Influencing the Divide: Political Polarization in Indian Influencer Discourse on X

Ziv Barretto, Agriya Yadav, Kyra Chhetri, Anirban Sen
Ashoka University

Abstract

This paper investigates partisan discourse on X (formerly Twitter), examining how political identities and affiliations shape online communication patterns. We analyze a large-scale dataset of tweets retweeted by Indian politicians from 2020-2023 to understand political polarization in social media. Our methodology combines influencer polarity scoring based on retweet behavior, automated keyword extraction using KeyBERT, and stance classification using a LoRA-adapted large language model. We identify politically polar keywords and classify tweet stances toward these topics to reveal patterns in how ruling and opposition parties engage with different political narratives.

1 Introduction

The rise of social media platforms has fundamentally transformed political communication and discourse. X (formerly known as Twitter) has emerged as a critical platform for political discussion, serving as a space where politicians, journalists, activists, and citizens engage in real-time conversations about political issues. However, this democratization of political discourse has coincided with growing concerns about political polarization and the formation of ideological echo chambers as pointed out by several recent research studies [? ?].

This polarization has several detrimental effects on the information ecosystem, and the society at large. The lack of spread and cross-pollination of ideas across communities lead to unbalanced and incorrect judgment of social and political issues, thereby biasing electoral outcomes. The problem is further compounded by virality of mal-information and toxic content, resulting in an undesirable impact on democracy [?].

More often than not, the boundaries of these echo chambers are fortified when the information is generated by a traditional influential figure (a prominent politician, business person, celebrity, etc.) with a significant network reach and follower count.

Moreover, with the increasing accessibility of digital technologies, a second tier of micro-influencers (hereafter referred to as *influencers*) has emerged as a significant contributor to polarization. Many of these individuals primarily derive their livelihoods from social media platforms and, in some cases, attain levels of popularity comparable to or exceeding those of traditional public figures. Content generated or disseminated by these influencers reaches a substantial audience, and consequently, ideological or opinion-based polarization among them further amplifies and compounds broader societal polarization. The political and social impact of these digital influencers can be estimated by the fact that influencer generated content accounted for around 36 % of all user actions across major social platforms (Facebook, Instagram, X, TikTok) in April 2025, reflecting their mass global engagement [?]. Political content alone drove over 30 % of influencer interactions in India. Additionally, the influencer base seems to be ever-growing on social media, with India alone seeing a significant rise in number of influencers starting 2020 [?].

This paper investigates one of the axes of polarization among such influencers, namely the political axis. In other words, we study the partisan influencer discourse on X, examining if and how political leaning of influencers determine their online communication patterns on various social and political issues. The primary research question we intend to address in this work is: *How does the political discourse of influencers on X vary on different social and political issues based on their political leaning?*

We currently focus on the Indian political context, although our methods are generalizable to any geography. We first develop a unique proxy for political leaning of influencers based on how politicians from the centrally ruling and opposition parties engage with their content. Next, we develop an *Aspect based Stance Classifier* based on a small language model (*Mistral-7B*)[?], which analyzes the stance of the influencer tweets corresponding to the primary aspects/subjects contained in them. We also carry out our analysis separately for the tweets written in English and Hindi, to compare our findings across the linguistic spectrum.

Our findings empirically reveal that Indian influencers exhibit bipolar polarization with respect to a wide range of social and political issues, as identified by the aspects present in their tweets. Furthermore, an influencer's political leaning is strongly correlated with the stance they adopt when tweeting on a given aspect. These observations raise several important questions and directions for future work, including how such polarization evolves over time, the mechanisms through which influencers shape and reinforce audience beliefs, and the extent to which polarized influencer discourse affects information diffusion and opinion formation among social media users. Future research could also investigate causal relationships between influencer stances and shifts in public opinion, as well as explore intervention strategies aimed at mitigating the amplification of polarization in social media ecosystems.

2 Related Work

Social media, lately, has become an indispensable source for news and information consumption [? ?] for the masses. A significant fraction of users stay updated with current affairs through social media, which leads to its impact on public opinion. Previous works report about this impact in various social and political fronts [? ?], thereby conclusively proving that social media carries the power to reconfigure the very foundations of democracy. However, increasing polarization in social media has a detrimental impact on the society. In this section, we discuss related work spanning the two primary components of our proposed framework – analysis of social media polarization, and stance classification.

2.1 Social Media Polarization

While originally envisioned as a tool to broaden the horizons of knowledge and ideas, social media has contributed to an increasingly polarized and fragmented global society [? ?]. Polarization leads to imbalance in opinion formation, hindering individuals from obtaining a holistic view of salient issues. Numerous examples of the detrimental social impact causes by a polarized social media have been highlighted in prior studies [ref]. With the increasing penetration of digital technologies, policy-makers, think-tanks, and governments are increasingly planning necessary interventions to tackle the undesirable effect social media brings on the society [? ?].

Furthermore, the impact of polarization is often accentuated by *micro-influencers* – social media creators with dedicated following focused on a specific niche. These micro-influencers (or influencers) usually have a significantly high number of regular followers forming closely knit communities. Prior work has sufficiently exemplified cases where these influencers have promoted certain political agenda through highly partisan content posted through their social media accounts [?]. Owing to their network reach and impact, such content is consumed by their follower base, strengthening the pre-existing echo chambers. Similar trends have also been observed in India [?], which forms the third highest user base for X, capturing around 36% of all social media interactions globally across platforms [?].

Social media polarization has been studied using both content [? ? ?] and network analysis methods [? ? ? ? ? ?] in extant literature. The connection between partisanship in social media content and polarization has been well established in prior work. Demszky et al. [?] use linguistic and semantic content analysis show that polarization is affected mainly by partisan differences in content framing. Flamino et al. [?] analyze tweets to examine how partisan news content circulated by media and influencer accounts increased ideological polarization between the 2016 and 2020 US elections. Dash et al. [?] quantify how influences engage with politically charged content in a partisan manner, showing that ideological polarization among influencers is linked to increased engagement and retweeting, thus amplifying partisan dynamics.

Motivated by prior work in this area, we study the political discourse generated by Indian influencers to assess the level of partisanship in it. For this purpose, we first manually annotate a set of influencer generated tweets to create a dataset for aspect based stance classification. It is ensured that these tweets are politically endorsed, i.e., have been retweeted at least once by a political entity. Next, we use this dataset to fine-tune a small language model (Mistral-7B) to develop an aspect based stance classifier. While existing works have focused on various stance classification methods including XXX [ref], XXX [ref], and XXX [ref], we use an SLM based approach to ensure that our framework is generalizable to low resource settings. We also evaluate partisanism in influencer generated content in both English and Hindi, to establish robustness of our findings.

2.2 Stance Classification

Stance classification has long been recognized as a challenging NLP task involving the identification of a speaker's holistic subjective disposition toward a topic or aspect. Prior work has shown that stance is not reducible to sentiment [?], and requires leveraging of significantly more nuanced linguistic features. Stance classification methods include both unsupervised [?], and supervised [? ? ? ?] approaches. Prior work has shown that transformer-based approaches often substantially outperform classical approaches for stance detection on social media data [? ?]. Recent work has additionally examined the use of LLMs for stance classification [?], showing that LLM performance is highly sensitive to prompt design and that parameter-efficient fine-tuning such as LoRA does not consistently outperform zero-shot prompting [? ?]. Prior work has also shown that effective stance classification requires rich representations and contextual modeling beyond surface lexical features [?]. Our work operationalizes these insights by extending this line of research to large-scale political influencer discourse, integrating LLM-based stance reasoning of multilingual tweet data. Specifically, we test the performance of both basic transformer and LLM based approaches (zero-shot, few-shot, and fine-tuned) in the task of stance classification, on influencer generated political tweets.

3 Data Collection

We utilize a dataset of tweets provided by Professor Joyojeet Pal, comprising retweets made by Indian politicians on X (Twitter) during the period 2020–2023. The dataset contains over 7.1 million retweet records, capturing the amplification behavior of politicians across the Indian political spectrum. Each record represents a retweet action containing the original tweet content, timestamp, retweeting politician's X handle, and the original author (influencer) whose content was retweeted.

To associate politicians with their party affiliations, we utilize an external mapping file containing X handles and corresponding party memberships. This mapping covers politicians from over 50 political parties, which we aggregate into two major blocs as shown in Table **??**.

Table 1: Political Party Bloc Mapping

Ruling Bloc	Opposition Bloc
BJP, JDU, LJP, HAM, JSP, NPP, AIADMK, AJSU, AGP, RPI, NPF, IPFT, NDPP, RSS, NDA, MNF, VHP, YSRCP, ABVP, BJD, BSCP	INC, AITC/TMC, DMK, SP, CPIM, RJD, AAP, JMM, IUML, CPI, NCP, TRS, Shiv Sena, TDP, JDS

The data processing pipeline consists of two stages. In the first stage (party labeling), we process the raw CSV in chunks of 100,000 rows and match each retweeting politician's handle to their party affiliation using lowercase normalization, producing a labeled dataset with the `retweet_party` column. In the second stage (polarity calculation), we compute yearly polarity scores for each influencer based on who retweets their content, then average across years to obtain a stable polarity measure, identifying 960 unique influencers with calculable scores. Table **??** presents the key statistics of our dataset.

Table 2: Dataset Statistics

Metric	Value
Total retweets	7,115,963
English tweets	2,939,716 (41.3%)
Hindi/Regional tweets	4,176,247 (58.7%)
Unique influencers (with polarity)	960
Influencer-year observations	3,712
Temporal range	2020–2023

The temporal distribution (Table **??**) shows consistent activity across all four years, with relatively balanced influencer-year records ranging from 908 to 949 per year.

Table 3: Temporal Distribution of Tweets

Year	Influencer-Year Records
2020	908
2021	944
2022	949
2023	911

Table **??** describes the key columns in our final processed dataset. The original dataset contains tweets in multiple languages, including Hindi and regional Indian languages. For the English analysis pipeline, we filter tweets using an ASCII-ratio heuristic: a tweet is classified as English if at least 85% of its alphabetic characters are ASCII letters, providing computationally efficient language detection for large-scale processing.

Table 4: Dataset Column Descriptions

Column	Description
timestamp	Date and time of the retweet (UTC)
tweet	Full text content of original tweet
retweet_author	X handle of retweeting politician
original_author	X handle of influencer
retweet_party	Political party of retweeter
year	Year extracted from timestamp
side	Political bloc (ruling/opposition/other)
polarity_avg	Influencer's average polarity score
tweet_label	Inferred political leaning of tweet

4 Methodology

4.1 Influencer Polarity

We calculate an influencer's political leaning (also termed as *polarity*) using the amount of endorsement their tweets receive from Indian politicians. Generally, the retweet behavior of politicians reveals their endorsement towards a tweet [ref]. Thus, by aggregating retweet patterns across politicians with known party affiliations, we can infer the political leaning of influencers. This methodology provides a more reliable signal of endorsement than passive following or hashtag co-occurrence, since hashtags can be appropriated or used sarcastically, while generic sentiment analysis methods struggle with political nuance and sarcasm.

For each influencer i and year y, we calculate a yearly political polarity score using the formula in Equation **??** where $R_{i,y}$ represents the count of retweets the influencer received from ruling politicians in year y, and $O_{i,y}$ represents the count from opposition politicians.

$$P_{i,y} = \frac{R_{i,y} - O_{i,y}}{R_{i,y} + O_{i,y}} \qquad (1)$$

The polarity score ranges from -1 (exclusively retweeted by opposition) to $+1$ (exclusively retweeted by ruling). We next compute an average polarity score across all available years using Equation **??**, where Y_i is the set of years in which influencer i was retweeted by at least one politician.

$$P_{avg}^i = E_y[P_{i,y}] = \frac{1}{|Y_i|} \sum_{y \in Y_i} P_{i,y} \qquad (2)$$

Based on the average polarity score, influencers are classified into three categories using a threshold of 0.5 (experimentally determined). Influencers with $P_{avg}^i \geq 0.5$ are classified

as Pro-Ruling, those with $P^i_{avg} \leq -0.5$ as Pro-Opposition, and those within $-0.5 < P^i_{avg} < 0.5$ as Neutral. This classification is propagated to all tweets authored by each influencer, effectively labeling tweets based on their author's political leaning.

Three independent annotators also manually analyzed the influencer polarities to validate the results. This exercise reached a high inter-annotator agreement (95% unanimous agreement), revealing the above-par performance of the procedrure.

4.2 Aspect Identification

To identify the topics or aspects discussed in influencer tweets, we employ KeyBERT[**?**] for automated keyword extraction from tweet text. KeyBERT leverages BERT embeddings to identify keywords and phrases that are semantically similar to the document as a whole, making it suitable for capturing the topical essence of short-form social media content.

Before aspect extraction, we preprocess the tweets to remove URLs, strip the "@" character while retaining usernames, perform hashtag segmentation using the `wordsegment` library (e.g., `#AatmaNirbharBharat` becomes *aatma nirbhar bharat*), and normalize whitespaces.

The extraction process uses the `paraphrase-multilingual-MiniLM-L12-v2` sentence transformer model as the KeyBERT backend. We configure n-gram range of 1–2, extract top three aspects per tweet, and apply Maximal Marginal Relevance (MMR) with diversity 0.7 to balance semantic similarity with keyword diversity.

Polar Aspect Identification: To understand if influencers with a certain political polarity preferentially tweet around certain political aspects, we analyze the aspects to identify *polar aspects* – topics disproportionately discussed by one political side. We first manually selected 15 aspects representing major political topics for annotation, chosen based on their political significance and representation across both ruling and opposition discourse. We examined the proportion of tweets from pro-ruling and Pro-Opposition sources for each selected aspect, and categorize the aspect as pro-ruling/opposition based on which side tweets more about it. This forms the seed set of polar aspects. Next, to expand coverage, we systematically analyzed aspect frequencies and identified 23 additional polar aspects exhibiting clear partisan skew. Table **??** presents the complete set of 38 keywords used for stance analysis.

Table 5: Complete Keyword Set for Stance Analysis

15 seed aspects identified manually			
caa	congress	farm_laws	farmers_protests
hindu	hindutva	kashmir	kashmiri_pandits
modi	muslim	new_parliament	rahulgandhi
ram_mandir	shaheen_bagh	china	

23 extended aspects identified using frequency analysis			
aatmanirbhar	ayodhya	balochistan	bhakts
democracy	demonetisation	dictatorship	gdp
hathras	inflation	islamists	lynching
mahotsav	minorities	msp	ratetvdebate
sangh	sharia	spyware	suicides
ucc	unemployment		

Aspects like *aatmanirbhar* (self-reliance campaign), *ucc* (Uniform Civil Code), *ayodhya*, and *mahotsav* are predominantly used by Pro-Ruling sources ($> 74\%$), while *unemployment*, *demonetisation*, *bhakts*, and *spyware* are predominantly discussed by Pro-Opposition sources ($> 80\%$).

4.3 Stance Classification

The goal of our stance classification system is to determine how influencers position themselves toward specific political topics. Given a tweet t and a target keyword e (such as *modi*, *caa*, or *farmers_protests*), the system predicts one of three stance labels: Favor (expressing support), Against (expressing opposition), or Neutral (no clear stance). This three-way classification avoids the ambiguity of an "unrelated" category, which in practice often overlaps with neutral stances.

Manual Annotation Dataset

A critical contribution of this work is the creation of a rigorously manually annotated stance classification dataset for Indian political discourse. Unlike prior work that often relies on weakly-supervised or automatically generated labels, we invested significant effort in creating high-quality ground truth data to serve as a reliable benchmark. This dataset plays a central role in our pipeline: it provides the core supervision signal for fine-tuning our baseline discriminative models (BERT, RoBERTa, PyABSA), serves as the source for retrieval-augmented few-shot prompts used by our generative models, and functions as the gold-standard held-out test set for rigorous performance evaluation.

The dataset creation involved a multi-stage process beginning with stratified sampling. For each of the 15 seed polar keywords, we sampled 100–150 tweets, ensuring a balanced representation of viewpoints from both Ruling and Opposition sources. This stratification technique is crucial as it prevents the model from learning specious correlations between political affiliation and stance, such as assuming all opposition tweets are inherently "Against". Following sampling, strict deduplication was applied to remove near-duplicate tweets, ensuring the dataset captures diverse linguistic expressions rather than repetitive coordinated behavior. Finally, each tweet was manually annotated with its stance toward the specific target aspect. Annotators were instructed to look for explicit expressions of support or opposition, labeling ambiguous or factual statements as Neutral.

The final dataset comprises 1,866 labeled tweet-keyword pairs. We adopted a strict train-test split strategy where 264 samples were set aside as a held-out test set (stratified by keyword and class). From the remaining training pool, we reserved 10 examples per keyword to serve as few-shot demonstrations, and utilized the rest for supervised training of baseline models.

Multi-Model Baseline Comparison

To rigorously evaluate the effectiveness of our approach, we established a comprehensive baseline comparison using diverse model architectures ranging from specialized aspect-based sentiment analysis systems to general-purpose encoders. Table **??** details the specific models and configurations used.

Table 6: Models Used for Stance Classification Comparison

Model	Type	Method
BERT-base-uncased	Encoder	Full Fine-tuning
RoBERTa-base	Encoder	Full Fine-tuning
PyABSA (FAST_LCF)	ABSA	Aspect-focused Fine-tuning
Mistral-7B-Instruct	Decoder	Zero-shot / Few-shot
Mistral-7B + LoRA	Decoder	LoRA Fine-tuning

For the encoder-based baselines, we fine-tuned `bert-base-uncased` using an aspect-aware input format: `[CLS] Aspect [SEP] Tweet [SEP]`. This allows the self-attention mechanism to directly model the relationship between the target topic and the tweet content. Similarly, for `roberta-base`, we employed a prompt-like input format by explicitly prepending the context: "Topic: {Aspect}. Tweet: {Content}", leveraging RoBERTa's strong semantic understanding. We also utilized the specialized PyABSA framework's `FAST_LCF_BERT` model, which uses a Local Context Focus (LCF) mechanism to dynamically weight context words based on their semantic and syntactic distance to the aspect. All baseline models were trained on the exact same training split (1,458 samples) and evaluated on the same held-out test set (264 samples) to ensure a strictly fair comparison.

LoRA-Adapted Mistral (Our Approach)

We employ Mistral-7B-Instruct as our primary backbone, enhanced with Low-Rank Adaptation (LoRA) for parameter-efficient fine-tuning.

Why LoRA?: Full fine-tuning of 7B+ parameter models is computationally prohibitive and prone to catastrophic forgetting. LoRA freezes the pre-trained weights W and injects trainable rank decomposition matrices A and B, such that the weight update is $\Delta W = BA$, where $B \in \mathbb{R}^{d \times r}$ and $A \in \mathbb{R}^{r \times k}$. We set the rank $r = 16$ and alpha $\alpha = 32$, targeting all linear projection layers (query, key, value, output, gate, up, down).

Training: The model was fine-tuned on the training split to generate structured JSON outputs containing both the stance label and a reasoning field. This instruction-tuning approach encourages the model to perform chain-of-thought reasoning before assigning a label, improving interpretability.

Evaluation Results

We evaluated all models on the held-out test set of 264 samples. Table **??** presents the comparative results. The LoRA-adapted Mistral model demonstrates competitive performance, effectively balancing precision and recall across classes.

Table 7: Stance Classification Evaluation Results

Metric	Value
Total test samples	264
Accuracy	78.0%
Precision (macro)	78.0%
Recall (macro)	75.3%
F1-Score (macro)	76.2%
F1-Score (weighted)	77.7%

Table **??** shows per-class performance. The model performs strongest on the "favor" class (F1=0.82), followed by "against" (F1=0.79). The "neutral" class shows lower recall (0.60), reflecting the inherent difficulty of distinguishing neutral stances from weakly expressed opinions.

Table 8: Per-Class Classification Performance

Class	Precision	Recall	F1	Support
Against	0.78	0.80	0.79	95
Favor	0.78	0.86	0.82	111
Neutral	0.78	0.60	0.68	58

Table **??** presents per-keyword accuracy, showing that some topics yield clearer stance signals than others. Keywords like *modi* achieve the highest accuracy (94.4%), followed by *caa* (89.5%) and *congress* (88.9%).

Table 9: Per-Keyword Classification Accuracy (Top 10)

Keyword	Accuracy	F1 (macro)
modi	94.4%	0.91
caa	89.5%	0.90
congress	88.9%	0.84
hindutva	85.7%	0.56
new_parliament	85.7%	0.82
rahulgandhi	81.3%	0.81
farmers_protests	80.0%	0.82
shaheen_bagh	78.9%	0.71
muslim	78.9%	0.79
china	73.7%	0.64

Final Inference Pipeline

For large-scale stance classification across all 38 keywords, we deploy an inference pipeline that combines the trained LoRA adapter with few-shot prompting. The pipeline loads the base Mistral-7B-Instruct model and applies the trained LoRA adapter using PEFT's `PeftModel`, effectively deploying our fine-tuned stance classifier without modifying the base model weights.

The inference prompt is constructed using the same 10 few-shot examples per keyword that were reserved during annotation. For each input tweet, the pipeline retrieves keyword-specific examples following the naming pattern `{prefix}_{keyword}_stance.json`. For the 23 extended keywords that lack dedicated annotations, the system falls back to a global set of representative examples, enabling stance classification across all 38 keywords with consistent prompt structure.

Inference uses deterministic settings (temperature 0, no sampling) to ensure reproducible outputs. The model returns JSON containing stance and reasoning, which is parsed by a robust extraction function that handles various output formats and applies label normalization. Batched processing with length-based bucketing enables efficient large-scale inference, while resume functionality and periodic checkpointing ensure reliability for processing the full dataset.

4.4 Hindi Tweet Analysis

We also analyzed the tweets written in Hindi (Devanagari and Roman fonts) to perform a comparative analysis of the trends around stance classification in Hindi with that observed for English tweets.

For this purpose, we first selected 500 unique Hindi tweets, by dropping retweets and duplicates. We also stratified the data by script×URL×hashtag, to prevent over-/under-representation exhibited in Naive Random Sampling. While there are NLP tools for aspect extraction and stance analysis for non-English/Hindi documents, in this work we use a translation pipeline to convert the Hindi tweets to their English equivalent, and then performing stance classification on this data. We performed a detailed sanity check of the translated data, and the final results of stance classification, to ensure robustness of our findings [to be reported]. A major advantage of this approach is the applicability of SOTA NLP approaches and their consistency in performance across both datasets (English and Hindi). Since we apply the same methods of aspect extraction and stance analysis, their performance on both datasets are easily comparable, ensuring generalizability of our findings.

We used the `facebook/nllb-200-distilled-600M` model [?] to translate the Hindi tweets to English [because...]. The model also has the additional advantage of efficient handling of both pure Devanagari and Roman Hindi scripts, both of which are abundant in our data. We used a batch size of 16 and set the `max-new-tokens` parameters = 128, experimentally. This resulted in translation of all 500 Hindi tweets to English.

Next, our goal was to perform quality check of the translated data. Standard qualitative analysis approaches to check translation quality often suffer from sampling issues, and the requirement of significant manual effort. To avoid this, we developed a quantitative method to assess the quality of translations, on top of a qualitative layer [to be provided].

Quantitative Quality Assessment: For the quantitative analysis, we first represent the two datasets in the same embedding space, and then observe the drift between each Hindi and its equivalent translated English data point, to assess their semantic coherence. If the average drift is lower than a certain threshold, we determine the translation quality to be above par, and vice versa.

We first embed the original and the translated texts using the same `paraphrase-multilingual-MiniLM-L12-v2` model[?]. [Write some details about this model and its multilingual ability]. This ensures that both the Hindi/Roman Hindi and English texts are embedded into a shared embedding space. We then measure the cosine similarity of the embeddings to see how semantically similar the original and translated data points are. We finally experimentally determine a set of similarity threshold scores. Specifically, if the two data points exhibit:

- A similarity score of above 0.9, they are considered to be semantically coherent

- A similarity score of below 0.85, they are considered to be semantically incoherent

- A similarity score $\epsilon[0.85, 0.9]$, the data points are considered to require further manual inspection

Our analysis revealed a median cosine similarity of 0.95, indicative of a significant number of semantically coherent translations. 15.8% of the data points exhibited a score of below 0.9, 1.8% between 0.8 and 0.85 (inspect), and 0.4% below 0.8 (incoherent). This indicates that the quality of most of our translations are significantly high.

We also use *Uniform Manifold Approximation and Projection* (UMAP) [?] to project the original and translated embeddings into a 2D space (figure ??), to observe the drifts between pair of data points. UMAP is a non-linear dimensionality reduction technique to visualize high-dimensional data in 2D or 3D. It is generally faster and better at preserving overall structure of data compared to other existing methods for dimensionality reduction. We see that both the Hindi and corresponding English data points are similarly distributed in the 2D space, and only a few exhibit significant drifts, once again corroborating our finding of satisfactory translation quality.

Figure 1: UMAP projection of original (Hindi/Roman) vs translated (English) tweets with drift arrows.

We also performed a manual check of the translated data, to doubly ensure the quality of translations generated. Two annotators (comfortable with Hindi) independently annotated 100 randomly selected examples for their translation quality. A label of 1 was assigned for satisfactory and semantically coherent translation, and 0 was assigned otherwise. This exercise revealed an inter-annotator agreement of [XXX%] and an accuracy of [XXX%], indicating an above par performance of the translation pipeline.

[– need to write about how we will handle values below 0.85 and 0.80- either manual review or treating as mistranslations. The thresholds need to be decided based on an analysis of the empirical distribution.]

Table **??** Showcases the justification for our use of multilingual sentence embeddings, we compared `paraphrase-multilingual-MiniLM-L12-v2` with a monolingual MiniLM baseline on 5,000 Hindi–English translation pairs. The multilingual model achieved high cross-lingual retrieval accuracy (R@1 = 0.837, MRR = 0.854) and near-perfect separation between true and mismatched translation pairs ($\delta = 0.539$, AUC = 0.960, Cohen's d = 2.88), while the monolingual model performed substantially worse (R@1 = 0.462, AUC = 0.717).

This behavior is consistent with the multilingual alignment objective described by Reimers and Gurevych [**?**], who show that sentence embeddings trained via knowledge distillation on parallel data form a shared semantic space across languages. Our results empirically confirm that only the multilingual model yields a sufficiently well-aligned cross-lingual embedding space for reliable translation quality assessment and multilingual stance analysis.

Connection to the Enligh pipeline

Multilingual Model Support

The Mistral-7B model supports multilingual input, enabling direct processing of Hindi tweets. The few-shot examples for Hindi tweets are constructed with Hindi text while maintaining the same JSON output format.

Fallback Mechanism

The pipeline implements hierarchical few-shot example retrieval:

1. Attempt to load keyword-specific examples

2. If unavailable, fall back to a global fallback JSON

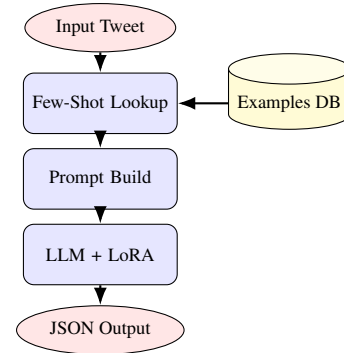This ensures robust inference even for keywords without dedicated few-shot examples.

Figure 2: Stance Detection System Architecture

5   Results

This section reports the empirical outcomes of the stance classification pipeline applied to Indian political influencer discourse on X. The analysis covers 680,847 tweet–keyword pairs across 38 political keywords, with influencers categorized by political affiliation into Pro-Ruling (462 influencers) and Pro-Opposition (498 influencers) groups.

Table 10: Embedding Model Evaluation

Metric	paraphrase-multilingual-MiniLM	monolingual-MiniLM
R@1	0.837	0.4618
R@5	0.8726	0.4676
R@10	0.8816	0.4716
MRR	0.854	0.465
POS_mean	0.831	0.606
POS_median	0.8805	0.5105
NEG_mean	0.2922	0.3031
NEG_median	0.2803	0.3077
Delta	0.5394	0.3029
AUC	0.9604	0.7168
Cohen_d	2.88	1.09

5.1 Model Performance

To validate our choice of the LoRA-adapted Mistral-7B model, we compared its performance against the baseline architectures described in the methodology. Figure **??** illustrates the comparative results across key metrics. The LoRA-tuned Mistral model consistently outperforms the baseline approaches, demonstrating superior capability in capturing the nuances of political stance.
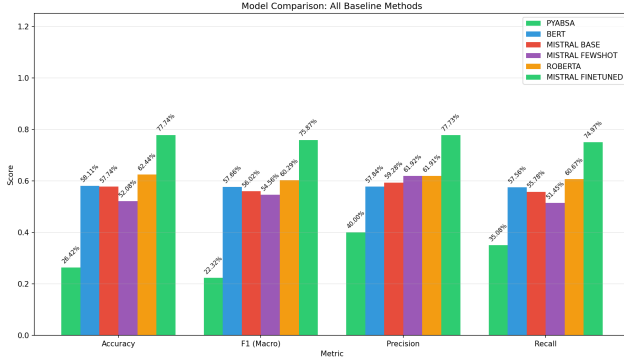

Figure 3: Performance comparison of the fine-tuned Mistral (LoRA) model against baseline architectures.

The confusion matrix presented in Figure **??** further elucidates the model's classification accuracy. We observe a strong diagonal, indicating high precision in correctly identifying Favor, Against, and Neutral stances. Notably, the model effectively minimizes confusion between opposing stances (Favor vs. Against), which is critical for accurate polarization analysis. The instances of misclassification are primarily concentrated between the Neutral category and polar stances, reflecting the inherent ambiguity often present in less explicit political expression.
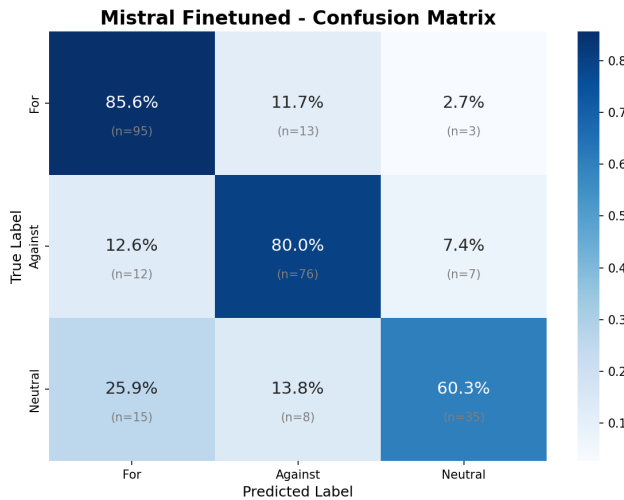

Figure 4: Confusion matrix for the fine-tuned Mistral model on the held-out test set.

5.2 Multilingual Stance Patterns

We extended our analysis to investigate whether political discourse patterns diverge across languages. Figure **??** presents a stacked stance distribution for English and Hindi tweets. A striking trend emerges: the stance distribution patterns are remarkably similar across both languages. This observation suggests that the polarization dynamics and narrative structures deployed by political influencers are consistent, irrespective of the linguistic medium. We further corroborated this finding by cross-referencing with a negativity scatter analysis (Figure **??**), which similarly indicated high congruence in the sentiment and negativity levels associated with specific keywords across both languages.
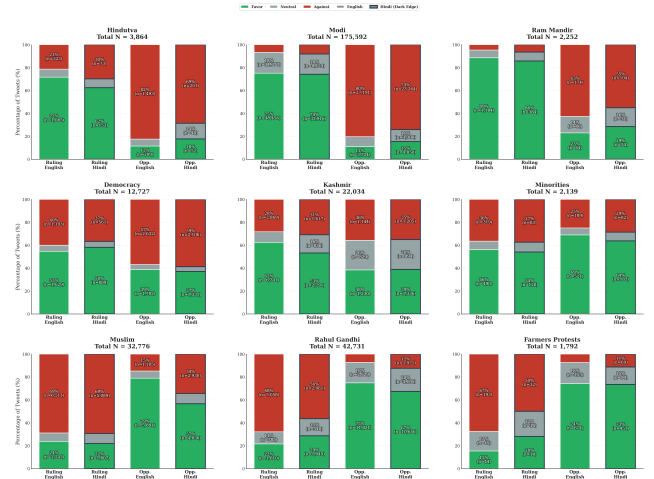

Figure 5: Stacked stance distribution comparison between English and Hindi tweets.

The visualization reveals three distinct bands of behavior. The top row highlights keywords where the Pro-Ruling faction exhibits extreme polarization, predominantly expressing strong favor. The middle row represents a "mixed" zone where both political parties show similar stance distributions, indicating areas of shared narrative or lower polarization. The bottom row serves as the counterpart to the first, showcasing keywords where the Pro-Opposition faction displays extreme polarization. Given this high degree of congruence between English and Hindi discourse, we appended the two datasets for the remainder of our analysis to provide a more comprehensive view of the political landscape.
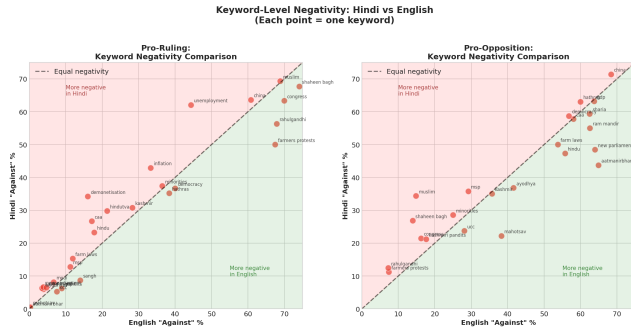
Figure 6: Comparison of negativity scores for keywords in English vs. Hindi tweets.

The scatter plot in Figure **??** plots the average negativity score of each keyword in English (X-axis) against its score in Hindi (Y-axis). We observe a strong positive correlation, with most keywords clustering along the diagonal. This indicates that topics evoking negative sentiment in English discourse tend to trigger similar levels of negativity in Hindi, reinforcing the hypothesis of a unified partisan narrative that transcends language barriers. The visual evidence suggests that the emotional valence of political topics is consistent across linguistic communities, further justifying the aggregation of datasets for subsequent analysis.

5.3 Normalized Favor Rates (Combined Data)

Using the combined English and Hindi dataset, we re-examined the normalized favor rates for each keyword. Figure **??** displays these rates using a butterfly chart, with Pro-Ruling favor percentages on the right and Pro-Opposition favor percentages on the left.
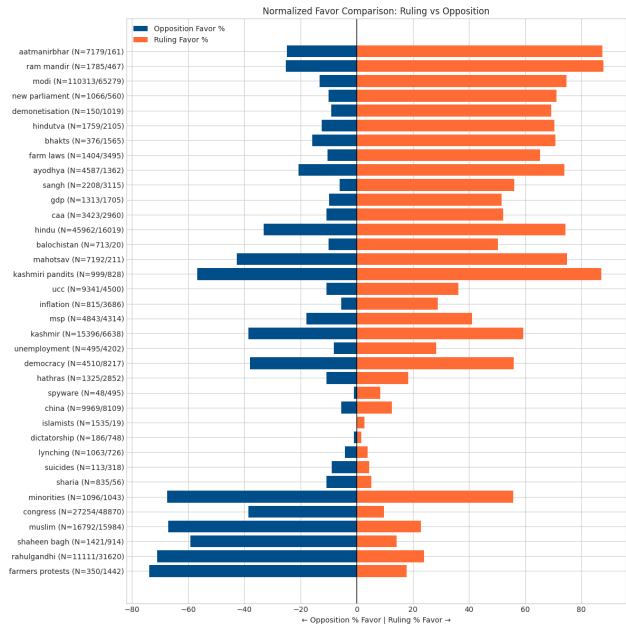

Figure 7: Butterfly chart showing normalized favor rates by keyword (combined English & Hindi data).

The combined analysis reinforces the bimodal nature of the discourse. We observe a set of keywords where the Pro-Ruling side exhibits near-complete dominance in favorable expression, contrasting sharply with keywords where the Pro-Opposition side dominates. Topics located near the center of the chart continue to exhibit more balanced favor distributions, but the overall trend confirms that most political topics in the Indian context are heavily skewed toward one side's narrative. This polarization is even more pronounced when leveraging the larger, combined dataset, validating the robustness of our initial findings.

5.4 Keyword-Level Stance Divergence

Figure **??** presents the keyword-level stance divergence for the combined corpus. This scatter plot maps each keyword based on the difference in favor rates (X-axis) and against rates (Y-axis) between the two political blocs.
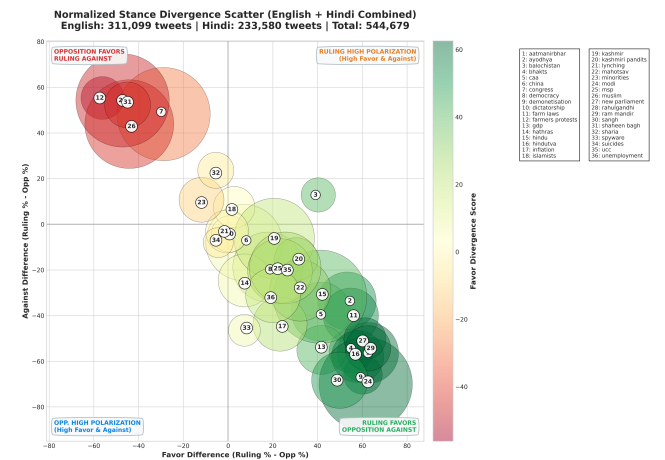

Figure 8: Keyword-level stance divergence scatter plot for the combined English and Hindi dataset.

The combined divergence analysis highlights the distinct "camps" occupied by different topics. We can clearly identify clusters of keywords that drive the most significant wedges between the political factions. Outliers in this plot represent topics that generate unique polarization patterns—for instance, topics where one side is highly favorable while the other is not necessarily hostile but silent, or where both are hostile. The distribution of points further confirms that divergence is not uniform; rather, specific high-volume topics drive the bulk of the polarization, while a "long tail" of issues remains relatively contested or neutral.

5.5 Complete Stance Distribution by Keyword

Figure **??** presents the full stance distribution for each keyword, disaggregated by political affiliation. Each keyword displays proportions of favor, neutral, and against stances for both Pro-Ruling and Pro-Opposition influencers.

../final_visualisations/4_stance_distribution_by_keyword.png

Figure 9: Stance distribution by keyword and political affiliation, showing proportions of favor, neutral, and against stances.

To organize keywords into meaningful thematic categories, we employed Google's Gemini model. The model was provided only with the list of 38 keywords and prompted to create thematic buckets based on its knowledge of Indian political discourse. Table **??** presents the resulting keyword categories.

Table 11: Keywords Grouped by Thematic Category (Generated using Gemini)

Category	Keywords
Favor-dominant (both)	ayodhya, mahotsav, ucc
Against-dominant (both)	inflation, unemployment, suicides
Split (PR favor, PO against)	modi, ram_mandir, aatmanirbhar, hindutva, farm_laws, caa
Split (PO favor, PR against)	rahulgandhi, congress, farmers_protests, shaheen_bagh, muslim
Mixed neutral (>20%)	china, gdp, democracy, minorities

Note: PR = Pro-Ruling, PO = Pro-Opposition

6 Discussion

We study political discourse of Indian influencers in this paper to understand if they exhibit polarization with respect to various social and political issues. In this direction, we first develop a proxy for their political leaning using the number of political retweets they receive around their tweets. Next, we analyze the stance of their tweets around different aspects, using an SLM fine-tuned on their tweets. Our findings conclusively prove that most Indian influencers are significantly partisan when it comes to their tweets on various aspects. The aggregate partisanism also is dichotomous in nature, wherein there exists two distinct communities of influencers, one in favor of the ruling dispensation and the other against it.

There are several nationally and regionally prominent political parties functioning in India. However, the dichotomy that we observe in the influencer discourse provides concrete indication of the highly bipolar nature of the Indian polity. We see that influencers with a significant political leaning towards the ruling dispensation disproportionately favor issues like *Hindutva, Hinduism, Kashmiri Pandits* and write against *farmer's protests, muslims, bharatjodoyatra*. On the other hand, the pro-opposition influencers follow the exact opposite trend around the same issues. We thus see a clear trend of influencers kotowing their favorite party lines. With the recent trend of increasing religious and political polarization in India [ref], this especially does not augur well. Influencers are highly popular figures, many of whom are revered and closely followed on social media by numerous users. A bipolar nature of the influencer discourse thus amplifies the extant biases in the society, thereby leading to a fragmented democracy.

We also see that these trends are uniformly observed for both English and Hindi tweets, indicative of the fact that the observed polarization holds across languages. In a country where a significant fraction of the social media population consumes tweets in Hindi [ref], this is a sign of a polarized political discourse cutting across linguistic borders.

The topical analysis of tweets reveals ...

This work comes as a timely intervention around influencer polarization in India, provided that several recent studies have targeted similar research questions. Our work acts as a formative step towards large-scale analysis of influencer discourse on social media (from 2020-2023). A major contribution of this study is the development of a generalizable research framework to study aspect based stance on social media. While we study influencer discourse in India corresponding to two languages, the framework is applicable to any geography and can be extended to work on other languages as well. The use of an SLM to perform stance analysis also highlights the need to incorporate low-infrastructure methods to perform large-scale data analysis.

As part of future work, we intend to extend the study to other regional languages in India, to see if similar trends are observed. Furthermore, while the current work focuses on an overall analysis of influencer discourse, a temporal study capturing the evolution of influencer leaning and discourse partisanism can be undertaken. The current study only considers influencer tweets that have received at least one political retweet (retweet by a politician). However, it would be interesting to include other tweets on similar aspects, followed by a stance analysis. A comparative analysis of the trends can then provide us with stronger empirical evidence of partisanism in tweets, irrespective of the political endorsements received by them.

7 Conclusion

In this paper, we examined the political discourse of Indian influencers to assess the extent and nature of polarization across a range of social and political issues. To this end, we first proposed a proxy to infer influencers' political leanings based on the political retweets received in response to their content. We then analyzed the stances expressed in their tweets with respect to different issue-specific aspects, leveraging a small language model fine-tuned on influencer-generated data. Our analysis provides strong empirical evidence that a majority of Indian influencers exhibit pronounced partisanship in their engagement with these issues. Furthermore, polarization at the aggregate level is distinctly dichotomous, revealing the presence of two well-defined communities of influencers—one broadly aligned with the ruling dispensation and the other positioned in opposition. These findings highlight the influential role of content creators in shaping and reinforcing polarized political discourse on social media platforms.

8   Conclusion

work.