

# Beyond the Ballot: TikTok Virality and Political Engagement in Nepal's 2022 Elections

Nima Thing

In partial fulfillment of the requirements for the degree of  
Master of Data Science for Public Policy

Hertie School

Supervisor: Prof. Dr. Simon Munzert

Academic Year: 2025

Matriculation Number: 231291

Word Count: 7657

April 2025

## Abstract

Nepal's 2022 elections marked a turning point in political communication, with social media platforms such as Facebook, YouTube, and TikTok emerging as a powerful tool for political engagement amid widespread youth disillusionment with traditional parties. Beyond mere platform adoption, virality on TikTok—measured through likes, shares, and comments—became a crucial mechanism for amplifying political narratives and shaping public discourse.

Despite TikTok's growing influence, little research has systematically examined how political content achieves virality on the platform, particularly in low-income, multilingual democracies like Nepal. Understanding and predicting virality is not only methodologically innovative—leveraging multimodal machine learning on audio, image, and text data—but also critical for unpacking how digital spaces shape electoral dynamics.

This study analyzes 2,964 political TikTok videos accessed via TikTok's Research API, applying machine learning (XGBoost) and statistical modeling (OLS regression) to predict engagement outcomes. The XGBoost model achieved approximately 60% accuracy, identifying sender characteristics as key predictors (mean  $|\text{SHAP}| \approx 0.4$ ), while also revealing algorithmic biases favoring established creators, raising equity concerns. Furthermore, communication styles such as humor and charismatic leadership, particularly when tied to independent political narratives, significantly boosted virality, reflecting Nepal's culturally diverse digital landscape.

The research provides an open-source multi-modal virality prediction pipeline, operationalized engagement metrics, and a cross-platform framework adaptable to platforms like YouTube Shorts. It offers critical insights for political strategists and policymakers aiming to promote equitable online political discourse and strengthen democratic participation in emerging digital democracies like Nepal.

# Contents

List of Abbreviations . . . . .	v
<b>1 Introduction</b>	<b>1</b>
<b>2 Literature Review</b>	<b>4</b>
2.0.1 Social Media and Political Communication . . . . .	4
2.0.2 Virality Prediction on Social Media . . . . .	5
2.0.3 Communication Styles and Content Themes . . . . .	5
<b>3 Research Questions</b>	<b>7</b>
<b>4 Research Design</b>	<b>8</b>
4.1 Data Sources and Collection Strategy . . . . .	8
4.1.1 Sampling . . . . .	9
4.2 Feature Engineering . . . . .	11
4.2.1 Operationalizing Virality (Dependent Variable) . . . . .	11
4.3 Multimodal Feature Extraction . . . . .	12
4.3.1 Textual Features . . . . .	12
4.3.2 Visual Features . . . . .	13
4.3.3 Platform Features . . . . .	13
4.3.4 Sender Features . . . . .	13
4.3.5 Feature Integration & Pre-Processing . . . . .	14
4.4 Sentence Embedding Model Selection . . . . .	15
4.5 Analytical Framework . . . . .	16
4.5.1 Machine Learning for Virality Prediction (RQ1) . . . . .	16

4.5.2	Content Analysis for Communication Styles and Political Affiliations (RQ2)	16
4.6	Data Annotation and Validation	17
4.6.1	Audio Transcripts	17
4.6.2	Video Descriptions	17
4.6.3	Image Embeddings	17
4.6.4	Style and Content Labels for RQ2	18
4.6.5	Discussion	18
4.7	Ethical Considerations	18
<b>5</b>	<b>Results</b>	<b>19</b>
5.1	Exploratory Analysis	19
5.1.1	Number of Videos Over Time	19
5.1.2	Post Volume Heat Map	20
5.1.3	Top Keywords in Video Descriptions	20
5.2	Virality Prediction Results	22
5.2.1	Model Performance	22
5.2.2	Performance of the Tuned XGBoost Model	23
5.3	Feature Importance Analysis	24
5.3.1	Sender Characteristics (Mean SHAP: 0.0480)	25
5.3.2	Content Characteristics	25
5.3.3	Platform Characteristics (Mean SHAP: 0.0004)	26
5.3.4	Local Explanation of a High-Virality Prediction	27
5.4	Error Analysis	28
5.5	Communication Styles and Political Content Themes	29
5.5.1	Model Performance	29
5.5.2	Dataset and Category Inclusion	29
5.5.3	Main Effects	31

5.5.4	Interaction Effects . . . . .	31
<b>6</b>	<b>Discussion</b>	<b>34</b>
6.1	Predictability of Political Virality (RQ1) . . . . .	34
6.1.1	Sender Influence and Algorithmic Bias . . . . .	34
6.1.2	Limits of Predictive Modeling . . . . .	35
6.2	Style–Theme Interactions in Political Communication (RQ2) . . .	36
<b>7</b>	<b>Conclusion</b>	<b>38</b>
7.1	Limitations . . . . .	40
7.2	Future Research . . . . .	43
<b>8</b>	<b>Appendix</b>	<b>i</b>
	Appendix . . . . .	ii
8.1	GitHub Repository Details . . . . .	x
	AI Disclosure . . . . .	xxii
	Statement of Authorship . . . . .	xxiii

# List of Tables

5.1	Model Comparison Results (Default Parameters)	22
5.2	Classification Report for Tuned XGBoost Model	23
5.3	Mean SHAP Values for Feature Groups	24
5.4	Top Communication Styles for <code>video_diversification_id = 10083.0</code>	26
5.5	Summary Statistics by Content Theme	30
8.1	Dataset summary statistics	i
8.2	Top Values of <code>audio_pca_0</code> and Corresponding Content	iii
8.3	Bottom Values of <code>audio_pca_0</code> and Corresponding Content	iii
8.4	Selected Diversification IDs and Virality Class Proportions	iii
8.5	Error Rates by Sender Characteristics	iv
8.6	TikTok Communication Style Codebook for Nepal's 2022 Local Elections	vi
8.7	Content Themes for Zero-Shot Classification	ix
8.8	Comparison of sentence embedding model performance on sample Nepali election text pairs	xi
8.9	Summary Statistics for OLS Regression Predicting <code>z_bc_virality</code>	xii
8.10	Main Effects for OLS Regression Predicting <code>z_bc_virality</code>	xiii
8.11	Interaction Effects for OLS Regression Predicting <code>z_bc_virality</code>	xiv
8.12	Video Diversification Labels and Counts	xv
8.13	Top Content Themes for <code>video_diversification_id = 10083.0</code>	xvi

## List of Abbreviations

Abbreviation	Definition
API	Application Programming Interface
CPN UML	Communist Party of Nepal (Unified Marxist-Leninist)
LaBSE	Language-agnostic BERT Sentence Embedding
MiniLM	Mini Language Model
OLS	Ordinary Least Squares
SHAP	SHapley Additive exPlanations
TikTok	Social media platform for short-form videos
XGBoost	eXtreme Gradient Boosting

# 1. Introduction

Nepal's 2022 local and provincial elections (2079 B.S.) marked a significant shift in the landscape of political communication, with TikTok emerging as a prominent platform alongside more established channels such as Facebook and YouTube (Nepali Times, 2025). These platforms increasingly competed for influence, effectively cannibalizing attention and engagement from traditional media as political parties sought to expand their digital reach. One of the most notable developments was the growing use of these platforms as agenda-setting tools, particularly in mobilizing and engaging the country's youth amid rising disillusionment with legacy political structures (Dahal, 2023; Explainers, 2024).

Nepal's local election cycle (prior to provincial/federal election), social media platforms, particularly TikTok, significantly amplified grassroots movements and youth-driven political narratives. Independent candidates, such as Balendra Shah (Balen Shah), capitalized on this shift, leveraging platforms to bypass traditional media and party structures (Deutsche Welle, 2022). The "lauro" (stick) hashtag, symbolizing the independent movement, gained traction, mobilizing voter support (The Annapurna Express, 2022). Influential creators like "Routine of Nepal Banda," with over 1 million TikTok followers, further boosted visibility for new generation candidates (Dahal, 2023). This new information ecosystem culminated in the unexpected victory of Balen Shah, a rapper without formal political experience, who secured the Kathmandu mayoral seat with 38.6% of the vote, outpacing major party candidates (Wikipedia contributors, 2022). While three big traditional parties - Nepali Congress and Communist Parties (CPN-UML and CPN-Maoist Centre) retained dominance overall, alongside their existing visibility on legacy network and social media campaigns, independent candidates, particularly Balen Shah gained significant visibility on these platforms, mobilizing youth and reshaping public perception outside traditional political frameworks (Deutsche Welle, 2022).



TikTok’s unique affordances, such as its algorithmic exposure of low-view videos, its automated “for You” page curation, and its built-in creative tools - has separated it from other legacy platforms that rely heavily on pre-existing networks or paid-content promotion (Guinaudeau et al., 2022). The centrality of TikTok’s algorithm in enabling “virality-from-nowhere”—unexpected viral success regardless of a creator’s status or affiliation—fundamentally reshapes political communication during elections (Guinaudeau et al., 2022). As a participatory and remixable platform, TikTok fosters an environment where influence is unpredictable, allowing anyone, from independent candidates to ordinary citizens, to become a political messenger (Guinaudeau et al., 2022). As global industries have adapted to this logic, even some top Nepalese politicians have also started to engage with TikTok as communicative campaigning tool (The Himalayan Times, 2025). This shift has enabled their followers to adopt and innvoate within a distinctly local stlye of political messaging - one that often blends politcal content with cultural expressions such as mimicry, humor, and entertainment. This practices align with communication styles like “Comedic” category analysed in this study (Umansky & Pipal, 2023).

However, despite TikTok’s rich metadata and increasing API accessibility (TikTok, 2025), there has been little research examining the factors driving political virality through a multi-modal lens. TikTok’s role in Nepal—a low-income, multi-lingual nation with a history of political volatility—remains largely underexplored as a politican communication tool, particularly following its controversial ban in 2023 and reinstatement in 2024 due to concerns over misinformation and digital governance (Lamichhane, 2024; Reuters, 2024). In the context of Nepal’s evolving digital democracy, addressing this research gap is essential to understand virality and its association with several forms of TikTok communication forms during electoral periods.

This study offers three key contributions:

- **Academic:** It provides a novel and open-source codebase for multimodal virality analysis, filling a gap in computational political communication research in Nepal.
- **Methodological:** It operationalizes virality through a weighted engagement score (combining views, likes, comments, shares, collects) and compares machine learning models (XGBoost, Random Forest) to identify optimal predictive approaches for TikTok data.

- **Practical:** By analyzing communication styles, cultural elements, and musical styles (e.g., rap, Nepali folk), it offers actionable insights for political candidates and content creators to enhance audience reach in Nepal's diverse digital landscape.

This research lays the groundwork for future cross-platform virality studies (e.g., TikTok, YouTube Shorts) and provides a framework for candidates to frame election content more effectively, fostering informed digital campaigning in low-income electoral contexts.

## 2. Literature Review

The rapid rise of social media has reshaped political communication, offering new avenues for voter engagement, agenda-setting, and grassroots mobilization. In Nepal, where traditional media has long dominated electoral discourse (Dahal, 2023), platforms like TikTok have introduced novel dynamics, particularly during the 2022 local and federal/provincial elections. This literature review synthesizes research on social media’s role in political communication, TikTok’s emergence as a political platform, virality prediction, and the interplay of communication styles and content themes, identifying critical gaps that this study addresses in the context of Nepal’s evolving digital democracy.

### 2.0.1 Social Media and Political Communication

Social media platforms have transformed political communication by enabling rapid information dissemination, amplifying user-generated content, and fostering direct candidate–voter interactions (Van Dijck & Poell, 2013). (McGregor, 2019) argue that platforms like Facebook and Twitter (now X) play a significant role in agenda-setting, influencing public perceptions of key issues during election cycles. For instance, studies in Western democracies have shown that social media activity can enhance voter mobilization, with increased engagement (e.g., likes, shares) correlating with higher voter turnout in some contexts (Bode & Dalrymple, 2016). However, findings are mixed, with (Piatak & Mikkelsen, 2021) noting that online engagement does not always translate into electoral success, particularly in regions with limited digital infrastructure.

In Nepal, social media has become a vital space for political expression, especially among youths frustrated with the country’s history of political volatility (Bhandari, 2024). (Dahal, 2023) highlight that unorganized groups and individual creators played a significant role in shaping public discourse, often amplifying anti-establishment sentiments. However, much of the research on social media in Nepal remains qualitative, focusing on user perceptions rather than systematic, data-driven analyses of engagement dynamics.

## 2.0.2 Virality Prediction on Social Media

Predicting virality on social media has been a focus of computational social science, with studies identifying key drivers such as content features, platform dynamics, and sender characteristics (Berger & Milkman, 2012). (Sah & Jordan, 2025) used multi-modal features (text, images, user metadata) to predict Reddit Meme virality, where certain temporal features and content layout was quite significant. Additional work include an explored video-based platforms, with (Nisa et al., 2021) applying XGBoost Learning algorithm to predict YouTube video popularity using visual and audio features, however, here popularity is primarily based on view\_counts. Besides, the study also doesn't formally operationalize virality, but rather exponential spread in short time, which may not fully capture overall virality as an distinct phenomenon. However, TikTok-specific studies on virality prediction on large dataset are even scarce, largely due to data access restrictions and relative novelty on opening up its research API.

In political contexts, virality prediction is particularly complex, as engagement often depends on emotional resonance and cultural relevance (Berger & Milkman, 2012). (Ingelstam, 2023) discusses in her paper a threshold of 250,000 views for “mildly viral” political content on TikTok, but such metrics lack standardization, especially in non-Western settings like Nepal where audience sizes and engagement patterns differ. Moreover, while TikTok’s Research API have enabled multi-modal analysis in regions like the U.S., EEA, UK or Switzerland (TikTok, 2025), access remains restricted in Nepal, limiting large-scale computational studies.

## 2.0.3 Communication Styles and Content Themes

The interplay between communication styles and content themes is a critical driver of engagement on TikTok, particularly in political contexts (Umansky & Pipal, 2023). (Schellewald, 2021) identified styles such as Comedic, Communal, and Meta on TikTok, noting that humorous content often garners higher engagement due to its emotional appeal. In political settings, (Cseri, 2024) found that styles like Charisma/Leadership—often through personalization and visuals rhetoric elicit higher engagement metrics. Styles like Critique/Frustration also resonate with audiences by amplifying anti-establishment sentiments, a trend observed in Nepal’s 2022 elections (Dahal, 2023).

However, existing code-books for TikTok communication styles (Umansky & Pipal, 2023) are often Western-centric, lacking cultural specificity for contexts like Nepal. Nepalese political content frequently incorporates local elements, such as mimicry, folk music, and rap, reflecting the country’s linguistic and cultural diversity. For instance, independent candidates like Balen Shah promotional campaign was mostly rap/hiphop to connect with youth voters, blending political themes with entertainment (Menge & Dhakal, 2022). Despite this, there is no standardized codebook for Nepalese election contexts, nor has research systematically examined how these styles interact with political themes (e.g., support for legacy parties vs. independents) to drive virality. This study addresses this gap by creating a culturally tailored codebook and employing regression analysis to investigate the interactions between communication styles and content themes in TikTok videos from Nepal’s 2022 Local Election.

### 3. Research Questions

The study addresses two research questions (RQs):

**RQ1: Can the virality of political TikTok videos be predicted using pre-upload content, platform, and sender characteristics?**

**RQ2: How do different communication styles and political content themes interact to influence the virality of TikTok videos during Nepal's local elections?**

## 4. Research Design

This study employs a mixed-methods approach, combining machine learning for virality prediction (RQ1) and regression-based content analysis for style–theme interactions (RQ2), to examine TikTok’s role in Nepal’s 2022 local elections. The design leverages multimodal data (text, audio, visual) and platform metadata, integrating computational techniques with political communication theory.

### 4.1 Data Sources and Collection Strategy

The dataset comprises 28,165 TikTok videos collected from March 15, 2022 (one month before candidacy registration) to May 13, 2022 (election day), capturing the campaign period of Nepal’s 2022 local elections. A hybrid approach was used, combining the Official TikTok Research API, the Unofficial TikTok-API, and a custom subclass of TikTok-Content-Scraper (Bukold, 2025). The Official API collected video metadata (e.g., views, likes, shares, hashtags, descriptions) and user metadata (e.g., follower count, verified status) but was limited by quotas (1,000 users/day, 100,000 videos/day). The Unofficial API supplemented user-level data, while the custom scraper captured platform-specific features like `video_diversification` labels, duet/stitch behavior, and music metadata.

Videos were collected via hashtag-based searches (e.g., `#NepalElection2022`, `#चुनाव`, `#LocalElection2022`), adapted from (Pinto et al., 2024) (see Appendix 8). After removing duplicates and irrelevant content, the final dataset included 28,165 videos. Initial features collected include:

- **Platform/Engagement Metrics:** Views, likes, shares, comments.
- **User Features:** Follower count, video count, verified status.
- **Content Metadata:** Hashtags, video descriptions, audio metadata.
- **Temporal Features:** Posting timestamps (UTC).

The data collection pipeline and schema are illustrated in Figures 4.1 and 4.2, respectively.

#### 4.1.1 Sampling

For multimodal analysis, a stratified sample of 3,000 videos was drawn from the 28,165 videos based on view count percentiles:

- **High-Virality:**  $\geq$  95th percentile ( $\geq$  8,000 views;  $\sim$ 5% of dataset).
- **Medium-Virality:** 60th to 95th percentile ( $\sim$ 35% of dataset).
- **Low-Virality:**  $\leq$  60th percentile ( $\sim$ 60% of dataset).

After removing videos that became private or irrelevant, the final sample was 2,964 videos.

As illustrated in Figure 4.2, the dataset is further enriched post-sampling with audio transcripts and image analysis. These additional features, derived through the processing pipeline depicted in Figure 4.1, complete the multimodal data collection for the sampled dataset (detailed in the following section).

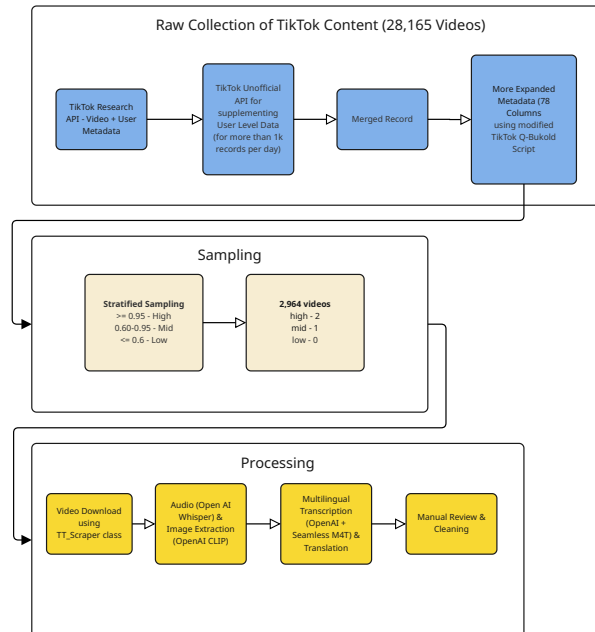


Figure 4.1: Data Collection Pipeline



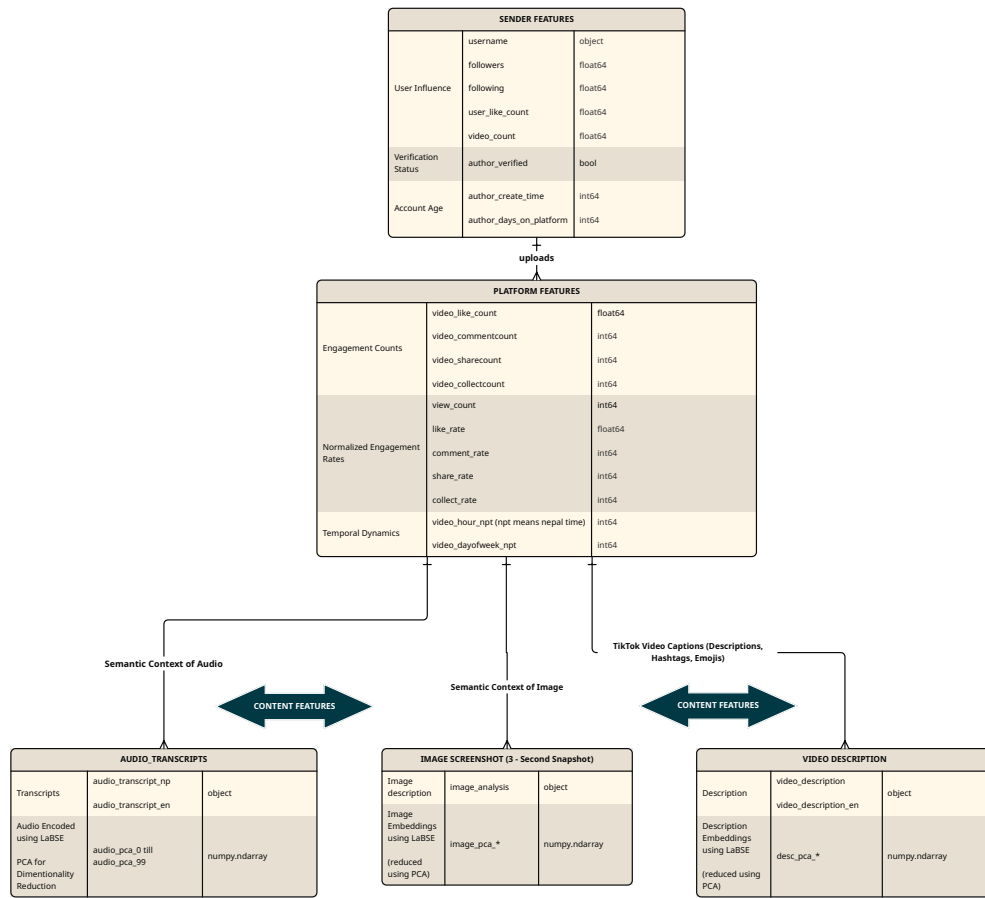


Figure 4.2: Data Schema Collection

## 4.2 Feature Engineering

### 4.2.1 Operationalizing Virality (Dependent Variable)

While there is no universally accepted standard for measuring virality on TikTok, several research studies and commercial tools have proposed their own metrics and predictive models (Roring, 2024; Sah & Jordan, 2025). Moreover, no prior research has applied standard virality criteria to TikTok videos in the context of Nepal's 2022 elections, necessitating a flexible approach to capture its role in shaping public attention. In this study, virality is operationalized as a composite score emphasizing relative audience engagement, prioritizing interaction over raw popularity. Four **normalized engagement rates** - like, comment, share, repost, were calculated.

The rates were combined with the logarithm of raw view count to form a **virality score**, using a weighted formula that emphasizes active user engagement. The final formula is:

$$\begin{aligned} \text{virality score} = & 0.35 \cdot \log(1 + \text{views}) + 0.25 \cdot \log(1 + \text{like rate}) + 0.20 \cdot \\ & \log(1 + \text{comment rate}) + 0.15 \cdot \log(1 + \text{share rate}) \\ & + 0.05 \cdot \log(1 + \text{collect rate}) \quad (4.1) \end{aligned}$$

Weightings of (0.35, 0.25, 0.20, 0.15, 0.05) were assigned to prioritize active engagement behaviors. To assess the robustness of this approach, a sensitivity analysis was conducted using equal weighting across components. The resulting agreement between the original and equal-weight virality labels was 99.53%, suggesting strong stability of the classification. Given the exploratory nature of this study and time constraints, weightings were fine-tuned iteratively based on observed distributions without external validation. The score was transformed into a ternary classification label for RQ1:

- **Low Virality (0):** Bottom third of scores.
- **Moderate Virality (1):** Middle third.
- **High Virality (2):** Top third.

For RQ2, a log-transformed version of the virality score was used as a continuous regression target to account for its skewed distribution.

## 4.3 Multimodal Feature Extraction

This study extracted multi-modal features from a stratified sample of 2,964 Tik-Tok videos to predict virality. Features were derived from three primary modalities: audio transcripts, video descriptions, and visual frames.

### 4.3.1 Textual Features

Textual or Content features aim to capture the semantic, emotional, and contextual characteristics of a video's actual media content. These include both linguistic and visual elements:

#### Audio Transcripts

- Transcribed using OpenAI's GPT-4o Whisper API (OpenAI, 2024), primarily in Nepali with minor instances of Hindi, Bhojpuri/Maithili, Urdu, and English. Manual corrections added contextual annotations (e.g., `{{Party_Name}}` Song). Transcripts were embedded using the BGE-m3 model (Chen et al., 2024) and reduced via PCA (`audio_pca_0` to `audio_pca_99`).

#### Video Descriptions

- **Video Descriptions:** Language detection via LangDetect (Nakatani, 2014) identified 494 English and 2,470 non-English (Romanized Nepali-English, Devanagari) descriptions. Non-English descriptions were translated to English using GPT-4o (OpenAI, 2024), preserving hashtags and mentions. Descriptions were embedded with BGE-m3 and reduced via PCA (`desc_pca_*`).

```
prompt = f"Translate the following text to English but keep all
hashtags and mentions (like #nepal, @username)
unchanged:\n\n{masked_text}"
```

### 4.3.2 Visual Features

- **Visual Features:** A frame at the 3-second mark was extracted (TikTok, 2023), captioned using GPT-4o Vision API (OpenAI, 2024), embedded with BGE-m3, and reduced via PCA (`image_pca_*`).
- **Sentiment Scores:** Extracted for audio, descriptions, and image captions using a pretrained sentiment model.
- **Lexical Features:** Included `audio_word_count`, `image_word_count`, and `hashtag_count`.

### 4.3.3 Platform Features

Platform features refer to platform-mediated affordances and behavioral traces that reflect how users engage with the content:

- **Temporal Dynamics:** `video_hour_npt`, `video_dayofweek_npt`.
- **Metadata:** `video_diversification` labels, duet/stitch indicators.

### 4.3.4 Sender Features

Sender features capture the video uploader’s social capital and credibility:

- **Creator Tier Metrics:** A categorical feature, `creator_tier`, was derived from follower counts to analyze non-linear effects of audience reach on virality. Using the empirical distribution of follower counts across 2,964 videos (25th percentile:  $\approx 983$ , 50th:  $\approx 2,191$ , 75th:  $\approx 8,120$ , 90th:  $\approx 48,740$ , 95th:  $\approx 108,700$ , 99th:  $\approx 376,500$ ), we defined tiers as: Regular ( $< 10,000$  followers), Micro Influencer (10,000–49,999), Macro Influencer (50,000–99,999), and Mega Influencer ( $\geq 100,000$ ). This data-driven approach ensures relevance to Nepal’s 2022 election context and enhances model interpretability.

- **Verification Status:** A binary feature, `author_verified`, indicates official TikTok verification, reflecting institutional presence or public figure status that may influence platform distribution and audience trust.
- **Account Age:** Derived from `author_create_time`, `author_days_on_platform` measures the creator's active duration on TikTok, capturing their tenure and experience, which may affect content strategy and reach.

### 4.3.5 Feature Integration & Pre-Processing

Predictor variables from platform metadata, sender attributes, and content modalities were integrated into a unified, model-ready dataset. Numerical features (e.g., follower count, interaction rates like `like_rate`, `comment_rate`) were z-score normalized for comparability, while categorical variables (e.g., day-of-week) were one-hot encoded. Multimodal embeddings from audio transcripts, image snapshots, and video descriptions were generated using a multilingual transformer model, reduced via PCA, and used in their compact latent form without further scaling, as they already captured normalized semantic similarity.

To prevent label leakage, post-performance variables (`video_like_count`, `video_commentcount`, `video_sharecount`, and their normalized counterparts `like_rate`, `comment_rate`, `share_rate`, `collect_rate`) were excluded, focusing strictly on pre-upload features like content semantics (text, image, audio) and user metadata (e.g., `creator_tier`, verification status). This structured predictor matrix ensured consistent model input and enabled nuanced evaluation of content, platform, and sender contributions to virality prediction.

## 4.4 Sentence Embedding Model Selection

We evaluate four sentence embedding models on Nepali election-related texts:

- LaBSE (Reimers & Gurevych, 2019)
- paraphrase-multilingual-MiniLM-L12-v2 (MiniLM-L12) (Reimers & Gurevych, 2020)
- all-MiniLM-L6-v2 (Reimers & Gurevych, 2021)
- BGE-M3 (Chen et al., 2024)

BGE-M3 demonstrates multi-lingual capabilities supporting 100+ languages with dense/sparse retrieval (Chen et al., 2024). The MiniLM variants offer optimized performance for semantic similarity tasks (Reimers & Gurevych, 2019).

Based on evaluation of sentence-embeddings (See Table 8.8) for 11-pairs of random sentence-text pairs, BGE-m3 was chosen for its more balanced performance in compared to other models.

## 4.5 Analytical Framework

The analysis addresses the two research questions through two complementary approaches:

### 4.5.1 Machine Learning for Virality Prediction (RQ1)

- **Model, Training, and Validation:** XGBoost was selected as the primary classifier over Random Forest and LightGBM (see Section 5.2.1). The stratified sample of 2,964 videos was split into 70% training and 30% test sets, evaluated using accuracy, precision, recall, F1-score, and AUC-ROC.
- **Explainability and Error Analysis:** SHAP (SHapley Additive Explanations) was applied to XGBoost to interpret feature contributions, where predictions are decomposed as  $f(x) = \text{base value} + \sum \text{SHAP values}$ , focusing on metadata (e.g., `user_like_count`, `author_days_on_platform`), visual/audio embeddings (e.g., `image_pca_*`, `audio_pca_*`), and descriptive text features (e.g., `desc_pca_*`, `sentiment scores`). Potential biases were assessed through qualitative error analysis of misclassified samples.

### 4.5.2 Content Analysis for Communication Styles and Political Affiliations (RQ2)

- **Communication Styles:** Videos were annotated with a revised codebook adapted from (Schellewald, 2021; Umansky & Pipal, 2023), categorizing 11 styles, with added Nepal-specific categories (e.g., Entertainment/Song, Charismatic/Leadership Appeal) to reflect cultural context and independent voices (Deutsche Welle, 2022) (see Appendix 8). Each video received a primary (`Style_1`) and secondary (`Style_2`) style, but `Style_2` was excluded due to limited coverage (40%) and time constraints.
- **Political Affiliations and Annotation:** Audio transcripts were categorized into 17 themes (e.g., campaign songs, comedic sketches) via GPT-4o zero-shot classification, then filtered into four Nepal 2022 election categories: *Nepali Congress*, *UML*, *Maoist*, and *Independent*. Labels were verified using multimodal cues (party symbols, candidate names, slogans, election songs, visuals).

- **Regression Modeling and Visualization:** OLS regression analyzed the effect of styles and affiliations on log-transformed, z-scored virality, using *Civic Awareness* as the reference style and *general election content* as the baseline theme, with dummy-encoded variables, assumed homoscedasticity, and 95% confidence intervals. A heatmap visualized mean z-virality-scores across style-theme pairs.

## 4.6 Data Annotation and Validation

A systematic annotation and validation process ensured the robustness of multimodal features and labels for the XGBoost model (RQ1) and OLS regression (RQ2) using 2,964 TikTok videos from Nepal’s 2022 elections.

### 4.6.1 Audio Transcripts

A 5% sample ( $n = 148$ ) of Nepali audio transcripts (`audio_transcript_np`), converted to PCA embeddings (`audio_pca_*`), was manually validated, achieving 87.2% accuracy by the primary author and a Cohen’s Kappa of 0.462 (moderate agreement) with a secondary evaluator (Landis & Koch, 1977). Despite dialectal challenges, embeddings were retained due to their complementary role with image (`image_pca_*`) and description (`desc_pca_*`) embeddings, with `audio_pca_0` showing a mean  $|\text{SHAP}| \approx 0.3$ .

### 4.6.2 Video Descriptions

A 5% sample ( $n = 148$ ) of translated video descriptions achieved 98.6% accuracy, preserving hashtags per best practices (Highfield & Leaver, 2016). These were converted to PCA embeddings (`desc_pca_*`) for the XGBoost model.

### 4.6.3 Image Embeddings

Image embeddings (`image_pca_*`) were generated using OpenAI’s CLIP API (Radford et al., 2021), recognized for high accuracy, and directly used in the XGBoost model without manual validation, with `image_pca_2` showing a mean  $|\text{SHAP}| \approx 0.1$  for high-virality predictions.



#### 4.6.4 Style and Content Labels for RQ2

Three annotators labeled all 2,964 videos for communication styles (`style_1_label`, e.g., Comedic) and content themes (`content_label`, e.g., party affiliation) per a predefined codebook. A 150-video sample yielded a Cohen’s Kappa of 0.622 (substantial agreement) for `style_1_label` (Landis & Koch, 1977), supporting its use in OLS regression. Content labels, derived via zero-shot classification and manually reviewed, lacked formal validation due to resource constraints.

#### 4.6.5 Discussion

Validation confirms reliability, with 98.6% accuracy for video descriptions and Kappa = 0.622 for style labels, though audio transcripts’ Kappa = 0.462 reflects linguistic challenges. CLIP embeddings align with computer vision standards, and zero-shot content labels were mitigated by manual review. This ensures dataset suitability for modeling virality and style–theme interactions, though future work could improve audio transcription using advanced multilingual NLP fine-tuned for Nepalese social media communication context.

### 4.7 Ethical Considerations

Data collection complied with TikTok’s privacy policies. No PII was collected. Approval was obtained from the Hertie Data Science Review Board. Dataset limitations and potential biases were transparently reported.

## 5. Results

### 5.1 Exploratory Analysis

Exploratory Analysis was conducted on the full dataset. For Dataset summary (see Table 8.1)

#### 5.1.1 Number of Videos Over Time

Figure 5.1 illustrates the volume of TikTok videos posted over this time frame. A noticeable rise in content production is observed as the election approaches, peaking sharply on Election Day (May 13, 2022).

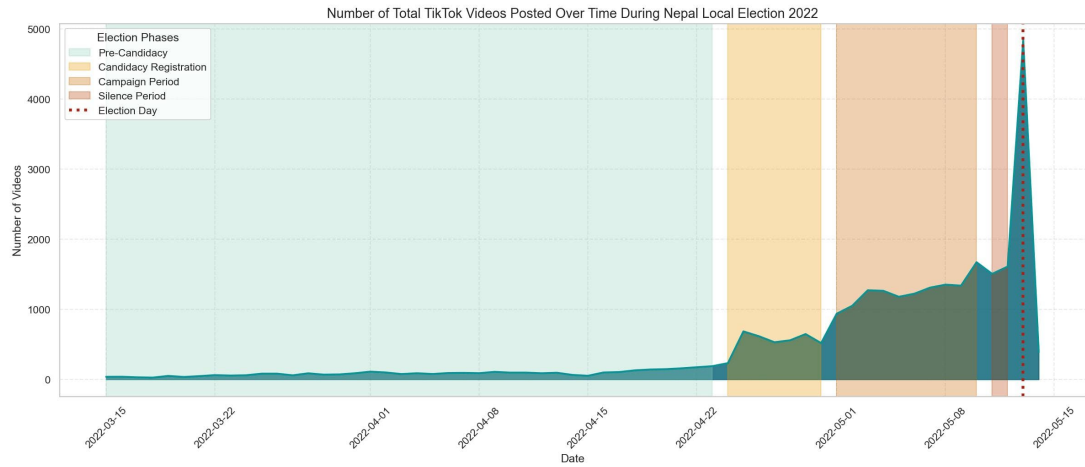


Figure 5.1: TikTok Video Post Over Time)

### 5.1.2 Post Volume Heat Map

Figure 5.2 reveals that most video posts occur on Fridays, particularly during the late afternoon and evening hours. A distinct posting peak is observed at 21:00 (9:00 PM NPT), suggesting that content creators might be strategically timing their posts to maximize visibility and engagement during the end of their weekdays.

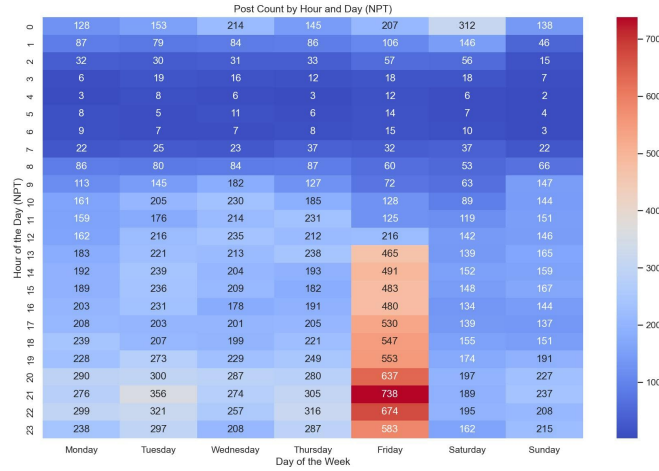


Figure 5.2: TikTok Video Post Over Time)

### 5.1.3 Top Keywords in Video Descriptions

The dataset includes 20,237 unique hashtags extracted from video descriptions.

After applying a custom stopwords filter to remove generic or repetitive terms (e.g., `fyp`, `viral`, `trending`), we generated a word cloud to visualize prominent hashtags. Figure 5.4 reveals Nepali keywords like *chunab* ("election") and *chunablagyo* ("election vibes"), as well as direct references to political campaigns. Note that some Devanagari (Nepali) script tokens are rendered as square boxes due to font encoding limitations.

Table 5.3 lists the 20 most frequently occurring hashtags in the political genre along with their respective counts.

Top\_20\_HashTags\_Political\_Support

Hashtag	Count
nepalicongress	732
cpnuml	329
congress	324
balen	299
voteforbalen	296
kpoli	286
balenshah	279
jaynepal	240
voteforchange	225
voteforsun	222
mayor	174
yemale	169
uml	167
nekapayemalay	166
prachanda	157
एमाले	123
नेकपा_एमाले_जिन्दाबाद	122
गठबन्धन	114
voteforbalenshah	97
votefortree	92

Figure 5.3: Top Political Hashtags Used During Local Election 2022



Figure 5.4: Top Filtered hashtags after removing trending terms

## 5.2 Virality Prediction Results

This section presents the results for our first research question, which aimed to study virality prediction using pre-upload content, platform, and sender characteristics. To focus on pre-upload features, post-upload engagement metrics (e.g., views, likes) were excluded. Results are organized into model performance, feature importance, ablation study, error analysis, and explainability using SHAP.

### 5.2.1 Model Performance

To select an appropriate model for predicting TikTok virality, three algorithms were compared: Random Forest, XGBoost, and LightGBM. These models were trained with default parameters on a stratified train-test split (70% train, 30% test,  $n=890$  test samples). Random Forest served as a baseline due to its simplicity, while XGBoost and LightGBM were chosen for their advanced gradient boosting capabilities, suitable for high-dimensional datasets with mixed features (numerical, categorical, embeddings). Performance was evaluated using accuracy, macro average F1-score, and AUC-ROC (one-vs-rest).

Performance metrics included accuracy, macro F1, and class-specific F1 scores for three virality classes: low (0), mid (1), and high (2). The baseline accuracy for random guessing in a three-class problem is 33.33%.

The comparison (Table 5.1) showed that LightGBM achieved the highest accuracy (0.61) and AUC-ROC (0.7972), followed closely by XGBoost (accuracy: 0.60, AUC-ROC: 0.7948), while Random Forest lagged (accuracy: 0.58, AUC-ROC: 0.7526). All models struggled with medium virality (Class 1, F1-scores: 0.48–0.52), but performed well on high virality (Class 2, F1-scores: 0.66–0.68). XGBoost was selected for further optimization due to its competitive performance, suitability with lesser data, robustness to noisy data, and flexibility for hyperparameter tuning (compared to LGB), which offered potential for improvement.

Table 5.1: Model Comparison Results (Default Parameters)

Model	Accuracy	Macro Avg F1-score	AUC-ROC (OvR)
Random Forest	0.58	0.58	0.7526
XGBoost	0.60	0.60	0.7948
LightGBM	0.61	0.60	0.7972

### 5.2.2 Performance of the Tuned XGBoost Model

The tuned XGBoost model was evaluated on the test set (n=890), achieving an accuracy of 0.60, a macro average F1-score of 0.60, and an AUC-ROC of 0.7871, slightly below the default LightGBM (AUC-ROC: 0.7972) but comparable to the default XGBoost (AUC-ROC: 0.7948). Detailed performance metrics are presented in Table 5.2

Table 5.2: Classification Report for Tuned XGBoost Model

Class	Precision	Recall	F1-score	Support
0 (Low-Viral)	0.56	0.72	0.63	294
1 (Mid-Viral)	0.53	0.49	0.51	293
2 (High-Viral)	0.75	0.60	0.67	303
<b>Macro Avg</b>	0.61	0.60	0.60	890
<b>Weighted Avg</b>	0.62	0.60	0.60	890

The tuned model maintained strong performance for low virality (Class 0, F1-score: 0.63) and high virality (Class 2, F1-score: 0.67), with high recall for Class 0 (0.72) and high precision for Class 2 (0.75). However, it continued to struggle with medium virality (Class 1, F1-score: 0.51), consistent with the baseline models, likely due to overlapping feature distributions in this range. The AUC-ROC of 0.7871 indicates robust discriminative ability across classes, though the slight decrease from the default XGBoost suggests that the tuned parameters may have prioritized balance over maximizing AUC-ROC.

### 5.3 Feature Importance Analysis

SHAP (SHapley Additive exPlanations) values were used to assess feature importance, measuring the average impact of each feature on the model's predictions across the three classes on the full test set (n=890). The top 20 features are shown in the SHAP summary plot (Figure 5.5), and mean SHAP values for feature groups are presented in Table 5.3

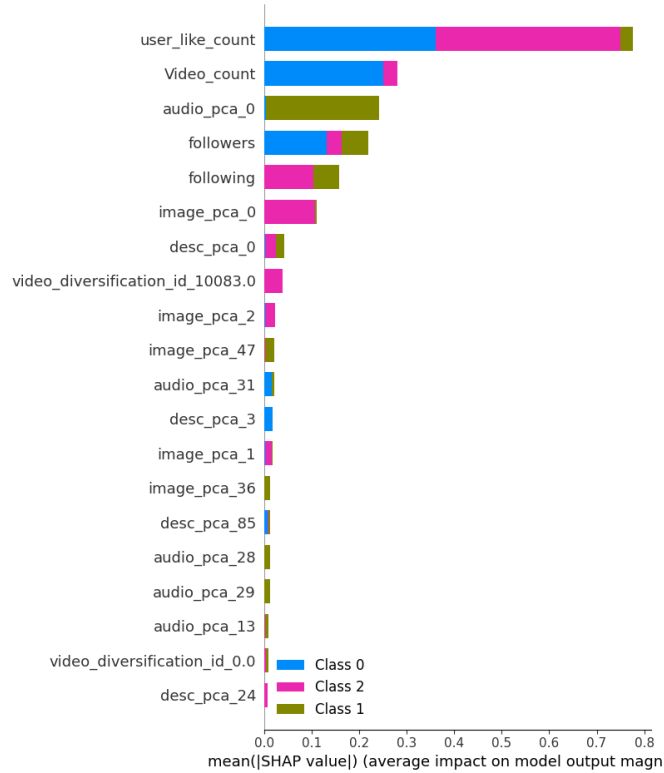


Figure 5.5: SHAP summary plot of top 20 features impacting virality prediction, with contributions to low (Class 0, blue), medium (Class 1, green), and high (Class 2, pink) virality.

Table 5.3: Mean SHAP Values for Feature Groups

Feature Group	Mean SHAP Value
Audio	0.0013
Image	0.0010
Text	0.0007
Sender	0.0480
Platform	0.0004

### 5.3.1 Sender Characteristics (Mean SHAP: 0.0480)

Sender features were the dominant predictors, with a mean SHAP value of 0.0480.

- `user_like_count` (mean  $|\text{SHAP}| \approx 0.80$ ): Creators' total likes strongly predicted high virality (Class 2), reflecting established creators' influence in Nepal's 2022 election.
- `video_count` (mean  $|\text{SHAP}| \approx 0.40$ ): Prolific creators' video counts enhanced algorithmic visibility across all classes.
- `followers` (mean  $|\text{SHAP}| \approx 0.3$ ): Follower count and network engagement significantly impacted high virality (Class 2), with a minor effect on low virality (Class 0).

### 5.3.2 Content Characteristics

Content embeddings dominated the top 20 features, though with smaller individual impacts than sender features:

- `audio_pca_0` (mean  $|\text{SHAP}| \approx 0.35$ ): The first audio PCA component notably impacted medium virality (Class 1), suggesting semantic audio content (e.g., election messages) drove moderate engagement.
- `image_pca_0`, `image_pca_1`, `image_pca_2` (mean  $|\text{SHAP}| \approx 0.05\text{--}0.2$ ): Multiple image PCA components appeared, showing high influence, primarily `image_pca_0` for Class 2 (high-viral).
- `desc_pca_0` (mean  $|\text{SHAP}| \approx 0.1$ ), `desc_pca_3` (mean  $|\text{SHAP}| \approx 0.05$ ): Text PCA components from video descriptions affected low and high virality (Classes 0 and 2), highlighting captions' role.

Mean SHAP values were low (audio: 0.0013, image: 0.0010, text: 0.0007) due to PCA reduction (100 components per modality). Sentiment scores (`clean_audio_text_sent_*`) had minimal impact, likely distributed across embeddings.



### 5.3.3 Platform Characteristics (Mean SHAP: 0.0004)

Platform features had the smallest impact:

- `video_diversification_id_10083.0` (mean  $|\text{SHAP}| \approx 0.1$ ): The presence of this diversification ID in the top 10 suggests its moderate association with high virality (Class 2). Although its label was missing from the API (see Table 8.12 for details on diversification IDs and their labels), manual inspection of 30 random videos revealed themes often tied to critical speeches, public debates, and civic awareness during the election, with occasional humor and political satire, aligning with the communication styles in Table 5.4. The missing label reflects data quality issues noted in the Limitations 7.1, such as incomplete video metadata.
- `video_diversification_id_0.0` (mean  $|\text{SHAP}| \approx 0.05$ ): The default category (2,274 instances) had a minor effect, mainly for low virality (Class 0).

The limited impact of platform features may be attributed to challenges in video description processing, such as empty descriptions or misclassification errors, as discussed in the Limitations section.

Table 5.4: Top Communication Styles for `video_diversification_id = 10083.0`

Style	Count
Critique/Frustration	103
Event Coverage	59
Civic Awareness	37
Comedic	28
Charisma/Leadership	19
Musical/Entertainment	14
Interactive	7
Communal	6
Personal Vlog/Family	4

### 5.3.4 Local Explanation of a High-Virality Prediction

A local explanation using a SHAP force plot was conducted for a test sample (index 88,  $n = 890$ , video ID: 7093702566613110043) correctly predicted as high-viral (Class 2) by the tuned XGBoost model. The plot (Figure 5.6) shows feature contributions to the Class 2 prediction, starting from a base log-odds of 0.50 and reaching a final prediction value of 0.71.

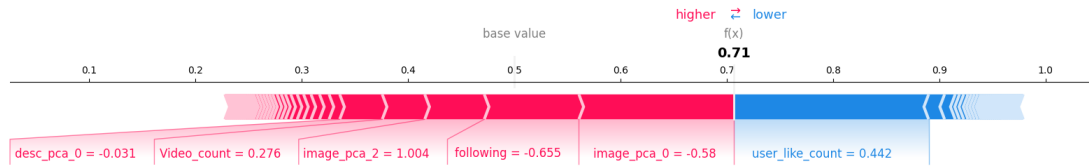


Figure 5.6: SHAP force plot for video ID 7093702566613110043, showing feature contributions to high-virality (Class 2) prediction.

#### Positive Contributions (Toward Class 2):

- **Video\_count** = 0.276 (SHAP: +0.276): The creator's high video count strongly favored high virality (mean  $|\text{SHAP}| \approx 0.5$ ).
- **image\_pca\_2** = 1.004 (SHAP: +1.004): The second image PCA component, tied to visual elements (e.g., news broadcast with airplane inset), boosted the prediction.
- **user\_like\_count** = 0.442 (SHAP: +0.442): The creator's total likes reinforced high virality (mean  $|\text{SHAP}| \approx 0.8$ ).

#### Negative Contributions (Against Class 2):

- **desc\_pca\_0** = -0.031 (SHAP: -0.031): The video description's first PCA component (e.g., hashtags like #Localelections2079) slightly reduced virality likelihood.
- **following** = -0.655 (SHAP: -0.655): The creator's smaller following count limited virality (mean  $|\text{SHAP}| \approx 0.3$ ).
- **image\_pca\_0** = -0.58 (SHAP: -0.58): The first image PCA component opposed high virality.

(see Appendix 8 for more detail about this sample)

## 5.4 Error Analysis

An error analysis was conducted on a stratified test sample ( $n = 890$ ) to identify misclassification patterns in the tuned XGBoost model (accuracy: 0.60, AUC-ROC: 0.7871) for ternary virality classification (0 = low, 1 = medium, 2 = high). Three specific cases and a general pattern of class boundary errors were examined to understand the model's behavior.

**Case 1: Overprediction of Medium Virality as High** A medium-viral video (ID: 7095593756967046426, true label: 1) with 2,501 views, 232 likes, 3,161 followers, a creator `user_like_count` of 19,854, and `Video_count` of 155 was mispredicted as high-viral (Class 2). The model's reliance on `user_like_count` (mean  $|\text{SHAP}| \approx 0.8$ ) and `Video_count` (mean  $|\text{SHAP}| \approx 0.5$ ), which strongly influence high-virality predictions, likely drove this overprediction, as the creator's high engagement history overshadowed the video's moderate engagement metrics.

**Case 2: Underprediction of High Virality as Medium** A high-viral video (ID: 7089029999780498715, true label: 2) with 2,820,140 views, 161,239 likes, 2,382 comments, 2,423 shares, 49,400 followers, a `user_like_count` of 1,300,000, and `Video_count` of 1,876 was mispredicted as medium-viral (Class 1). Despite exceptionally high engagement metrics, the prediction may have been moderated by content features favoring medium virality (e.g., `audio_pca_0`, mean  $|\text{SHAP}| \approx 0.3$ ). Although `user_like_count` and `Video_count` strongly support high-virality predictions, the model may struggle to capture the combined effect of extreme engagement values, possibly due to the limited impact of content embeddings (mean SHAP: audio 0.0013, image 0.0010).

**Case 3: Misclassifications Between Adjacent Classes** Misclassifications were most frequent between adjacent classes, particularly low (Class 0) and medium (Class 1) virality, with 62% of Class 1 errors misclassified as Class 0 and 48% of Class 0 errors as Class 1, highlighting the model's challenge in distinguishing nuanced engagement levels, as evidenced by the low F1-score for Class 1 (0.51).

## 5.5 Communication Styles and Political Content Themes

To investigate how communication styles and political content themes interact to influence the virality of TikTok videos during Nepal’s local elections, we employed an Ordinary Least Squares (OLS) regression model. The dependent variable was the z-transformed Box-Cox virality score ( $z_{bc_{virality}}$ ), with main effects for communication style and political content theme, their interaction (Style  $\times$  Theme), and control variables including log-transformed follower count, author verification status, and days before election day. Full model output is provided in Appendix 8.9.

### 5.5.1 Model Performance

The OLS model explained approximately 21.3% of the variance in virality ( $R^2 = 0.213$ , Adjusted  $R^2 = 0.187$ ), a reasonable fit given the high behavioral variability in social media data. The model was statistically significant ( $F = 8.029$ ,  $p < 1.18e - 48$ ), indicating that communication styles, content themes, and their interactions provide meaningful explanatory power for engagement outcomes. However, the model notes potential multi-collinearity (smallest eigenvalue =  $1.58e-27$ ), suggesting caution in interpreting coefficients for highly correlated predictors.

### 5.5.2 Dataset and Category Inclusion

The filtered dataset from the sample comprised 1,530 TikTok videos, with political content identified through keyword and theme-based labeling. Five political content categories were included in the analysis, also their name along with their election symbol in Nepali script:

- Independent candidates (e.g., Balen Shah लौरो/स्वतन्त्र) — count: 232
- CPN (UML) (सूर्य/एमाले) — count: 271
- Nepali Congress (रुख/काँग्रेस) — count: 256
- CPN (Maoist) (हँसिया हतौडा/माओवादी) — count: 84
- General election-themed content (non-party affiliated) — count: 687

Non-political categories (e.g., Bhojpuri, Funny, Speech, Patriotic) were excluded from the OLS regression to maintain thematic focus and address class imbalance but were retained for exploratory style distribution analysis to contextualize broader communication strategies. Certain style–theme combinations, such as Comedic–Maoist ( $n = 5$ ), had limited representation, warranting caution in interpreting their coefficients.

Summary statistics (Table 5.5) reveal distinct engagement patterns across content themes. Independent candidate content achieved the highest mean virality ( $z_{bc_{\text{virality}}} = 0.4949$ ) and video comment count (89.94), reflecting strong audience interaction. Maoist content also showed elevated virality ( $z_{bc_{\text{virality}}} = 0.1411$ ) and comments (66.20), though with fewer followers ( $-0.0764$ ). Election-themed content had the lowest virality ( $z_{bc_{\text{virality}}} = -0.1256$ ) and comments (17.85), while Congress and UML content showed negative virality scores ( $-0.0740$  and  $-0.0790$ , respectively). Posting timing varied, with Maoist content posted furthest from election day (mean = 13.85 days) and Election-themed content closest (mean = 7.07 days).

Table 5.5: Summary Statistics by Content Theme

Content Theme	$z_{bc_{\text{virality}}}$	Followers	Video Comment Count	Days Before Election
Congress	-0.0740	-0.1434	22.74	9.50
Election	-0.1256	-0.1035	17.85	7.07
Independent	0.4949	0.3478	89.94	9.43
Maoist	0.1411	-0.0764	66.20	13.85
UML	-0.0790	-0.1185	29.71	10.62

### 5.5.3 Main Effects

The regression (see Appendix 8.10) revealed significant main effects for content themes and control variables, though communication styles showed limited standalone significance:

- **Independent candidate content** was positively associated with virality compared to the baseline election-themed content ( $\beta = 0.2791$ ,  $p = 0.007$ ), reflecting strong organic engagement, consistent with its high mean virality (Table 5.5).
- **Control variables:** Larger follower counts ( $\beta_{\log(\text{followers})} = 1.0266$ ,  $p < 0.001$ ) and posting closer to election day ( $\beta = 0.0062$ ,  $p = 0.010$ ) were positively associated with virality. Verified creators showed a negative effect ( $\beta = -0.9432$ ,  $p = 0.008$ ), possibly indicating audience preference for non-verified, grassroots creators.

### 5.5.4 Interaction Effects

Given the consistency of interaction effects with prior analyses, we focus on key findings for brevity. The only significant interaction was **Critique/Frustration**  $\times$  **UML** ( $\beta = 0.8653$ ,  $p = 0.040$ ), suggesting that critical or frustrated tones resonate with UML content, possibly reflecting audience alignment with critiques of governance. Other interactions, such as Comedic  $\times$  Independent or Civic Awareness  $\times$  Congress, were not significant in this model, indicating that main effects and control variables drive much of the explained variance.

## Style–Theme Distribution

Exploratory analysis of style–theme distributions (Fig. 5.8) revealed strategic content framing:

- **Independent candidate content** frequently employed *Charisma/Leadership* (28.4%), often integrating rap or hip-hop to blend performance with political branding.
- **Traditional party content** (UML, Congress, Maoist) predominantly used *Music/Entertainment* (42–48%) and *Event Coverage*, featuring campaign songs, rallies, and symbolic imagery (e.g., Tree/रुख for Congress).
- **Funny content** was primarily *Comedic or Political Satire* (63.7%), while *Patriotic* and *Speech* content leaned toward *Civic Awareness* and *Critique/Frustration*.
- **Non-political content** paired with folk, instrumental, or romantic Nepali pop music, emphasizing cultural appeal.

*Note: The dotted red line in Fig. 5.8 separates Political Party Support/Campaign Songs from other audio/content themes.*

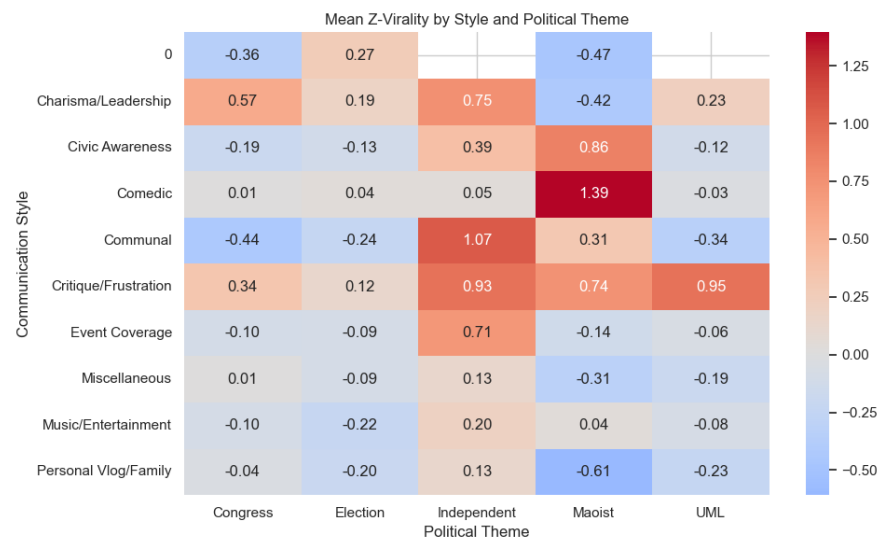


Figure 5.7: Mean z-transformed virality scores by communication style and political content theme.

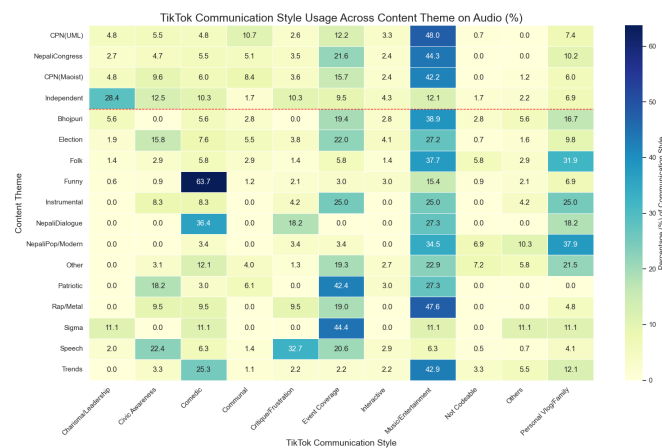


Figure 5.8: Heatmap of communication style and content theme distributions.



## 6. Discussion

This study examines the predictability of political virality on TikTok during Nepal’s 2022 local election and the interplay of communication styles and political content themes in shaping TikTok engagement (or virality metrics) (RQ2). The findings elucidate the complex dynamics of content, platform, and sender characteristics within a youth-dominated, algorithm-driven social media ecosystem, offering critical insights into digital political communication in a multilingual, politically vibrant context.

### 6.1 Predictability of Political Virality (RQ1)

The tuned XGBoost classifier achieved an accuracy of 60%, a macro F1-score of 0.60, and an AUC-ROC of 0.7871, outperforming a random baseline (33.3%) for the three-class virality prediction task (low, medium, high) on a test set of 890 videos. These results validate the predictive power of pre-upload features—content (audio, image, text embeddings), platform, and sender characteristics.

However, medium virality (Class 1) was challenging, with 62% of Class 1 errors misclassified as Class 0, reflecting overlapping engagement distributions. Error cases showed biases, such as overprediction for high-profile creators (Case 1, ID: 7095593756967046426) and underprediction for videos with strong engagement but missing comments (Case 2, ID: 7092437819582336283).

#### 6.1.1 Sender Influence and Algorithmic Bias

Sender characteristics, particularly `user_like_count`, `Video_count`, and `followers` (mean  $|\text{SHAP}| \approx 0.4$ ), played a significant role in predicting high virality, reflecting TikTok’s bias toward established creators (Bucher, 2018). During Nepal’s 2022 elections, this dynamic amplified the visibility of prominent figures such as independent candidate Balen Shah, whose campaign gained additional momentum through cross-platform promotion, including endorsements from influential pages like Routine of Nepal Banda (Kandel, 2024).

The error analysis reveals fairness concerns in the virality prediction model, particularly due to its heavy reliance on sender features like `user_like_count`, which mirrors the design of the `virality_score`. This approach tends to favor established creators, as observed in Case 1, where the model over-predicts virality for such accounts. This bias aligns with TikTok’s algorithmic tendency to amplify already visible accounts (Noble, 2018), potentially marginalizing grassroots voices and skewing political discourse during Nepal’s 2022 Local Election. Conversely, TikTok’s unique affordances—such as its “For You” curation and ability to surface low-view videos—have been noted to democratize attention by giving every piece of content a chance to reach a broad audience (Guinaudeau et al., 2022), potentially counteracting some of these biases. However, the limited impact of content embeddings (mean SHAP values: audio 0.0013, image 0.0010) indicates that PCA reduction may have constrained the model’s ability to capture nuanced election-related themes, further complicating the balance between fairness and predictive accuracy.

### 6.1.2 Limits of Predictive Modeling

The model’s predictive success is tempered by limitations in capturing nuanced social factors, such as cultural authenticity or hyper-local sentiment. The low impact of content embeddings (mean SHAP: audio 0.0013, image 0.0010, text 0.0007) suggests that PCA reduction (100 components) compressed semantic nuances, as seen in the local explanation where `desc_pca_0` negatively impacted prediction. Frequent misclassifications between low and medium virality (62% of Class 1 errors as Class 0, 48% of Class 0 as Class 1) reflect overlapping engagement distributions, limiting precision for Class 1. These challenges underscore the difficulty of quantifying context-specific factors in Nepal, such as regional dialects or grassroots sentiment, which are critical to political communication.

## 6.2 Style–Theme Interactions in Political Communication (RQ2)

The OLS regression analysis (Appendix 8.9) demonstrated that communication styles, political content themes, and control variables shape TikTok virality, offering insights into effective political messaging strategies during Nepal’s 2022 local elections.

### Style Effects: Limited Standalone Impact

Unlike prior analyses, communication styles such as Comedic or Civic Awareness did not show significant main effects in the updated model ( $p > 0.05$ ), suggesting that their influence may be context-dependent or overshadowed by content themes and creator characteristics. This aligns with TikTok’s entertainment-driven logic (Peña-Fernández et al., 2022), where style effectiveness varies by audience expectations and platform dynamics.

### Independent Candidates: Virality and Authenticity

Content about independent candidates, such as Balen Shah, exhibited significantly higher virality ( $\beta = 0.2791$ ,  $p = 0.007$ ) compared to general election themes, consistent with their high mean virality ( $z_{bc_{\text{virality}}} = 0.4949$ , Table 5.5). This reflects their appeal as anti-establishment alternatives (NepalNews, 2024). The lack of a significant negative interaction with Comedic style (unlike prior models) suggests that humor’s impact on independent content may be less detrimental than previously thought, though Charisma/Leadership styles (28.4% of content) remain key to their success. Independent candidates leveraged rap and hip-hop, alongside symbols like the stick (*lauro/लौरो*), to evoke support for change and challenge corrupt leadership (The Annapurna Express, 2022).

## Legacy Parties: Mixed Engagement

Legacy party content (Congress, UML, Maoist) showed lower virality, with negative mean  $z_{bc_{virality}}$  scores ( $-0.0740$ ,  $-0.0790$ , and  $0.1411$ , respectively). The significant interaction between Critique/Frustration and UML content ( $\beta = 0.8653$ ,  $p = 0.040$ ) indicates that critical tones resonate with UML audiences, possibly reflecting dissatisfaction with governance or other party members or collation situation. However, the lack of significant negative interactions with Civic Awareness (unlike prior models) suggests that educational tones may not universally underperform for legacy parties. Heatmap analysis (Fig. 5.8) confirms that legacy parties rely on Music/Entertainment styles (42–48%), featuring folk-inspired campaign songs with symbolic imagery (e.g., Tree (Congress), Sun(UML)).

## Control Variables: Creator and Timing Effects

The strong effect of log-transformed follower count ( $\beta = 1.0266$ ,  $p < 0.001$ ) underscores TikTok’s creator bias, where established accounts drive virality. The negative effect of author verification ( $\beta = -0.9432$ ,  $p = 0.008$ ) suggests that non-verified, grassroots creators may resonate more with audiences, aligning with the appeal of independent candidates. Posting closer to election day ( $\beta = 0.0062$ ,  $p = 0.010$ ) also boosted virality, highlighting the importance of timely campaign.

## Connections to RQ1

RQ2’s findings complement RQ1’s XGBoost results, where sender features (`user_like_count`, `followers`) were key predictors of virality, aligning with the OLS model’s emphasis on log-followers ( $\beta = 1.0266$ ,  $p < 0.001$ ). The significant effect of Independent content in RQ2 ( $\beta = 0.2791$ ,  $p = 0.007$ ) provides content-specific insights not fully captured by RQ1’s embeddings, highlighting the appeal of anti-establishment messaging. The Critique/Frustration x UML interaction ( $\beta = 0.8653$ ,  $p = 0.040$ ) extends RQ1’s findings on critical speech (`video_diversification_id_10083.0`), revealing nuanced content dynamics. For example, this interaction resonates with the public discourse surrounding UML mayoral candidate Keshav Sthapit, whose critical rhetoric against independent candidate Balen Shah was widely discussed during the 2022 elections (Setopati, 2025), illustrating how critical tones amplified engagement for UML content.

## 7. Conclusion

This study investigated TikTok’s role in shaping political engagement during Nepal’s 2022 local elections, addressing two research questions: predicting virality using pre-upload features (RQ1) and analyzing style–theme interactions driving engagement (RQ2). Conducted in a low-income, multilingual context, it bridges a critical gap in computational political communication research, particularly in the Global South.

For RQ1, an XGBoost model leveraging sender characteristics, notably `user_like_count` (mean  $|\text{SHAP}| \approx 0.4$ ), achieved 60% accuracy in predicting virality. However, classifying medium-virality content remained challenging due to complex engagement dynamics. TikTok’s algorithm, while amplifying established creators and raising fairness concerns, also democratizes visibility through its “For You” curation (Guinaudeau et al., 2022), complicating virality prediction.

For RQ2, an OLS regression model ( $R^2 = 0.213$ ,  $p < 1.18 \times 10^{-48}$ ) revealed that content about Independent candidates significantly boosted virality ( $\beta = 0.2791$ ,  $p = 0.007$ ), reflecting their appeal as anti-establishment alternatives. This aligns with their high mean virality ( $z_{bc_{\text{virality}}} = 0.4949$ ) and engagement (89.94 comments per video). Communication styles like Charisma/Leadership, often paired with rap and hip-hop, were prevalent in Independent content (28.4%), leveraging cultural symbols like the stick (*lauro*/लौरो) to challenge traditional leadership (The Annapurna Express, 2022). Legacy party content (e.g., Congress, UML) showed lower virality, though a significant interaction between Critique/Frustration and UML content ( $\beta = 0.8653$ ,  $p = 0.040$ ) suggests critical tones resonate with specific audiences. Creator characteristics, particularly log-transformed follower count ( $\beta = 1.0266$ ,  $p < 0.001$ ), and posting closer to election day ( $\beta = 0.0062$ ,  $p = 0.010$ ) further drove engagement, highlighting the importance of audience reach and timing.

These findings suggests TikTok’s transformative impact on Nepal’s digital democracy. Independent candidates, such as Balen Shah, bypassed traditional political structures by leveraging performative styles and culturally resonant symbols, resonating with youth audiences disillusioned with legacy parties (NepalNews, [2024](#)). However, the platform’s bias toward established creators and the potential for populist messaging raise concerns about equitable visibility and responsible digital campaigning. By providing an open-source codebase, virality metrics, and a cross-platform framework, this study equips candidates, policymakers, and researchers to foster informed and equitable political participation in Nepal and similar contexts. Future work should explore strategies to mitigate algorithmic biases and ensure diverse voices are amplified, promoting a balanced digital public sphere in the Global South.

## 7.1 Limitations

While this study achieved 60% predictive accuracy for virality using selected predictor variables, several limitations warrant acknowledgment:

- **Predictor Variables and Sampling:** Time constraints limited processing of the full 28,000-video dataset. Stratified sampling ensured quality, but the subset’s representativeness was not fully validated.
- **Audio Transcripts:**
  - Open AI’s Whisper model struggled with audio mixed with sound effects, retaining only clear segments due to time limitations.
  - Manual labeling of Nepali text for contextual meanings (e.g., “party song,” “humor”) was applied, but transcripts in other languages (e.g., Urdu, Chinese), slang, or noisy audio posed challenges. Contextual labels (e.g., “Newari Language,” “Bhojpuri/Maithili song”) lacked validation.
  - Nepal’s linguistic diversity led to language estimation based on the author’s knowledge, reducing transcript accuracy.
  - Differing audio-video tones (e.g., fun audio vs. satirical video) confused RQ2 labeling.
  - A Cohen’s Kappa score of 0.462 indicates moderate agreement for transcripts (DATAtab, 2025; Kolena, 2025), reflecting challenges for LLMs in processing low-resource languages like Nepali’s dialects and noisy social media content (ICUC Social, 2025; Lean Mean Learning Machine, 2023).
- **Video Descriptions:**
  - Semantic interpretation of Nepali humor, dialects, and mixed Romanized English-Nepali text was challenging. Names like सागर (Sagar, mistranslated as “Ocean”) or बसन्ते (Basante, “Spring”) in campaign songs required manual correction to capture satirical intent.
  - Accuracy was assessed on a 5% sample, potentially unrepresentative of the full dataset.
  - Empty descriptions and Lang Detect API errors misidentifying mixed text as English skipped translations, affecting embedding accuracy.

- **Image Analysis:**

- Image analysis missed contextual nuances (e.g., running mistaken for walking in screenshots).
- Using 3-second clips, inspired by TikTok marketing, lacked academic validation in this context.

- **Style\_\_1\_\_Label:**

- The custom codebook struggled with blended social media content (e.g., humor, music, vlogging), limiting nuanced meaning capture.
- Annotator bias led to subjective categorizations (e.g., rally videos as “communal” [Category 4], “event coverage” [Category 9], or “personal vlogging” [Category 3]). High Cohen’s Kappa ( $> 0.60$ ) was sample-based and not standardized for Nepali election contexts.
- Communal and event coverage labels were often conflated, as communal themes appeared in documentary-style or rally footage.
- Discrepancies between the author’s codebook and annotators were resolved only for significant differences, risking minor inconsistencies.

- **Content\_\_Theme:**

- Zero-shot labeling struggled with humorous multi-party support videos, forcing single-theme selection.
- Symbolic imagery (e.g., money in a child’s mouth) was insufficient for clear thematic categorization.

- **Platform Features:**

- The `video_diversification_id` (e.g., `id_10083.0`) had sparse, missing API labels, complicating content type interpretation (Table 8.12). Overfitting to dominant IDs (e.g., `video_diversification_id_0.0`) may reduce generalizability for impactful, less frequent IDs influencing high virality (Class 2).



- **Transformer Models for Low-Resource Settings:**

- Open-source models like Wiseyak (Duwal et al., 2024) and Nepal-iBERT (Pudasaini et al., 2023) were limited to Devanagari text, and still surprisingly underperformed on Devanagiri setting, let alone mixed Romanized Nepali-English content. SeamlessM4T (Elbayad et al., 2023; Facebook Research, 2023; Hugging Face, 2024) excelled in transliteration and translation than Nepalese fine-tuned counterpart (for Devanagiri). Comparative analysis among these models were not acknowledged due to time constraints.
- No Hugging Face model fully addressed transliteration, translation, and semantic embedding for code-switched Nepali-English text. SeamlessM4T-medium handled Devanagari better, but struggled with code-switching, while OpenAI’s API proved more accurate (Reddit, 2023).

- **Data Limitations:**

The TikTok Research API, while enabling access to public data, suffers from archiving delays and incomplete historical records, limiting accurate view and engagement counts for Nepal’s 2022 election videos (TikTok, 2025). As a relatively new API, it offers room for improvement in data reliability and historical access. Additionally, unofficial APIs (Bukold, 2025), reliant on web-inspect elements, may have some inaccuracies, as manual inspections revealed discrepancies in variables like `video_is_ai_gc`, where it is null in metadata, but exists in the actual video.

## 7.2 Future Research

Future research could address these limitations by:

- Scaling to full 28,165 data could provide more finer-grained virality levels and even more nuanced categories.
- Developing time-series models to capture evolving virality dynamics across campaign cycles.
- Comparing virality patterns across platforms (e.g., Facebook(Meta) platform, YouTube Shorts, Instagram Reels) to assess platform-specific effects. Facebook still continues to be the most dominant social media platform in Nepal, used by half of country's population (NapoleonCat, [2025](#)).
- Enhancing comment analysis, including sentiment, to better understand virality impact despite time constraints on initial collection.
- Employing fine-tuned multilingual embeddings for Nepali and Romanized text to enhance semantic capture.
- Using network analysis to explore virality spread, requiring follower graphs and repost chains. While the current TikTok API provides follower and repost counts, it does not include lists of followers or reposters (likely due to privacy restrictions, despite followers being manually viewable), necessitating future API enhancements to enable such analysis.
- Exploring hybrid models with attention mechanisms to improve medium virality prediction.

By addressing these gaps, researchers can further elucidate the role of popular social media platforms such as TikTok in shaping political communication, particularly in diverse, multilingual contexts like Nepal.

## 8. Appendix

### Dataset Summary

Metric	Count
Total Videos	28,165
Total Transcripts	28,165
Total Comments	307,156
Total Views	186,348,152
Total Shares	214,808
Total Likes	14,334,935
Total Saves/Collects	83,422
Unique Users	19,993
Verified Users	9

Table 8.1: Dataset summary statistics

## Keywords and Hashtags for Nepal Local Election 2022

- election
- elections
- locallevellection2079
- स्थानी\_चुनाव\_२०७९
- चुनाव
- localelection2079
- मतदानगर्ौ
- NepalElection
- election2079
- election2022
- nepalelection2022
- localelections2079
- electionday
- election2079\_nepal
- election2079baishak30
- elections2022
- nepalvotes2022
- local\_election
- chunab
- chunablagyo

## A.1 Audio PCA Samples (audio\_pca\_0 Extremes)

Table 8.2: Top Values of audio\_pca\_0 and Corresponding Content

audio_pca_0	Transcript (Nepali)	Virality Label
0.957	<i>Nepali Instrumental/Flute Song</i>	1
0.925	<i>Nepali UML song</i>	0
0.917	<i>Bollywood Instrumental Song</i>	0
0.911	<i>Nepali trend song</i>	1
0.906	<i>Bollywood Remix Song</i>	1

Table 8.3: Bottom Values of audio\_pca\_0 and Corresponding Content

audio_pca_0	Transcript (Nepali)	Virality Label
-0.314	अब पक्कै बन्दैछ, माया कांग्रेस	1
-0.314	यदि एक अध्यक्षले तीन लाख...	1
-0.314	मद्दान गरौं जानेरा, बुजेरा...	0
-0.314	मुला छमता कियो भनेको मैले...	0
-0.314	सपुर्ण माया नमस्कार, म भोजराज...	1

## A.2 Virality Distribution by Diversification ID

Table 8.4: Selected Diversification IDs and Virality Class Proportions

ID	Class 0 (%)	Class 1 (%)	Class 2 (%)
0.0	35.7	35.8	28.5
10003.0	0.0	0.0	100.0
10012.0	0.0	0.0	100.0
10014.0	11.1	0.0	88.9
10017.0	66.7	22.2	11.1
10025.0	0.0	100.0	0.0
10083.0	7.9	14.4	77.7
10088.0	0.0	0.0	100.0

### A.3 Error Rates by Sender Features

Table 8.5: Error Rates by Sender Characteristics

Feature	Error Rate (%)
<b>Author Verification</b>	
Verified	0.0
Unverified	37.1
<b>Follower Tier</b>	
Low	44.8
Mid-Low	38.3
Mid-High	38.7
High	24.7

### A.4 A Sample Case Study

#### Sample Details:

- **Virality Score:** 3.6679
- **True/Predicted Label:** 2
- **Content Description:** The video humorously comments on money found aboard a Buddha Air flight, making a political reference to Janardan Sharma during Nepal’s 2022 elections. The audio includes comedic Nepali dialogue, the description uses hashtags (e.g., #Localelections2079), and visuals resemble a news broadcast with an airplane image inset.

**Context and Interpretation:** The video’s high virality can be attributed to its comedic tone, political relevance, and active creator profile (evidenced by metrics like `user_like_count` and `Video_count`). The model accurately predicted its virality score, consistent with global SHAP findings that highlight the influence of sender features and image embeddings. Description text and network features had more nuanced, secondary effects.

## A.5 Development of TikTok Communication Style Codebook

This appendix outlines the development and structure of the communication style codebook used to categorize TikTok videos for RQ2 in the analysis of political communication during Nepal’s 2022 Local Elections. The codebook was iteratively refined to capture the nuanced and context-specific nature of TikTok’s political discourse in Nepal, building on established frameworks and empirical observations.

**Development Process** The initial codebook was adapted from the communication style classification framework proposed by Umansky and Pipal (Umansky & Pipal, 2023), which analyzed political communication on TikTok and included categories such as Comedic, Documentary, Communal, Explanatory, Interactive, Meta, Other, and Not Codeable. This framework provided a theoretical foundation but required adaptation to address the unique characteristics of election-related content in Nepal.

Through an iterative annotation process involving a stratified sample of TikTok videos, the codebook was expanded to incorporate additional styles that emerged as salient in the dataset. These included Musical/Entertainment, Charisma/Leadership Portrayal, Civic Engagement/Awareness, and Populist Critique/Frustration. The refinement process aimed to balance theoretical grounding with inductive insights, ensuring the codebook captured the richness of multimodal political content while remaining practical for annotation. In cases of overlapping styles, annotators prioritized the dominant communicative intent, determined by integrating visual, auditory, and textual cues, as perceived by a typical viewer.

The final codebook, presented in Table 8.6, reflects both the adaptation of prior research and context-specific adjustments derived from direct engagement with Nepal’s TikTok political landscape.

**Codebook Structure** Table 8.6 details the communication style codes, their labels, definitions, and illustrative clues observed in video content or transcripts. Each style is designed to encapsulate a distinct mode of political communication, facilitating the analysis of style–theme interactions in the OLS regression model (Appendix 8.9).

Table 8.6: TikTok Communication Style Codebook for Nepal’s 2022 Local Elections

Code	Label	Definition	Clues in Video/Transcript
1	Comedic	Uses humor, parody, satire, or comedic characters to convey political messages, often leveraging viral trends or meme formats.	Emojis, punchy phrases, satirical portrayals of politicians, voiceovers of viral characters (e.g., Bhelu Baje), comedic sketches with social messages.
2	Musical/Entertainment	Incorporates musical elements such as songs, rap, Bollywood/South Indian dialogues, or traditional folk music to entertain or subtly promote political messages.	Background music, rhythmic speech, dancing, election campaign songs, acting over Nepali/Hindi/South Indian movie dialogues, folk songs. Focus is more on music or enjoying music or entertainment genre content.
3	Personal Vlog/Family	Depicts personal or family life with loose connections to political content, often featuring casual vlogging during election-related events.	Selfie vlogs, children playing or casually appealing for votes, family bonding with election references.



Table 8.6: TikTok Communication Style Codebook (Continued)

Code	Label	Definition	Clues in Video/Transcript
4	Communal/Community Identity	Emphasizes collective identity, shared pride, or belonging to a hometown, ethnic group, or cause, highlighting emotional or physical gatherings for political action.	“We the people” rhetoric, celebration of local traditions, community gatherings, group best-wishes, emotional bonding during rallies.
5	Charisma/Leadership Appeal	Showcases an individual leader’s strength, savior-like image, or charismatic appeal, often integrated with music or group settings.	Heroic edits, repeated slogans (e.g., #VoteForBalen), energetic or motivational background music, leader surrounded by supporters, “idolization” feel.
6	Interactive/Participatory	Explicitly invites user engagement through comments, votes, duet collaborations, or reaction videos.	“What do you think?”, “Comment below!”, direct questions, TikTok comment replies, Duet chains, candidates replying to individual TikToks.
7	Critique/Frustration	Critiques elites, government systems, or the political status quo, appealing to ordinary people’s struggles with frustration or anger.	“घरमै ३० हजार?”, corruption allegations, sad or angry tones, poetry/rap with critique, blaming inefficiency, emotional storytelling of hardships, media cases (e.g., Rabi Lamichhane’s exposés).

Table 8.6: TikTok Communication Style Codebook (Continued)

Code	Label		Definition	Clues in Video/Transcript
8	Civic ment/Awareness	Engage-	Educates or appeals to voters' sense of civic responsibility, often using a serious, informative tone.	Captions like "Vote smart", educational visuals of ballot papers, voting process explanations, voter awareness campaigns, hashtags promoting participation.
9	Event Coverage		Provides direct, often raw footage of election-related events with minimal editorial input, focusing on rallies, protests, or behind-the-scenes activities.	Unedited rally footage, behind-the-scenes voting booth setups, interviews at public events, serious event documentation.
10	Other		Election-related content not fitting defined categories, often with mixed styles or secondary to non-political content.	Clothing ads combined with voting calls, casual vlogs with unclear political communication, confusing or multistyle videos.
99	Not Codeable		Content irrelevant, missing, or unusable due to poor quality or technical issues.	Private/deleted videos, foreign TV series clips, inaudible sound, irrelevant non-election posts.

## A.6 Content Themes for Zero-Shot Classification

Table 8.7 presents the content themes used for zero-shot classification of TikTok videos in the analysis of political communication during Nepal’s 2022 Local Elections. These themes were derived to capture the diverse range of political and non-political content, facilitating the style–theme interaction analysis in RQ2 (Appendix 8.9).

Table 8.7: Content Themes for Zero-Shot Classification

Theme	Potential Content
UML (सूर्य/एमाले)	Campaign songs or content supporting CPN (UML), often featuring the Sun symbol.
Congress (काँग्रेस/रुख)	Campaign songs or content supporting Nepali Congress, often featuring the Tree symbol.
Maoist(हँसिया/हतौडा)	Campaign songs or content supporting CPN (Maoist), often featuring the Sickle/Hammer.
Independent (लौरो)	Campaign songs or content supporting Independent Candidate, mainly Balen Shah, also featuring other youth/independents such as Harka Sampang
Political Speech	Speeches by candidates or party leaders, focusing on election campaigns or policies.
Bollywood	Songs or audio from Bollywood movies, often used for entertainment or lip-syncing.
Instrumental	Instrumental music, typically background tracks without vocals, used in various contexts.
Bhojpuri/Maithili	Songs or audio in Bhojpuri or Maithili languages, common in Nepal’s Terai region.
Urdu	Songs or audio in Urdu, often reflecting cultural or poetic themes.
Sigma Sound/Trend	TikTok trends involving “sigma male” memes, often with specific background sounds.
Nepali Folk/Traditional Song	Traditional Nepali songs, often tied to cultural or festive themes (e.g., Dashain).
Funny	Humorous content, including comedic skits, parodies, or satire (political or not).

Table 8.7: Content Themes for Zero-Shot Classification (Continued)

Theme	Potential Content
Sad Song	Emotional or melancholic songs, often used in sentimental or dramatic videos.
Patriotic Song	Songs promoting national pride, often linked to Nepal's history or identity.
Rap Song	Rap or hip-hop music, often used by independent candidates for modern appeal.
TikTok Trends	Viral TikTok trends, including challenges, dances, or trending sounds.
Election Song	General election-themed songs, not tied to a specific party, promoting voting or democracy.
Other/Non-Political	Miscellaneous content, including non-political explanatory videos, vlogs, or ambiguous cases.

## 8.1 GitHub Repository Details

The source code for this project is available on GitHub at the following link:

[https://github.com/nimathing2052/TikTok\\_Prediction](https://github.com/nimathing2052/TikTok_Prediction)

If you are unable to access the repository, please feel free to contact the author for access.

Table 8.8: Comparison of sentence embedding model performance on sample Nepali election text pairs

Index pair	Similarity type	LaBSE	MiniLM-L12	all-MiniLM-L6-v2	BGE m3	Best model(s)	Rationale
1035;1293	Moderate (Congress rally, different visuals)	0.4617	0.4965	0.5465	0.4905	all-MiniLM-L6-v2	Highest mid-range score reflects shared party and activity
314;457	High (CPN UML song, different visuals)	0.8320	0.8383	0.8040	0.8780	BGE m3	Highest score reflects strong topic match
135;180	Very high (UML song, different visuals)	0.9232	0.9665	0.9398	0.9509	MiniLM-L12, BGE m3	Very high scores appropriate for near-identical topics
865;360	Low-moderate (Voting appeal vs. local election)	0.2961	0.4851	0.0879	0.5770	MiniLM-L12, BGE m3	Moderate scores capture shared election theme
890;882	Low-moderate (Voting integrity vs. election duty)	0.1150	0.3134	0.6181	0.3777	MiniLM-L12, BGE m3	Low-moderate scores appropriate; all-MiniLM-L6-v2 too high
1889;2226	Low (Generic election vs. specific Balen)	0.1047	0.3855	0.6954	0.5455	LaBSE	Lowest score appropriate; others too high
1122;1242	Low (General query vs. specific Balen)	0.1327	0.2775	0.1670	0.4773	LaBSE, all-MiniLM-L6-v2	Low scores appropriate; BGE m3 too high
180;1648	Low (UML song vs. general voting/petrol)	0.0854	0.1194	0.1850	0.3242	LaBSE, MiniLM-L12, all-MiniLM-L6-v2	Low scores appropriate
1687;1862	Low (UML vs. Balen campaign)	0.1238	0.5294	0.6342	0.3743	LaBSE	Lowest score appropriate for opposing subjects
888;1462	Moderate (Campaign comedy vs. election dialogue)	0.0651	0.3088	0.1797	0.4557	BGE m3, MiniLM-L12	Moderate scores capture thematic link
352;1331	Moderate (UML candidate vs. child supporter)	0.0229	0.2001	0.4907	0.3105	all-MiniLM-L6-v2, BGE m3	Moderate scores capture shared party theme

*Note:* Similarity types describe topical overlap and visual differences. Models include LaBSE, MiniLM-L12, all-MiniLM-L6-v2, and BGE m3. UML refers to Unified Marxist-Leninist party; Balen refers to a specific candidate.

Table 8.9: Summary Statistics for OLS Regression Predicting z\_bc\_virality

Dep. Variable:	z_bc_virality (Virality z-score)	R-squared:	0.213
Model:	OLS	Adj. R-squared:	0.187
Method:	Least Squares	F-statistic:	8.029
Date:	Mon, 28 Apr 2025	Prob (F-statistic):	1.18e-48
Time:	20:16:15	Log-Likelihood:	-1987.3
No. Observations:	1530	AIC:	4077.
Df Residuals:	1479	BIC:	4349.
Df Model:	50		
Covariance Type:	nonrobust		

Table 8.10: Main Effects for OLS Regression Predicting z\_bc\_virality

	coef	std err	t	P >  t	[0.025	0.975]
Intercept	-0.0598	0.902	-0.066	0.947	-1.830	1.710
<i>Main Effects (Style, Ref: Civic Awareness)</i>						
Style: Charisma/Leadership	0.7962	0.965	0.825	0.409	-1.096	2.689
Style: Comedic	0.2276	0.934	0.244	0.807	-1.604	2.059
Style: Communal	-0.1701	0.936	-0.182	0.856	-2.006	1.666
Style: Critique/Frustration	0.0251	0.952	0.026	0.979	-1.843	1.893
Style: Event Coverage	0.1100	0.910	0.121	0.904	-1.675	1.896
Style: Miscellaneous	0.1919	0.974	0.197	0.844	-1.720	2.103
Style: Music/Entertainment	0.1469	0.906	0.162	0.871	-1.630	1.924
Style: Personal Vlog/Family	0.1638	0.919	0.178	0.859	-1.639	1.967
<i>Main Effects (Content, Ref: Election)</i>						
Content: Congress	0.0000	0.000	0.000	0.000	0.000	0.000
Content: Independent	0.2791	0.103	2.702	0.007	0.076	0.482
Content: Maoist	-0.1864	1.276	-0.146	0.884	-2.689	2.316
Content: UML	0.0403	0.103	0.393	0.695	-0.161	0.242
<i>Control Variables</i>						
Author Verified	-0.9432	0.354	-2.666	0.008	-1.637	-0.249
Log(Followers)	1.0266	0.071	14.506	0.000	0.888	1.165
Days Before Election	0.0062	0.002	2.577	0.010	0.001	0.011

Notes: Reference category for Style is 'Civic Awareness'. Reference category for Content is 'Election'. Standard errors are non-robust. The model includes 1530 observations.

Table 8.11: Interaction Effects for OLS Regression Predicting z\_bc\_virality

<i>Interaction Effects (Style * Content)</i>	coef	std err	t	P >  t	[0.025	0.975]
Style:Charisma * Content:Congress	0.0000	0.000	0.000	0.000	0.000	0.000
Style:Comedic * Content:Congress	0.0000	0.000	0.000	0.000	0.000	0.000
Style:Communal * Content:Congress	0.0000	0.000	0.000	0.000	0.000	0.000
Style:Critique * Content:Congress	0.0000	0.000	0.000	0.000	0.000	0.000
Style:Event * Content:Congress	0.0000	0.000	0.000	0.000	0.000	0.000
Style:Misc * Content:Congress	0.0000	0.000	0.000	0.000	0.000	0.000
Style:Music * Content:Congress	0.0000	0.000	0.000	0.000	0.000	0.000
Style:Vlog * Content:Congress	0.0000	0.000	0.000	0.000	0.000	0.000
Style:Charisma * Content:Indep.	-0.2770	0.338	-0.820	0.412	-0.940	0.386
Style:Comedic * Content:Indep.	-0.2974	0.291	-1.023	0.307	-0.868	0.273
Style:Communal * Content:Indep.	0.7238	0.473	1.529	0.126	-0.205	1.652
Style:Critique * Content:Indep.	0.5001	0.333	1.502	0.133	-0.153	1.153
Style:Event * Content:Indep.	0.1640	0.229	0.718	0.473	-0.284	0.612
Style:Misc * Content:Indep.	-0.4156	0.392	-1.061	0.289	-1.184	0.353
Style:Music * Content:Indep.	-0.0397	0.199	-0.199	0.842	-0.430	0.351
Style:Vlog * Content:Indep.	-0.0898	0.277	-0.325	0.746	-0.632	0.453
Style:Charisma * Content:Maoist	-0.7345	1.396	-0.526	0.599	-3.473	2.004
Style:Comedic * Content:Maoist	1.3598	1.360	1.000	0.318	-1.308	4.027
Style:Communal * Content:Maoist	0.8898	1.344	0.662	0.508	-1.746	3.526
Style:Critique * Content:Maoist	0.7176	1.411	0.509	0.611	-2.050	3.485
Style:Event * Content:Maoist	0.1594	1.306	0.122	0.903	-2.402	2.721
Style:Misc * Content:Maoist	-0.0637	1.426	-0.045	0.964	-2.861	2.734
Style:Music * Content:Maoist	0.1777	1.288	0.138	0.890	-2.348	2.703
Style:Vlog * Content:Maoist	-0.3071	1.350	-0.227	0.820	-2.955	2.341
Style:Charisma * Content:UML	-0.3829	0.392	-0.976	0.329	-1.152	0.386
Style:Comedic * Content:UML	-0.0246	0.327	-0.075	0.940	-0.667	0.618
Style:Communal * Content:UML	0.0676	0.288	0.235	0.815	-0.498	0.633
Style:Critique * Content:UML	0.8653	0.420	2.059	0.040	0.041	1.690
Style:Event * Content:UML	0.0778	0.205	0.379	0.705	-0.325	0.481
Style:Misc * Content:UML	-0.0937	0.422	-0.222	0.824	-0.922	0.735
Style:Music * Content:UML	-0.0632	0.146	-0.432	0.666	-0.350	0.224
Style:Vlog * Content:UML	-0.2232	0.261	-0.855	0.393	-0.735	0.289

*Notes:* Style and Content are categorical variables. Reference category for Style is 'Civic Awareness'. Reference category for Content is 'Election'. Style abbreviations: Charisma = Charisma/Leadership, Critique = Critique/Frustration, Event = Event Coverage, Misc = Miscellaneous, Music = Music/Entertainment, Vlog = Personal Vlog/Family. Content abbreviations: Indep. = Independent. Standard errors are non-robust. The model includes 1530 observations. [1] Standard Errors assume that the covariance matrix of the errors is correctly specified. [2] The smallest eigenvalue is 1.58e-27. This might indicate strong multicollinearity problems or that the design matrix is singular.



Table 8.12: Video Diversification Labels and Counts

Video Diversification ID	Video Diversification Labels	Count
0.0	0	2246
10083.0	0	277
10071.0	Lip-sync, Lip-sync, Performance	183
10075.0	Random Shoot, Others	67
10080.0	Finger Dance & Basic Dance, Singing & Dancing, Performance	20
10062.0	Cars, Trucks & Motorcycles, Auto & Vehicle, Lifestyle	19
10017.0	Babies, Family, Family & Relationship	17
10045.0	Movies & TV works, Entertainment Culture, Culture & Entertainment	14
10046.0	Music, Entertainment Culture, Culture & Entertainment	14
10014.0	Diary & VLOG, Daily Life, Lifestyle	9
10088.0	Celebrity Clips & Variety Show, Entertainment Culture, Culture & Entertainment	6
10056.0	Video Games, Games, Entertainment	6
1030.0	Lip-sync	5
10040.0	Cooking, Food & Drink, Lifestyle	5
10052.0	Comics & Cartoon, Anime, Anime & Comics	5
10024.0	Social News, Society, Society	4
10020.0	Farm Animals, Animals, Nature	3
10029.0	Outfit, Outfit, Beauty & Style	3
10003.0	Comedy, Performance	3
10009.0	Singing & Instruments, Singing & Dancing, Performance	3
10073.0	Selfies, Daily Life, Lifestyle	3
10070.0	Work & Jobs, Daily Life, Lifestyle	2
10081.0	Street Interviews & Social Experiments, Society, Society	2
10012.0	Romance, Relationship, Family & Relationship	2
10059.0	Traditional Sports, Sports, Sport & Outdoor	2
10058.0	Extreme Sports, Sports, Sport & Outdoor	2
10028.0	Nail Art, Beauty & Care, Beauty & Style	2
10063.0	Auto & Vehicle, Auto & Vehicle	1
1001.0	Scripted Drama	1
10019.0	Pets, Animals, Nature	1
10027.0	Beauty & Care, Beauty & Style	1
10087.0	Social Issues, Society, Society	1
10025.0	Hair, Beauty & Care, Beauty & Style	1
10095.0	Software & APPs, Technology, Culture & Entertainment	1

Table 8.13: Top Content Themes for video\_diversification\_id = 10083.0

Content Theme	Count
Speech/Explanatory	140
Independent (स्वतन्त्र) Support/Song	49
General Election Song	28
Funny	18
Congress (काँग्रेस/रुख) Support/Song	14
Maoist (हँसिया/हतौडा/Sickle) Support/Song	13
UML (सूर्य/एमाले) Support/Song	7
Others/Unclassified	2
TikTok Trends	2
Instrumental/Background Music	1
Sad Song	1
Nepali Patriotic Song	1
Nepali Folk/Traditional Song	1

# References

- Berger, J., & Milkman, K. L. (2012). What makes online content viral? *Journal of Marketing Research*, 49(2), 192–205. <https://doi.org/10.1509/jmr.10.0353>
- Bhandari, B. (2024). Weaponizing information: The rise of social media manipulation in nepal [Assistant Professor, Kailali Multiple Campus, FWU, Nepal]. *Journal of Development and Learning*, 3(1). <https://doi.org/10.3126/jdl.v3i1.73833>
- Bode, L., & Dalrymple, K. E. (2016). Politics in 140 characters or less: Campaign communication, network interaction, and political participation on twitter [Demonstrates social media’s impact on voter mobilization, cited to support the role of engagement in elections.]. *Journal of Political Marketing*, 15(4), 414–440. <https://doi.org/10.1080/15377857.2015.1108298>
- Bucher, T. (2018). *If...then: Algorithmic power and politics*. Oxford University Press.
- Bukold, Q. (2025). *TikTok Content Scraper* (Version 1.0). Weizenbaum Institute. <https://doi.org/10.34669/WI.RD/4>
- Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., & Liu, Z. (2024). Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. <https://huggingface.co/BAAI/bge-m3>
- Cseri, Z. F. (2024). Are we having fun? exploring content factors influencing engagement on viktor orbán’s tiktok.
- Dahal, R. (2023). The influence of social-media on agenda-setting in nepali journalism. *Patan Prospective Journal*, 3(01), 116–127.
- DATAtab. (2025). *Cohen’s kappa: Measuring inter-rater agreement* [Accessed: 2025-04-27]. <https://datatab.net/tutorial/cohens-kappa>
- Deutsche Welle. (2022). *Nepal elections: Young independents look to make big gains* [Published: 2022-11-17. Accessed: 2025-04-27]. <https://www.dw.com/en/nepal-elections-young-independents-look-to-make-big-gains/a-63789472>

- Duwal, S., Prasai, S., & Manandhar, S. (2024). Domain-adaptative continual learning for low-resource tasks: Evaluation on nepali. <https://arxiv.org/abs/2412.13860>
- Elbayad, M., Wang, C., Bapna, A., Ma, X., Bhosale, S., Adi, Y., Pino, J., Goyal, N., Chaudhary, V., Ott, M., & Schwenk, H. (2023). Seamlessm4t: Massively multilingual & multimodal machine translation. *arXiv preprint arXiv:2308.11596*. <https://arxiv.org/abs/2308.11596>
- Explainers, F. (2024, March 5). *13 governments in 16 yrs: How nepal continues to see political churn* [Accessed: 2025-04-26]. Firstpost. Retrieved April 26, 2025, from <https://www.firstpost.com/explainers/pushpa-kamaldaha-new-alliance-13-governments-16-years-nepal-political-churn-13745366.html>
- Facebook Research. (2023). *Seamlessm4t* [Accessed: 2025-04-27]. [https://github.com/facebookresearch/seamless\\_communication](https://github.com/facebookresearch/seamless_communication)
- Guinaudeau, B., Munger, K., & Votta, F. (2022). Fifteen seconds of fame: Tiktok and the supply side of social video. *Computational Communication Research*, 4(2), 463–485.
- Highfield, T., & Leaver, T. (2016). *Instagrammatics and digital methods*. Routledge.
- Hugging Face. (2024). *Facebook/seamless-m4t-medium* [Accessed: 2025-04-27]. <https://huggingface.co/facebook/seamless-m4t-medium>
- ICUC Social. (2025). *Overcoming the challenges of audio moderation* [Accessed: 2025-04-27]. <https://icuc.social/resources/blog/overcoming-the-challenges-of-audio-moderation/>
- Ingelstam, F. (2023). Populism from below: Mapping user-centered populist content on tiktok.
- Kandel, S. (2024). Weaponizing information: The rise of social media manipulation in nepali politics [Discusses Balen Shah’s use of social media and the role of platforms like Routine of Nepal Banda in the 2022 elections]. *Journal of Durgalaxmi (JDL)*, 3(1), 1–18. <https://www.nepjol.info/index.php/jdl/article/download/73833/57195/216545>
- Kolena. (2025). *Cohen’s kappa - testing with kolena* [Accessed: 2025-04-27]. <https://docs.kolena.com/metrics/cohens-kappa/>
- Lamichhane, S. (2024). Tik tok ban in nepal: Reactions of the public. *Sprink Journal of Arts, Humanities and Social Sciences*, 3(4), 24–27.
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159–174.

- Lean Mean Learning Machine. (2023). *Audio analysis with deep learning: Techniques and challenges* [Accessed: 2025-04-27]. <https://lean-mean-learning-machine.com/machine-learning/audio-analysis-with-deep-learning-techniques-and-challenges/>
- McGregor, S. C. (2019). Social media as a tool for political agenda-setting [Examines social media's role in agenda-setting during elections, cited to contextualize TikTok's impact.]. *Political Communication*, 36(4), 521–539. <https://doi.org/10.1080/10584609.2019.1619639>
- Menge, J., & Dhakal, D. (2022). *Kathmandu's rapper revolution* [Accessed: 2025-04-28]. <https://asia.fes.de/news/kathmandus-rapper-revolution.html>
- Nakatani, S. (2014). *Langdetect: Language detection library ported from google's language-detection* [Accessed: 2025-04-27]. <https://pypi.org/project/langdetect/>
- NapoleonCat. (2025, March). *Social media users in nepal – 2025* [Accessed: 2025-04-27]. <https://napoleoncat.com/stats/social-media-users-in-nepal/2025/>
- Nepali Times. (2025). *Election infowar goes digital* [Accessed: 2025-04-28]. <https://nepalitimes.com/here-now/election-infowar-goes-digital>
- NepalNews. (2024). *Everything you need to know about kathmandu mayor balen shah: His battle with the establishment, core team, achievements and future plans* [Accessed: 2025-04-26]. <https://nepalnews.com/s/long-reads/everything-you-need-to-know-about-kathmandu-mayor-balen-shah-his-battle-with-the-establishment-core-team-achievements-and-future-plans/>
- Nisa, M. U., Mahmood, D., Ahmed, G., Khan, S., Mohammed, M. A., & Damaševičius, R. (2021). Optimizing prediction of youtube video popularity using xgboost. *Electronics*, 10(23), 2962. <https://doi.org/10.3390/electronics10232962>
- Noble, S. U. (2018). *Algorithms of oppression*. NYU Press.
- OpenAI. (2024). *Gpt-4o model documentation* [Accessed: 2025-04-27]. <https://platform.openai.com/docs/models/gpt-4o>
- Peña-Fernández, S., Larrondo-Ureta, A., & Morales-i-Gras, J. (2022). Current affairs on tiktok. virality and entertainment for digital natives [Accessed: 2025-04-26]. *Profesional de la Información*, 31(1), e310106. <https://doi.org/10.3145/epi.2022.ene.06>
- Piatak, J., & Mikkelsen, I. (2021). Does social media engagement translate to civic engagement offline? *Nonprofit and Voluntary Sector Quarterly*, 50(5), 1079–1101.

- Pinto, G., Bickham, C., Salkar, T., Luceri, L., & Ferrara, E. (2024, July). Tracking the 2024 us presidential election chatter on tiktok: A public multimodal dataset. Retrieved April 22, 2025, from <https://ssrn.com/abstract=4883401>
- Pudasaini, S., Shakya, S., Tamang, A., Adhikari, S., Thapa, S., & Lamichhane, S. (2023). Nepalibert: Pre-training of masked language model in nepali corpus. *2023 7th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, 325–330. <https://doi.org/10.1109/I-SMAC58438.2023.10290690>
- Radford, A., Kim, J. W., Hallacy, C., et al. (2021). Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.
- Reddit. (2023). *Gpt-4's nepali translation is on point* [Accessed: 2025-04-27]. [https://www.reddit.com/r/Nepal/comments/12j6twf/gpt\\_4s\\_nepali\\_translation\\_is\\_on\\_point/](https://www.reddit.com/r/Nepal/comments/12j6twf/gpt_4s_nepali_translation_is_on_point/)
- Reimers, N., & Gurevych, I. (2019). Sentence-bert. *EMNLP*.
- Reimers, N., & Gurevych, I. (2020). Sentence-transformers/paraphrase-multilingual-minilm-l12-v2 [Accessed: 2025-04-28].
- Reimers, N., & Gurevych, I. (2021). Sentence-transformers/all-minilm-l6-v2 [Accessed: 2025-04-28].
- Reuters. (2024, August). *Nepal lifts tiktok ban after app addresses cyber crime concerns* [Accessed: 2025-04-28]. <https://www.reuters.com/world/asia-pacific/nepal-lifts-tiktok-ban-after-app-addresses-cyber-crime-concerns-2024-08-22/>
- Roring, R. S. (2024). *Decoding social media virality: Measuring impact, influence, and engagement in the digital age* (tech. rep.) (Accessed: 2025-04-28. Email: riovan@universitasmulia.ac.id). Faculty of Computer Science, Information System Program, Mulia University, Indonesia. [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=4875869](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4875869)
- Sah, T., & Jordan, K. (2025). Decoding reddit memes virality [Published online ahead of print assignment]. *International Journal of Data Science and Analytics*. <https://doi.org/10.1007/s41060-025-00772-5>
- Schellewald, A. (2021). Communicative forms on tiktok: Perspectives from digital ethnography. *International Journal of Communication*, 15, 21.
- Setopati. (2025, May 1). *Sthapit calls balen shah fraud accusing him of hiding madhesi identity* [Accessed: 2025-05-01]. <https://en.setopati.com/political/158415>

- The Annapurna Express. (2022). *Balen and sampang effect in federal elections* [Accessed: 2025-04-26]. <https://theannapurnaexpress.com/story/30628/>
- The Himalayan Times. (2025). *Prime minister kp sharma oli makes tiktok debut* [Accessed: 2025-04-28]. <https://thehimalayantimes.com/entertainment/prime-minister-kp-sharma-oli-makes-tiktok-debut>
- TikTok. (2023). *Tiktok creative research 2023* [Accessed: 2025-04-27]. <https://www.tiktok.com/business/en/creativecenter/insights>
- TikTok. (2025). *Research api / tiktok for developers* [Accessed: 2025-04-28]. <https://developers.tiktok.com/products/research-api/>
- Umansky, N., & Pipal, C. (2023). Dances, duets, and debates: Analysing political communication and viewer engagement on tiktok [September 21, 2023].
- Van Dijck, J., & Poell, T. (2013). Understanding social media logic [Provides a framework for social media's role in political communication, used in the literature review.]. *Media and Communication*, 1(1), 2–14. <https://doi.org/10.12924/mac2013.01010002>
- Wikipedia contributors. (2022). *2022 nepalese local elections — wikipedia, the free encyclopedia* [Accessed: 2025-04-28]. [https://en.wikipedia.org/wiki/2022\\_Nepalese\\_local\\_elections](https://en.wikipedia.org/wiki/2022_Nepalese_local_elections)

# AI Disclosure

In accordance with the Hertie School's AI Guidelines (February 2023), I declare that I have used AI tools such as ChatGPT, Grok, and Google AI Studio for assistance in coding and thesis preparation. These tools were utilized in the following manner:

1. Code generation and debugging
2. Drafting and structuring certain sections of the text
3. Assistance with LaTeX formatting
4. Help with language refinement and clarity

I confirm that all AI-generated content has been carefully reviewed, edited, and verified by me, and I take full responsibility for all content, including any factual errors or inaccuracies that may exist in the final document. The use of these tools has been in accordance with the Hertie School's policy on academic integrity, with all significant contributions properly acknowledged.



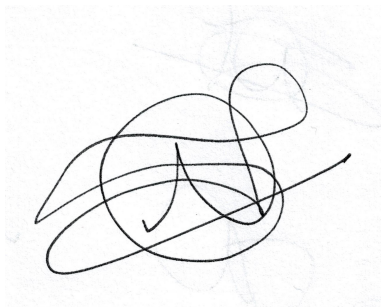
# Statement of Authorship

I hereby confirm and certify that this master thesis is my own work. All ideas and language of others are acknowledged in the text. All references and verbatim extracts are properly quoted and all other sources of information are specifically and clearly designated. I confirm that the digital copy of the master thesis that I submitted on 28th April 2025 is identical to the printed version I submitted to the Examination Office on 29th April 2025.

**DATE:** 28th April 2025

**NAME:** Nima Thing

**SIGNATURE:**

A handwritten signature in black ink, consisting of several overlapping loops and a long horizontal stroke extending to the right. The signature is written on a light blue grid background.