are not identical to live broadcast and may be strategically edited, and our topic extraction from titles and clustering introduces additional measurement error. We also note that because our entire analytical pipeline depends on VLM-generated descriptions, any systematic bias in how these models interpret low-resource languages or culturally

specific visual content could propagate through topic modeling and cross-platform matching results.

Finally, the rapid evolution of generative models presents a moving target. The models used in this study (late 2024/early 2025) may soon be surpassed. However, the fundamental framework we propose for converting multimodal data into semantic text descriptions for unified analysis remains a robust and adaptable strategy.

In sum, our contribution is to show that multimodal LLM/VLM tooling can make large-scale WhatsApp analysis possible without pretending that the hardest interpretive tasks have been solved. Multimodality changes the empirical picture; model-based annotation expands what can be measured but requires humility and auditing; and multimodal narrative matching opens a new way to study how political communication moves across mainstream media and encrypted social networks.

Generative AI usage. The authors acknowledge the use of generative AI tools like Gemini for writing assistance in parts of the paper. All AI generated text was double checked by the authors before being included in the paper.

References

Albladi, A.; Islam, M.; Das, A.; Bigonah, M.; Zhang, Z.; Jamshidi, F.; Rahgouy, M.; Raychawdhary, N.; Marghitu, D.; and Seals, C. 2025. Hate speech detection using large language models: A comprehensive review. *IEEE Access*.

Banaji, S.; Bhat, R.; Agarwal, A.; Passanha, N.; and Sadhana Pravin, M. 2019. WhatsApp vigilantes: An exploration of citizen reception and circulation of WhatsApp misinformation linked to mob violence in India.

Breuer, A.; Dietrich, B. J.; Crespin, M. H.; Butler, M.; Pryse, J.; and Imai, K. 2025. Using AI to Summarize US Presidential Campaign TV Advertisement Videos, 1952–2012. *Scientific data*, 12(1): 1552.

Bui, M. D.; Von Der Wense, K.; and Lauscher, A. 2025. Multi3Hate: Multimodal, multilingual, and multicultural hate speech detection with vision–language models. In *Proceedings of the 2025 Association for Computational Linguistics*, 9714–9731.

Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; and Kegelmeyer, W. P. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16.

Del Vigna12, F.; Cimino23, A.; Dell'Orletta, F.; Petrocchi, M.; and Tesconi, M. 2017. Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the first Italian conference on cybersecurity*, 86–95.

Feng, F.; Yang, Y.; Cer, D.; Arivazhagan, N.; and Wang, W. 2022. Language-agnostic BERT Sentence Embedding. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *ACL*. Dublin, Ireland: Association for Computational Linguistics.

Garimella, K.; and Chauchard, S. 2024. WhatsApp explorer: A data donation tool to facilitate research on WhatsApp. *Mobile Media & Communication*, 20501579251326809.

Garimella, K.; and Datta, A. 2024. Unraveling the Dynamics of Television Debates and Social Media Engagement: Insights from an Indian News Show. In *ICWSM*, volume 18.

Garimella, K.; and Eckles, D. 2020. Images and misinformation in political groups: Evidence from WhatsApp in India. *Harvard Kennedy School Misinformation Review*.

Girdhar, R.; El-Nouby, A.; Liu, Z.; Singh, M.; Alwala, K. V.; Joulin, A.; and Misra, I. 2023. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Guo, L.; and Vargo, C. 2020. "Fake news" and emerging online media ecosystem: An integrated intermedia agenda-setting analysis of the 2016 US presidential election. *Communication Research*, 47(2): 178–200.

Harder, R. A.; Sevenans, J.; and Van Aelst, P. 2017. Intermedia agenda setting in the social media age: How traditional players dominate the news agenda in election times. *The international journal of press/politics*, 22(3): 275–293.

International Journalists' Network. 2024. Corporate and political influence undermines media's editorial independence in India. *IJNet*. Accessed 2026-01-13.

Jakesch, M.; Garimella, K.; Eckles, D.; and Naaman, M. 2021. Trend alert: A cross-platform organization manipulated Twitter trends in the Indian general election. *CSCW*.

Khattab, O.; Singhvi, A.; Maheshwari, P.; Zhang, Z.; Santhanam, K.; Haq, S.; Sharma, A.; Joshi, T. T.; Moazam, H.; Miller, H.; et al. 2024. Dspy: Compiling declarative language model calls into state-of-the-art pipelines. In *ICLR*.

Kiela, D.; Firooz, H.; Mohan, A.; Goswami, V.; Singh, A.; Ringshia, P.; and Testuggine, D. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*. PMLR.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

Mahapatra, S.; and Plagemann, J. 2019. Polarisation and politicisation: The social media strategies of Indian political parties.

McCombs, M. E.; and Shaw, D. L. 1972. The Agenda-Setting Function of Mass Media. *The Public Opinion Quarterly*, 36(2): 176–187.

Narayanan, V.; Kollanyi, B.; Hajela, R.; Barthwal, A.; Marchal, N.; and Howard, P. N. 2019. News and information over Facebook and WhatsApp during the Indian election campaign. *Data Memo*, 2.

Ousidhoum, N.; Lin, Z.; Zhang, H.; Song, Y.; and Yeung, D.-Y. 2019. Multilingual and Multi-Aspect Hate Speech Analysis. In *Proceedings of the 2019 EMNLP*, 4675–4684.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Reporters Without Borders. 2025. India: RSF calls for press freedom in the world's largest democracy.

Roy, S.; Harshvardhan, A.; Mukherjee, A.; and Saha, P. 2023. Probing LLMs for hate speech detection: strengths and vulnerabilities. In *Findings of the association for computational linguistics: EMNLP 2023*, 6116–6128.

Shih, G. 2023. Inside the vast digital campaign by Hindu nationalists to inflame India. *The Washington Post*. Accessed: 2025-10-01.

Trauthig, I. K.; and Woolley, S. C. 2025. "On WhatsApp I say what I want": Messaging apps, diaspora communities, and networked counterpublics in the United States. *New Media & Society*, 27(4): 2050–2067.

Varanasi, R. A.; Pal, J.; and Vashistha, A. 2022. Accost, accede, or amplify: attitudes towards COVID-19 misinformation on WhatsApp in India. In *CHI*.

Veneti, A.; Jackson, D.; and Lilleker, D. G. 2019. *Visual political communication*. Springer.

Woolley, S. C.; and Howard, P. N. 2018. *Computational propaganda: Political parties, politicians, and political manipulation on social media*. Oxford University Press.

Ethics checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes
   (e) Did you describe the limitations of your work? Yes
   (f) Did you discuss any potential negative societal impacts of your work? Yes
   (g) Did you discuss any potential misuse of your work? No
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? NA
   (b) Have you provided justifications for all theoretical results? NA
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA
   (e) Did you address potential biases or limitations in your theoretical framework? NA
   (f) Have you related your theoretical results to the existing literature in social science? Yes
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? Yes
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? No
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? Yes
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? NA
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? No
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? Yes
   (b) Did you mention the license of the assets? Yes
   (c) Did you include any new assets in the supplemental material or as a URL? No
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see **?**)? Yes
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see **?**)? NA
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? No

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? Yes

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? Yes

(d) Did you discuss how data is stored, shared, and deidentified? Yes

A Secondary keywords

Figure 10 shows the top 20 secondary keywords in our dataset.

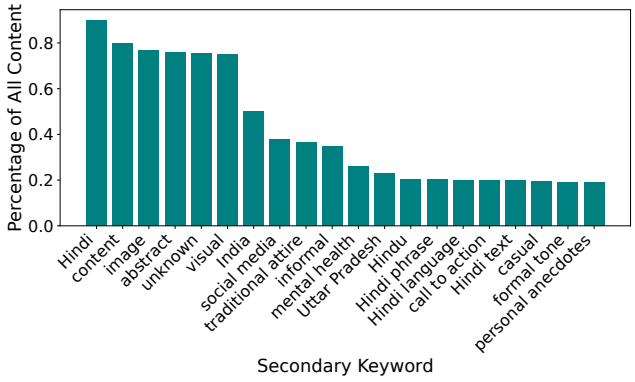


Figure 10: Top 20 secondary keywords as a percentage of all extracted keywords.

B Group coverage

To assess how content production is distributed across groups, we measured the number of distinct groups required to account for 90% of all messages within each major topic category. For every primary topic, groups were ranked by their message volume, and the cumulative distribution was computed to determine the minimal set achieving 90% coverage. The resulting counts, shown in Figure 11, provide a compact measure of topical concentration: smaller values indicate that discussion is dominated by a few highly active groups, whereas larger values imply a more diffuse participation pattern. As visible in the plot, topics such as *Commerce*, *Religion*, and *Politics* exhibit high group counts, reflecting broad and distributed engagement, while categories like *Employment* and *Poetry* reach 90% coverage with far fewer groups, suggesting more localized or niche conversation clusters.

C Experiment Definitions

C.1 Caste Detection

Prompt:

```
Task: Determine whether this
[Text/Image/Video] contains casteist
elements.
```

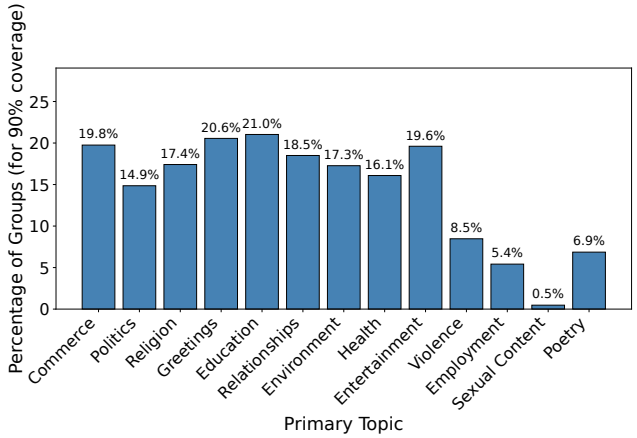


Figure 11: Number of distinct groups required to cover 90% of all messages for each primary topic. Lower values indicate higher topical concentration within fewer groups.

```
  Definition: Content is casteist if it
makes any reference to caste|explicitly
or implicitly|including stereotypes,
discrimination, or social hierarchies based
on caste.
  Instructions:
1. Analyze the content thoroughly
2. Does this contain casteist content?
(Yes/No)
```

Output Schema:

```
{
"type": "object",
"properties": {
"caste": {"type": "boolean"}
},
"required": ["caste"]
}
```

C.2 Claim Detection

Prompt:

```
Task: Determine whether this
[Text/Image/Video] contains a false or
misleading claim.
  Definition: Content is misleading if
it misleads the audience. This includes:
misrepresenting context, omitting important
details, manipulative framing/editing, false
information as fact.
  Instructions:
1. Analyze all elements
2. Does it contain false/misleading claims?
(Yes/No)
```

Output Schema:

```
{
"type": "object",
"properties": {
"is_claim_misleading": {"type": "boolean"}
},
"required": ["is_claim_misleading"]
}
```

C.3 Hateful Content Detection

Prompt:

```
Task: Determine whether this
[Text/Image/Video] contains hateful
elements.
  Definition: Content that promotes violence
or incites hatred against individuals/groups
based on: race, ethnic origin, religion,
disability, gender, age, veteran status, or
sexual orientation/gender identity.
  Hateful content includes: dehumanizing
speech or imagery, allegations of
serious immorality/criminality, slurs
or harmful stereotypes, serious
insults/contempt/disgust, calls for
exclusion or segregation.
  Note: Criticism of political
parties/leaders (BJP/Congress/Modi/Rahul
Gandhi) or organizations (RSS) is not
hateful.
  Instructions:
1. Analyze completely
2. Is this hateful content? (Yes/No)
3. If Yes: Is it anti-Muslim? Does it align
with
common anti-Muslim narratives?
```

Output Schema:

```
{
"type": "object",
"properties": {
"hateful_content": {"type": "boolean"},
"anti_muslim": {"type": "boolean"},
"common_theories": {
"type": "string",
"enum": [
"Love Jihad",
"Population Jihad",
"Forced Conversions",
"Muslim Appeasement",
"None"
]
}
},
"required": [
"hateful_content",
"anti_muslim",
"common_theories"
]
}
```

C.4 Health Advice Detection

Prompt:

```
Task: Determine whether this
[Text/Image/Video] contains health-related
advice or claims.
  Definition: Any content that makes claims
linked to health, including: medical
advice or treatments, health-related
claims or remedies, disease prevention
or cure suggestions, dietary or lifestyle
recommendations, traditional medicine
practices.
  Instructions:
1. Analyze all content
2. Does this contain health-related
content/advice?
(Yes/No)
```

Output Schema:

```
{
"type": "object",
"properties": {
"health_advice": {"type": "boolean"}
},
"required": ["health_advice"]
}
```

C.5 News Detection

Prompt:

```
Task: Determine whether this
[Text/Image/Video] contains news-related
information.
  Definition: Content which shares: news
clips or TV news footage, reporting
on current events, updates about
incidents/developments, news-style
presentations/formats, screenshots/clips
from news sites/apps.
  Instructions:
1. Analyze the content
2. Does this contain news-related content?
(Yes/No)
```

Output Schema:

```
{
"type": "object",
"properties": {
"news": {"type": "boolean"}
},
"required": ["news"]
}
```

C.6 Political Content Detection

Prompt:

```
Task: Determine whether this
[Text/Image/Video] is political and analyze
partisan characteristics.
  Definition: Political content that: refers
to political parties or leaders, discusses
politically relevant issues, shows political
rallies/speeches/campaigns, can be explicit
or implicit.
  Partisan Content: Political content taking
a party/ideology side.
- Non-partisan: Content discussing pros/cons
of a
law without party refs
- Partisan: Election results with captions
praising
specific parties
  Propaganda Classification (if partisan):
BJP (praising Modi/Yogi/Amit Shah, BJP
rallies), Congress (praising Rahul Gandhi,
Congress events), Other (Hindutva-aligned
without explicit party).
  Instructions:
1. Analyze all aspects
2. Is this political? (Yes/No)
3. If Yes: Is it partisan? (Yes/No)
If partisan, which party/group?
(BJP/Congress/Other)
```

Output Schema:

```
{
"type": "object",
"properties": {
"political": {"type": "boolean"},
"partisan": {"type": "boolean"},
"propaganda_for": {
"type": "string",
"enum": ["BJP", "Congress", "Other"]
}
},
"required": [
"political",
"partisan",
"propaganda_for"
]
}
```

C.7 Religious Content Detection

Prompt:

```
Task: Determine whether this
[Text/Image/Video] contains religious
elements and iconography.
  Definition: Content relating to: religious
groups, religious beliefs, religious
images/symbols/ceremonies, explicit or
implicit references.
  Religious Iconography includes:
images/videos of gods and goddesses,
religious symbols displayed or worn,
religious stories or narratives, religious
ceremonies or rituals, religious buildings
or places of worship.
  Instructions:
1. Analyze all content
2. Does this contain religious content?
(Yes/No)
3. If Yes: Does it contain religious
iconography?
(Yes/No)
```

Output Schema:

```
{
"type": "object",
"properties": {
"religious": {"type": "boolean"},
"religious_iconography": {"type": "boolean"}
},
"required": [
"religious",
"religious_iconography"
]
}
```

C.8 Misinformation Detection

Prompt:

```
Task: Determine whether this
[Text/Image/Video] contains misinformation
and assess verifiability.
  Definition: False information spread
through content, regardless of intent to
mislead.
  Classification Levels: Yes (definitely
contains misinformation), Maybe (potentially
contains misinformation), No (does not
contain misinformation).
  Verifiability: For Yes/Maybe, determine
if misinformation is theoretically
verifiable. Theoretically verifiable:
A single investigative journalist could
fact-check the claim. Unverifiable: Cannot
be definitively proven/disproven (e.g.,
"Group X always does Y").
  Instructions:
1. Analyze all content
2. Does this contain misinformation?
(Yes/Maybe/No)
3. If Yes or Maybe: Can it be theoretically
verified?
```

Output Schema:

```
{
"type": "object",
"properties": {
"misinformation_part": {
"type": "string",
"enum": ["yes", "no", "maybe"]
},
"can_theoretically_verify": {
"type": "boolean"
}
},
"required": [
"misinformation_part",
"can_theoretically_verify"
]
}
```

C.9 Content Description

Prompt:

```
Please provide a detailed description of
this [Text/Image/Video].
  Instructions:
1. Analyze thoroughly including: visual
content and
scenes, audio content (speech, music,
sounds),
text overlays or captions, overall narrative
or
message.
2. Provide comprehensive description
capturing: main
topic/subject matter, key
points/arguments/story,
tone and style, notable
features/editing/production
elements.
3. Keep description clear and concise (3-4
sentences).
```

Output Schema:

```
{
"type": "object",
"properties": {
"description": {"type": "string"}
},
"required": ["description"]
}
```

C.10 Topic Modeling

Prompt:

```
Task: Identify the primary topic and five
descriptive keywords for the given text,
image, or video.
  Primary Topic: Choose ONE from the
following list of 14 topics: Religion,
Politics, Entertainment, Commerce, Poetry,
Employment, Violence, Health, Education,
Environment, Greetings, Relationships,
Sexual Content, Miscellaneous
  Secondary Keywords: Extract a list of
keywords describing the content.
  Instructions:
1. Select exactly one primary topic.
2. Return a JSON array of keyword phrases.
3. Select between 5 and 10 keywords.
4. Keywords should be neither too specific
nor too
generic.
```

Output Schema:

```
{
"type": "object",
"properties": {
"primary_topic": {
"type": "string",
"enum": [
"Religion", "Politics", "Entertainment",
"Commerce", "Poetry", "Employment",
"Violence", "Health", "Education",
"Environment", "Greetings",
"Relationships", "Sexual Content",
"Miscellaneous"
]
},
"secondary_keywords": {
"type": "array",
"items": {"type": "string"},
"minItems": 5,
"maxItems": 10
}
},
"required": [
"primary_topic",
"secondary_keywords"
]
}
```

D Examples from Dataset

D.1 Text

Message 1: Commerce (Non-viral, 09/03/2024 02:14)

Original Text (Hindi): 12 July ko Anant [ENCRYPTED] Mukesh Ambani ji de rahe hain pure Bharat ko free mein Rs.749 wala 3 m[ENCRYPTED] link par click karke apne number par recharge kare. https://tirangagame.today/

Description: This text is a promotional message for a free recharge offer by Mukesh Ambani. It is written in Hindi and includes an encrypted link. The message is promotional in nature and uses a casual tone.

Topic: Commerce **Keywords:** recharge offer, Mukesh Ambani, promotional message, Hindi language, encrypted link

Detection Category	Result
Caste	False
Claim	Misleading: False
Hateful	Content: False, Anti-Muslim: False, Theories: None
Health	Advice: False
News	False
Political	Political: False, Partisan: False, Propaganda: Other
Religious	Iconography: False, Religious: False
Misinformation	Present: Maybe, Verifiable: False
SCAW	Present: False, Verifiable: False, Scope: National

Message 2: Politics (Viral, 16/07/2024 02:43)

Original Text (Hindi): Subject: Statement on the extreme oppression of General Category people. Honorable Prime Minister and nation's [ENCRYPTED] leader, please emphasize this system so that General Category people are not oppressed... [Long message comparing experiences of General Category vs Reserved Category students across entrance exams, college fees ([ENCRYPTED]), mess fees ([ENCRYPTED]), pocket money ([ENCRYPTED]), CAT scores, campus recruitment, and GATE scores. Ends with call to forward message to save country from "reservation termites"]

Description: The text is a letter or statement written in Hindi, addressing the issue of discrimination against General Category students in India. It starts with a formal salutation to the Prime Minister, narrates comparative experiences of a General Category student and a reserved category student highlighting disparities in academic and professional achievements. The tone is passionate and emotional, with a strong call to action urging readers to forward the message.

Topic: Politics **Keywords:** discrimination, General Category students, Prime Minister, call to action, academic achievements, equal treatment

Detection Category	Result
Caste	False
Claim	Misleading: False
Hateful	Content: False, Anti-Muslim: False, Theories: None
Health	Advice: False
News	False
Political	Political: True, Partisan: True, Propaganda: BJP
Religious	Iconography: False, Religious: False
Misinformation	Present: Yes, Verifiable: True
SCAW	Present: True, Verifiable: True, Scope: National

Message 3: Politics (Non-viral, 06/05/2024 12:58)

Original Text (Hindi): "..Main apne Bahujan Samaj ke logon ko kahunga ki Bahujan Samaj Party ko vote dena apna vote kharab karna hai, wo BJP ka saath de rahe hain chahe saamne se chahe peeche se.." Akhilesh Yadav ka Mayawati par seedha hamla. Jaunpur aur Basti ke BSP ke ticket achanak badle gaye to Akhilesh Yadav ne BSP par direct attack kar diya. Jaunpur se Dhananjay Singh ki patni Shri Kala Singh aur Basti se Daya Shankar Mishra ka ticket badal kar dusre p[ENCRYPTED] nomination ka aakhiri din tha.

Description: The text discusses political maneuvering in India, specifically mentioning Akhilesh Yadav's direct attack on Mayawati and BSP in the context of the Uttar Pradesh elections. It highlights the sudden change in BSP tickets for candidates in Jaunpur and Basti, and mentions it was the last day of nomination.

Topic: Politics **Keywords:** elections, political maneuvering, Uttar Pradesh, Akhilesh Yadav, BJP, Samajwadi Party, BSP, campaigning

Detection Category	Result
Caste	False
Claim	Misleading: False
Hateful	Content: False, Anti-Muslim: False, Theories: None
Health	Advice: False
News	True
Political	Political: True, Partisan: True, Propaganda: Congress
Religious	Iconography: False, Religious: False
Misinformation	Present: No, Verifiable: False
SCAW	Present: True, Verifiable: True, Scope: National

Message 4: Miscellaneous (Non-viral, 17/09/2023 06:19)

Original Text (Hindi): Shilp vidya ke pravartak hone ke karan Bhagwan Vishwakarma ko nirman aur srijan ka aradhya mana jata hai. [ENCRYPTED] ko khubsurti di. Kisi angadh vastu ko akarshak banane ka kaam shilpkar karta hai. Is tarah ke karyon mein pratyakshat [ENCRYPTED] hai jaisa mastishk nirdesh deta hai. Manushya sharir mein sarvadhik mahatva mastishk ka hi hai. Isliye vidwanon ne sahitya, sangeet aur kala se anabhigya vyakti ko pashu saman kaha hai... Jai Shri Sitaram, Jai Vishwakarma Bhagwan [ENCRYPTED]

Description: The text discusses the importance of craftsmanship and creativity, associating it with the Hindu god Vishwakarma, the architect of the gods. It emphasizes the role of the mind in creating appealing objects and compares the knowledgeable person to a fragrant flower spreading the scent of knowledge. The tone is reverent and educational, concluding with religious invocations.

Topic: Miscellaneous **Keywords:** Vishwakarma, craftsmanship, creativity, knowledge, Hindu deity

Detection Category	Result
Caste	False
Claim	Misleading: False
Hateful	Content: False, Anti-Muslim: False, Theories: None
Health	Advice: False
News	False
Political	Political: False, Partisan: False, Propaganda: Other
Religious	Iconography: True, Religious: True
Misinformation	Present: No, Verifiable: False
SCAW	Present: False, Verifiable: False, Scope: Local

Message 5: Commerce (Non-viral, 05/07/2024 13:56)

Original Text (Hindi): Lucknow - Rajdhani mein Nagar Nigam bana raha hai Mango Park. Raibareli Road par Kisan Path ke paas hoga nirman. Lagbhag 18 crore ki lagat se banega Mango Park. Karib 15 acre mein viksit kiya ja raha Mango Park. 108 prajatiyon ke 2068 paudhon ka ropan kiya jayega. 400 square meter mein Mango Museum banane ki yojana.

Description: The text describes the development of a mango park in Lucknow, India. The park will be located near the farmer's path on Raibareli Road, and is expected to cost approximately 18 crore rupees. It will cover about 15 acres and will feature 2068 mango trees of 108 varieties. Additionally, a 400 square meter mango museum is planned.

Topic: Commerce **Keywords:** mango park, Lucknow, India, mango trees, mango museum, 15 acres, 108 varieties, 18 crore rupees

Detection Category	Result
Caste	False
Claim	Misleading: False
Hateful	Content: False, Anti-Muslim: False, Theories: None
Health	Advice: False
News	False
Political	Political: False, Partisan: False, Propaganda: Other
Religious	Iconography: False, Religious: False
Misinformation	Present: No, Verifiable: True
SCAW	Present: False, Verifiable: False, Scope: Local

D.2 Images

Image 1: Commerce (Non-viral, 16/01/2024 05:46)

Description: The image displays a stack of women's kurtas in various pastel colors – beige, pink, green, and red – with intricate embroidered designs. Each kurta has price tags indicating they cost Rs. 567 each, while one tag on top shows Rs. 907. The overall presentation suggests this is likely an advertisement for clothing retail, showcasing affordable fashion options.

Topic: Commerce **Keywords:** women's clothing, kurtas, pastel colors, intricate embroidery, price tags, retail advertisement, affordable fashion

Detection Category	Result
Caste	False
Claim	Misleading: False
Hateful	Content: False, Anti-Muslim: False, Theories: None
Health	Advice: False
News	False
Political	Political: False, Partisan: False, Propaganda: Other
Religious	Iconography: False, Religious: False
Misinformation	Present: No, Verifiable: False
SCAW	Present: True, Verifiable: True, Scope: Local

Image 2: Politics (Non-viral, 04/06/2024 03:32)

Description: The image is a live election results screen from an Indian news channel, likely covering the 2024 elections. It displays seat counts for different political alliances – NDA (National Democratic Alliance), INDIA (a coalition of opposition parties), and OTH (others) – across various states like Uttar Pradesh, Maharashtra, and Bihar. The visual layout emphasizes numerical data with bold colors to highlight each alliance's performance in specific regions alongside images of politicians involved.

Topic: Politics **Keywords:** election results, Indian news channel, political alliances, 2024 elections, seat counts

Detection Category	Result
Caste	False
Claim	Misleading: False
Hateful	Content: False, Anti-Muslim: False, Theories: None
Health	Advice: False
News	True
Political	Political: True, Partisan: True, Propaganda: BJP
Religious	Iconography: False, Religious: False
Misinformation	Present: Yes, Verifiable: True
SCAW	Present: True, Verifiable: True, Scope: National

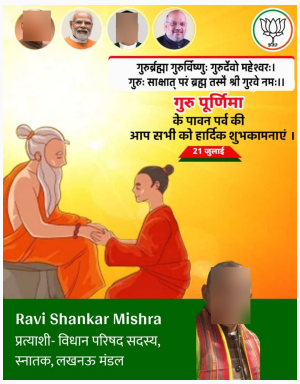
Image 3: Politics (Non-viral, 20/08/2024 14:38)

Description: The image shows a scanned newspaper clipping in Hindi discussing a Supreme Court order dated 23/06/2023 regarding land disputes, specifically concerning Wakf Boards. The article details how the court ruled against recognizing post-partition claims on Wakf properties by Bangladesh (formerly East Pakistan), stating these claims are not valid as they lack proper documentation; it also touches upon historical context dating back to 1947 and emphasizes government authority over such lands.

Topic: Politics **Keywords:** Supreme Court, land disputes, Wakf Boards, post-partition claims, historical context

Detection Category	Result
Caste	False
Claim	Misleading: True
Hateful	Content: False, Anti-Muslim: False, Theories: None
Health	Advice: False
News	True
Political	Political: True, Partisan: True, Propaganda: Other
Religious	Iconography: False, Religious: True
Misinformation	Present: Maybe, Verifiable: True
SCAW	Present: True, Verifiable: True, Scope: National

Image 4: Politics (Non-viral, 21/07/2024 04:58)



Description: This is a political greeting card for Guru Purnima, featuring Indian Prime Minister Narendra Modi alongside BJP party symbols and imagery related to Hindu spirituality. The central image depicts two figures in prayer with text wishing recipients well on July 21st, while at the bottom it includes Ravi Shankar Mishra's name and his position as a candidate for Legislative Council member from Lucknow District. The overall tone is respectful and celebratory, aiming to connect religious observance with political messaging.

Topic: Politics **Keywords:** greeting card, Guru Purnima, Indian Prime Minister, BJP party, Hindu spirituality, political messaging, Legislative Council, candidate, Lucknow District, religious observance

Detection Category	Result
Caste	False
Claim	Misleading: False
Hateful	Content: False, Anti-Muslim: False, Theories: None
Health	Advice: False
News	False
Political	Political: True, Partisan: True, Propaganda: BJP
Religious	Iconography: True, Religious: True
Misinformation	Present: No, Verifiable: False
SCAW	Present: True, Verifiable: True, Scope: National

Image 5: Violence (Non-viral, 19/12/2023 05:08)

Description: The image depicts a tense standoff between police officers and a crowd of people, likely during a protest or public disturbance in an urban setting. Several uniformed policemen are attempting to control individuals within the densely packed group, with visible expressions of frustration and resistance on both sides; the scene is dimly lit suggesting it is nighttime or indoors, contributing to a somber and potentially volatile atmosphere.

Topic: Violence **Keywords:** police, crowd, protest, control, resistance, urban, tension, nighttime

Detection Category	Result
Caste	False
Claim	Misleading: True
Hateful	Content: False, Anti-Muslim: False, Theories: None
Health	Advice: False
News	True
Political	Political: True, Partisan: False, Propaganda: Other
Religious	Iconography: False, Religious: False
Misinformation	Present: Maybe, Verifiable: True
SCAW	Present: True, Verifiable: True, Scope: Local

D.3 Videos

Video 1: Religion (Viral, 06/05/2024 11:34)

Video Link: http://bit.ly/3LKaS62

Description: The video shows a religious procession in front of a temple. A woman in a pink and green sari performs a classical Indian dance in front of a large group of people, mostly women, who are singing and carrying small oil lamps. The group is singing a devotional song in Hindi, and the dancer's movements are synchronized with the music. The temple is decorated with lights, and there is a colorful rangoli on the ground. The procession ends with a decorated cart carrying statues of deities.

Topic: Religion **Keywords:** religious procession, classical Indian dance, devotional song, Hindi, temple decoration

Detection Category	Result
Caste	False
Claim	Misleading: False
Hateful	Content: False, Anti-Muslim: False, Theories: None
Health	Advice: False
News	False
Political	Political: False, Partisan: False, Propaganda: Other
Religious	Iconography: True, Religious: True
Misinformation	Present: No, Verifiable: False
SCAW	Present: True, Verifiable: True, Scope: Local

Video 2: Politics (Viral, 10/05/2024 16:26)

Video Link: https://bit.ly/4bH7Xp8

Description: The video is a political commentary criticizing the current Indian government led by Prime Minister Narendra Modi. It features a woman speaking directly to the camera, interspersed with black and white clips of Modi speaking, as well as images and video clips highlighting various issues. The woman argues that Modi's policies have led to increased unemployment, rising fuel prices, high gas cylinder costs, a large national debt, increased poverty, and the favoring of his friends with the country's wealth. The video concludes by urging viewers not to give Modi another chance in power.

Topic: Miscellaneous **Keywords:** political commentary, Modi criticism, unemployment, inflation, farmers protests

Detection Category	Result
Caste	False
Claim	Misleading: True
Hateful	Content: False, Anti-Muslim: False, Theories: None
Health	Advice: True
News	True
Political	Political: True, Partisan: True, Propaganda: Other
Religious	Iconography: True, Religious: True
Misinformation	Present: Maybe, Verifiable: True
SCAW	Present: True, Verifiable: True, Scope: National

Video 3: Miscellaneous (Viral, 01/03/2024 17:37)

Video Link: https://bit.ly/4pF0bzC

Description: The video begins with a man in a blue shirt sitting across from a jeweler, who is calculating the cost of jewelry the man is selling. The jeweler tells him the total is 355,000, but the man says he is still 150,000 short. The man explains that his daughter is sick and needs an operation costing 300,000 to 500,000. The jeweler remembers UNM Children's Hospital in Kamrej that offers subsidized or free operations for children. The video shows the family visiting the hospital, meeting Dr. Rajesh Shah and Dr. Manish Jain, who assure them the operation can be done at subsidized rates.

Topic: Miscellaneous **Keywords:** hospital promotion, subsidized healthcare, children's surgery, family struggle, medical costs

Detection Category	Result
Caste	False
Claim	Misleading: False
Hateful	Content: False, Anti-Muslim: False, Theories: None
Health	Advice: True
News	True
Political	Political: True, Partisan: True, Propaganda: BJP
Religious	Iconography: True, Religious: True
Misinformation	Present: Maybe, Verifiable: True
SCAW	Present: True, Verifiable: True, Scope: Local

Video 4: Politics (Non-viral, 06/04/2024 10:22)

Video Link: https://bit.ly/4qVxrn5

Description: The video depicts large-scale protests in Israel against Prime Minister Benjamin Netanyahu's government over its handling of recent events involving war and hostage negotiations. The first scene shows an aerial view of massive crowds gathered on city streets at night under bright streetlights. Subsequent frames capture close-up shots of protesters holding signs with messages like "FREE ISRAEL FROM NETANYAHU" and images depicting political figures. Protesters express frustration through slogans such as "This government is doing horrendous things."

Topic: Politics **Keywords:** protests, government, Netanyahu, dissent, accountability

Detection Category	Result
Caste	False
Claim	Misleading: True
Hateful	Content: False, Anti-Muslim: False, Theories: None
Health	Advice: False
News	True
Political	Political: True, Partisan: False, Propaganda: Other
Religious	Iconography: False, Religious: False
Misinformation	Present: Maybe, Verifiable: True
SCAW	Present: True, Verifiable: True, Scope: National

Video 5: Environment (Non-viral, 04/07/2024 15:34)

Video Link: https://bit.ly/4sHmIOX

Description: The video shows close-up footage of someone wearing orange gloves gently handling a small white bird with red markings on its head and wings. The person appears to be examining or caring for the bird carefully while holding it in their hands. In the background, there is an outdoor setting featuring potted plants with green leaves, suggesting a garden-like environment. There is no visible text overlay or audio present.

Topic: Environment **Keywords:** bird care, garden setting, orange gloves, red markings, outdoor environment

Detection Category	Result
Caste	False
Claim	Misleading: True
Hateful	Content: False, Anti-Muslim: False, Theories: None
Health	Advice: False
News	False
Political	Political: False, Partisan: False, Propaganda: Other
Religious	Iconography: False, Religious: False
Misinformation	Present: No, Verifiable: True
SCAW	Present: False, Verifiable: True, Scope: National

E Topic Explanations

Table 3 contains the primary topic explanations.

Topic	Definition
Commerce	Messages related to buying, selling, pricing, promotions, or commercial transactions.
Politics	Discussions of political actors, parties, elections, public policy, or governance.
Religion	Content referencing religious beliefs, practices, institutions, or identities.
Greetings	Short social interactions such as greetings, wishes, congratulations, or casual salutations.
Education	Messages concerning schools, exams, learning, teaching, or educational institutions.
Entertainment	Content related to movies, music, celebrities, sports, or leisure activities.
Environment	Discussions about nature, climate, pollution, conservation, or environmental issues.
Health	Messages about physical or mental health, medicine, diseases, or healthcare services.
Violence	References to physical harm, threats, conflict, or violent acts.
Employment	Content related to jobs, recruitment, work conditions, or career opportunities.
Poverty	Discussions of economic hardship, inequality, unemployment, or lack of basic resources.
Sexual Content	Messages containing sexual references, explicit language, or adult themes.

Table 3: Primary topic categories with one-line definitions used for chat message classification.

F Performance Tables

Figure 12 shows performance of different experiments across different modalities and viralities and the SMOTE vs baseline performance.

G Topics Identified from News and Used for Matching

Table 4 has the list of topics analyzed and information regarding the matched messages.

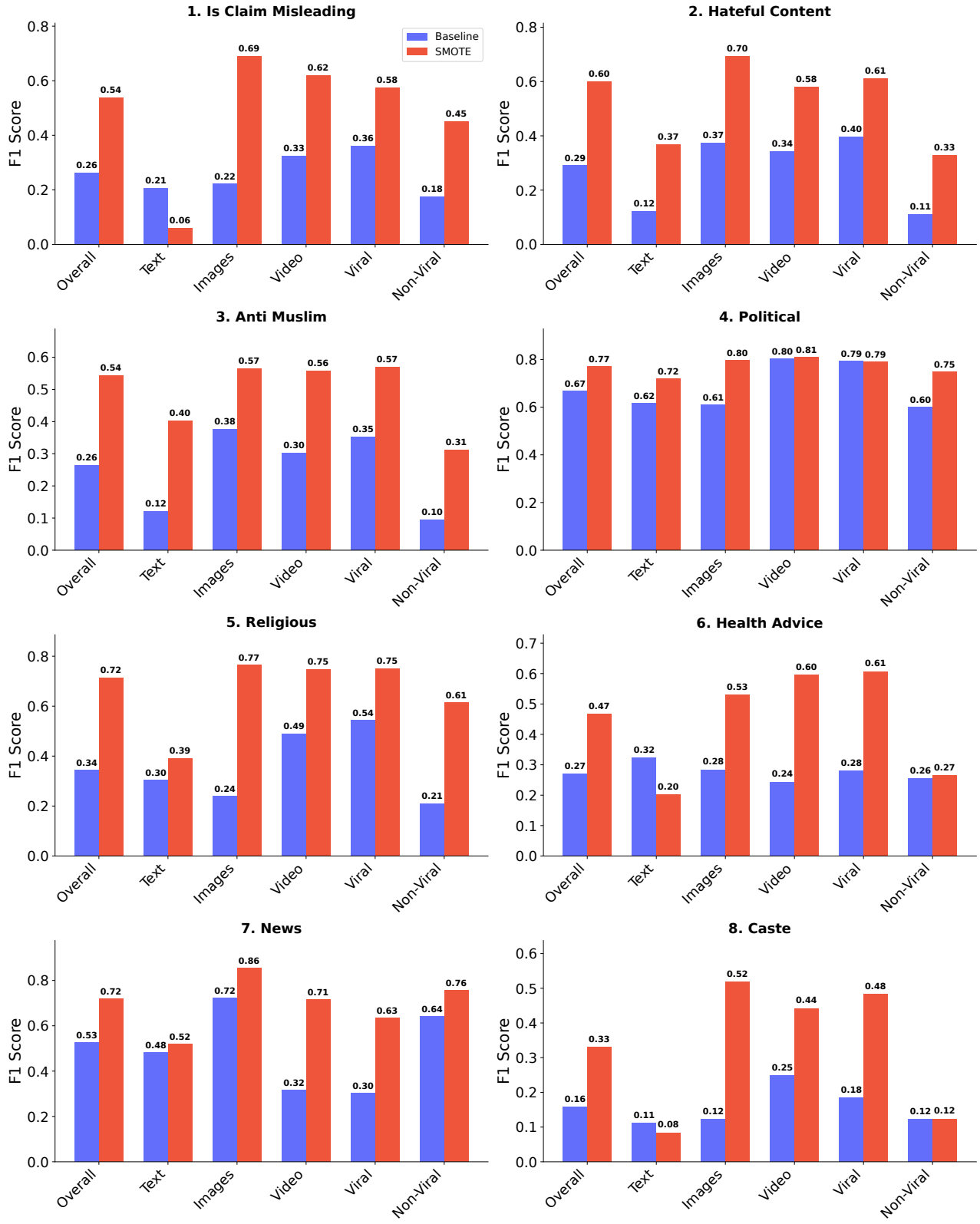


Figure 12: Performance of the large language models on various categories like Misinformation, Hate speech and topics like Politics, religion, etc.

#	Topic	Description	Stance	Msgs	Viral %
1	Liquorscam	Delhi liquor policy controversy	Pro-Gov	397	38.04
2	Swati Maliwal	AAP MP assault allegations	Pro-Gov	211	47.39
3	Modi 3 Oath	PM Modi third term swearing-in	Pro-Gov	173	27.75
4	Pm Kanyakumari	PM meditation at Vivekananda Rock	Pro-Gov	149	51.68
5	Raga Amethi	Rahul Gandhi Amethi election	Anti-Gov	133	40.60
6	NEET Controversy	Medical entrance exam irregularities	Anti-Gov	127	13.39
7	Modi Russia Austria	PM foreign visits to Russia and Austria	Pro-Gov	117	48.72
8	Pune Porsche	Teen drunk driving accident	Neutral	108	24.07
9	Sam Pitroda	Congress leader controversial statements	Anti-Gov	102	50.00
10	Ambani Wedding	Celebrity wedding of Mukesh Ambani's son Anant	Neutral	98	25.51
11	Kejriwal Release	Delhi CM Arvind Kejriwal release from custody	Anti-Gov	97	37.11
12	Ram Navmi	Hindu festival celebrations	Neutral	97	16.49
13	India T20 Win	Cricket World Cup victory	Neutral	95	4.21
14	Muslim Reservation	Debate on reservation policies	Pro-Gov	93	53.76
15	Union Budget	2024 Union Budget announcements	Pro-Gov	90	26.67
16	PM Modi Campaigns	PM election rallies and speeches	Pro-Gov	85	60.00
17	Delhi Water	Water crisis in national capital	Neutral	65	32.31
18	G7	PM Modi at G7 summit	Pro-Gov	58	39.66
19	Padma Awards	Topic related to Padma Awards	Neutral	45	8.89
20	Hathras Incident	Stampede at religious gathering	Anti-Gov	42	19.05
21	Heatwave	Extreme heat conditions across India	Neutral	42	35.71
22	OBC Reservation	OBC reservation debates	Pro-Gov	42	47.62
23	Babu Swearing In	Topic related to babu swearing in	Neutral	34	35.29
24	Kargil Diwas	Kargil War commemoration	Pro-Gov	29	17.24
25	Owaisi Jai Palestine	AIMIM leader pro-Palestine statement	Pro-Gov	25	48.00
26	Kanwar Yatra	Hindu pilgrimage and related controversies	Neutral	24	8.33
27	Ebrahim Raisi Death	Iranian President helicopter crash	Neutral	24	12.50
28	Delhi Coaching Flood	Topic related to Delhi coaching flood	Neutral	22	22.73
29	Naveen Swearing In	Topic related to Naveen swearing in	Neutral	22	18.18
30	Intl Yoga Day	International Yoga Day celebrations	Neutral	21	9.52
31	Amit Shah Fake	Controversy over fake video of Home Minister	Pro-Gov	19	57.89
32	Arvind Kejriwal Jail	Delhi CM arrest and legal proceedings	Pro-Gov	19	42.11
33	Raaj Anand	Political figure developments	Neutral	18	27.78
34	Gourav Vallabh	Congress leader joining BJP	Pro-Gov	17	35.29
35	Ram Sacrificial Goat	Religious controversy	Neutral	16	31.25
36	Airport Roof Collapse	Infrastructure failure at Delhi Airport	Anti-Gov	15	40.00
37	Evm Manipulation	Topic related to EVM manipulation	Neutral	14	21.43
38	Trump Shot	Topic related to Trump being shot	Neutral	13	0.00
39	Train Derailed	Train accident in Assam	Neutral	13	0.00
40	Jagan Mohan	Andhra Pradesh political developments	Neutral	13	23.08
41	Assam Campaign	Assam CM Himanta Biswa election campaigning	Pro-Gov	13	7.69
42	Keshav Maurya	Topic related to Keshav Maurya	Neutral	13	15.38
43	Kangana Slap	BJP MP slapped at airport	Pro-Gov	12	16.67
44	Prajwal Revanna	Karnataka politician controversy	Pro-Gov	12	16.67
45	Covishield Sideffects	COVID vaccine side effects controversy	Anti-Gov	10	20.00

Table 4: Complete list of topics analyzed with descriptions and political stance classification.