BTPD: A Multilingual Hand-curated Dataset of Bengali Transnational Political Discourse Across Online Communities

DIPTO DAS, University of Toronto, Canada
SYED ISHTIAQUE AHMED, University of Toronto, Canada
SHION GUHA, University of Toronto, Canada

Understanding political discourse in online spaces is crucial for analyzing public opinion and ideological polarization. While social computing and computational linguistics have explored such discussions in English, such research efforts are significantly limited in major yet under-resourced languages like Bengali due to the unavailability of datasets. In this paper, we present a multilingual dataset of Bengali transnational political discourse (BTPD) collected from three online platforms, each representing distinct community structures and interaction dynamics. Besides describing how we hand-curated the dataset through community-informed keyword-based retrieval, this paper also provides a general overview of its topics and multilingual content.

1 Introduction

Computer-mediated communication in online communities profoundly shapes contemporary political discourse. Prior computer-supported cooperative work (CSCW) research has studied the discussions of political issues with one's peers in the context of Bengali communities, especially focusing on their postcolonial conditions, decolonial efforts, and intriguing cultural and political dynamics in the Global South [15, 18, 21]. Often dubbed as "adda," such political discourse is a "something quintessentially Bengali, ... an indispensable part"of Bengali practices [9]. The Bengali people are the third largest ethnic group in the world [4], native to South Asia. Through the postcolonial partition of the region, Bengali communities were divided between Bangladesh and India [11], particularly in the states of West Bengal, Assam, and Tripura. Such a colonial formation of transnational dynamics among the Bengali communities makes the political perceptions and discourse historically complex [10]. The region has also become geopolitically important in recent times due to its strategic position in South Asian trade and connectivity, migration dynamics, and being situated within and near emerging global powers (e.g., China, India) and major stakeholders in many global issues (e.g., climate change) [2, 27]. Moreover, with around half a million Bangladeshi Bengalis living in the US[1, 63], as well as many Indian Bengalis who often identify as Indian-Americans, and with growing Bengali communities in major Canadian cities [8], the influence of such large ethnic enclaves on North American politics and the economy is steadily growing [45].

CSCW community often develops datasets of computer-mediated political discussions and empirically studies those interactions [32, 62, 66]. As contemporary political discourse among the Bengalis frequently takes place online [15], a dataset of their political discourse would enable CSCW researchers to study its transnational dynamics, complexities, and significance. However, despite Bengali being the sixth-largest native language [42] and having a strong web presence, there exist fewer resources in different linguistic data sites and consortia for Bengali than for other major languages [40]. Given their diverse backgrounds, Bengali communities adopt different online platforms based on platform-specific affordances, i.e., perceived and actual interaction possibilities, and tailor their political discussions accordingly. In this paper, we develop a dataset[1] of Bengali transnational political discourse (BTPD) with multilingual support, collected from three online platforms of different community structures and objectives, namely Reddit, Politics Stack Exchange[2], and Bengali Quora[3]. In the following sections, we will explain ways to conceptualize different types of online communities and review existing research, outline our data collection, preprocessing, and organization strategies, and describe the dataset using natural language processing (NLP) methods.

[1] Publicly available upon the paper's acceptance. [2] https://politics.stackexchange.com/ [3] https://bn.quora.com/

2 Literature Review

While most existing research on political discussions online focuses on online platforms, such as Twitter and Facebook [14, 34, 39] that are typically understood as "social media," Bruckman argues that these online communities should be viewed as a prototype-based category [6], defined not by rigid inclusion and exclusion rules, but by their prototypical members. Though social media platforms are more representative of online communities, which are often riddled with political misinformation [65], Question-and-Answer (Q&A) platforms, which usually offer better information quality with adequate support for fostering connections among users [26], can also be viewed as online communities. Moreover, the degree to which a platform embodies the prototypicality of a community can be viewed as a cultural construct [6]. For example, while prior studies on Quora focused on collective wisdom, reputation, quality of answers–objectives that are typical for Q&A sites [52, 53, 67], Das and colleagues demonstrated how the Bengali users from Bangladesh and India fostered transnational communities based on their linguistic and cultural similarities and participated in sociopolitical discussions through this platform [19, 21]. While most political discourse datasets rely on news sources or single platforms [49], that may not reflect the holistic online political discourse. Considering that different platforms facilitate user interaction differently, any datasets on political discourse should be curated across multiple sites.

Similar to many other fields adjacent to CSCW, such as human-computer interaction (HCI), NLP, and algorithmic fairness [40, 44, 58], most research and resources in computational social science, for example, in studying political discourse, overwhelmingly focus on the Global North contexts [7, 12, 23, 33, 38, 43]. With recent studies focused on discourse in the Global South, countries like India, Brazil, Indonesia, and Nigeria [25, 47, 48, 50] have facilitated cross-cultural analyses of global political participation [41, 51]. However, there exists a dearth of datasets for studying the transnational Bengali ethnolinguistic communities in Bangladesh and India. While most NLP datasets in this under-resourced language have focused on tasks like sentiment analysis and hate speech detection [22, 54, 56], some recent datasets have focused on bias evaluation [16, 17], Q&A [60], machine translation [31], etc. Though some recent single platform-sourced datasets of public opinion in Bengali exist [13, 30], they primarily feature product reviews or discussions on global events rather than political discourse directly relevant to the Bengali people. Moreover, most Bengali datasets are shaped by the construct of nationality, framing it either as a language specific to Bangladesh or as a regional language in India. This paper seeks to address this gap by foregrounding the transnational Bengali political discourse in Bangladesh and India. Following recommendations in human-centered data science [3, 36] and common practices in HCI and CSCW [20, 61], we hand-curated the dataset from multiple online communities through keyword search.

3 Dataset Creation

3.1 Choice of Platforms

Drawing on Bruckman's argument that community is best understood as a prototype-based category [6], we chose three online platforms that exhibit varying degrees of different prototypicality as communities. For our paper on preparing a corpus of Bengali political discussions online, we collected political discussions from Reddit, the Politics Forum on Stack Exchange (PoliticsSE), and Bengali Quora (BnQuora). Among these, Reddit aligns most closely with traditional notions of community due to its persistent user identities, subreddit-based governance structures, and ongoing interactions centered around shared interests [68]. In contrast, though users build reputations and some expert-driven communities develop around specific topics on Stack Exchange, this platform prioritizes high-quality information exchange through structured Q&A rather than sustained interaction for social bonding [55] and, thus, is a less prototypical example of a community. Compared to these two, BnQuora falls somewhere in between. Though it has a looser sense of community compared to Reddit, as discussions in Q&A threads are more individual-driven than group-based, [19] have found its effectiveness in fostering a sense of social relationship among Bengali users from different regions based on their cultural similarities, reinforcing Bengali ethnolinguistic identity, and facilitating political discourses.

3.2 Data Collection

We collected data from these platforms through keyword searches. The list of keywords and data collection process varied across platforms based on their technical scaffolds and topical focus.

3.2.1 Reddit. Reddit facilitates decentralized discussions through subreddits, which are often geographically anchored or based on similar cultures and interests. Therefore, we could look for subreddits related to politics and Bengali contexts. Since most politics-related subreddits (e.g., `r/politics`) are US-centered or strongly guided by US-adjacency (e.g., `r/Ask_Politics`) as found by [28], to collect posts on Bengali politics, following [21], we included the subreddits related to the geographic regions where Bengali people live (`r/bangladesh`, `r/westbengal`) and their political centers (`r/Dhaka`, `r/kolkata`) and their shared linguistic backgrounds (`r/bengalilanguage`) as communities where discussions on politics in Bengali social contexts are likely to occur,. These subreddits are moderated and use flairs (e.g., “Seeking advice/পরামর্শ”) that indicate the type and topic of the content. For our data collection, we looked for posts in those subreddits that used the flairs: “Politics/রাজনীতি”, “Discussion/আলোচনা,” and “News/সংবাদ.”

We used the Python Reddit API Wrapper (PRAW) to collect data from January 25, 2025, to February 17, 2025. This period reflects the code execution time, not the posting times of the posts. We collected the posts’ titles, URLs, bodies, flair, times of posting, and comments. While Reddit employs a nested branching structure for comments, we stored the comments as a flat list. Based on our long membership in the previously mentioned subreddits, we have observed certain differences in how flairs are used in various subreddits. After data collection, we similarly noticed how different subreddits used flair more or less frequently to indicate political posts and how the same flair in various subreddits resulted in differing numbers of political posts as well as posts unrelated to politics. For example, political posts in `r/bangladesh` often bear the flair ”Politics/রাজনীতি”, whereas `r/westbengal` uses the flair “News/সংবাদ” and uses the flair “Politics/রাজনীতি” less frequently. In both subreddits, the flair “Discussion/আলোচনা” is used in posts related to politics as well as other topics. Hence, to keep the corpus relevant to Bengali political discussions, we excluded posts on other topics (e.g., posts bearing the “Discussion/আলোচনা” flair but focusing on different topics). Table 1 lists the subreddits and numbers of members and collected posts.

Table 1. Subreddits and their number of members (top x% of largest communities on Reddit) and political posts from there included in our dataset.

Subreddit	#Members	#Political posts
`r/bangladesh`	75K (2%)	601
`r/westbengal`	5.5K (10%)	206
`r/Dhaka`	54K (3%)	309
`r/kolkata`	331K (1%)	49
`r/bengalilanguage`	26K (4%)	55

3.2.2 PoliticsSE. In contrast to the diverse topics discussed in our selected subreddits, PoliticsSE is a Q&A forum solely for political discourse, ensuring the inherent topical relevance of its posts. As such, for our data collection, we can prioritize the contextual relevance of the data to Bengali communities rather than concerns about the broader topical focus. We retrieved PoliticsSE’s latest data dump from the Internet Archive, which includes data from the platform's launch until December 31, 2024. As before, we used the keywords mentioning the regions where the Bengali people are native (e.g., `Bangladesh`, `West Bengal`), their political centers (e.g., `Dhaka`, `Kolkata`) and their language and community name (`Bengali`) to identify posts on PoliticsSE related to the context of Bengali communities based on their titles, bodies, and tags. We manually read through the posts and only retained the unique posts while excluding the ones not directly related to political discussion in the Bengali context (e.g., posts that mention Bangladesh as a passing example while generally discussing different parliamentary structures around the world). We also retained their metadata, such as URLs and posting time. Table 2 shows the number of posts identified using keywords from the PoliticsSE data dump and the ones relevant to Bengali politics.

Table 2. Keywords, number of posts mentioning those keywords, and number of posts relevant to Bengali political discussion among those identified posts.

Keyword	**#Posts identified through keywords**	**#Posts identified as relevant**
`Bangladesh`	234	209
`West Bengal`	14	14
`Dhaka`	10	10
`Kolkata`	5	5
`Bengali`	19	17

3.2.3 BnQuora. Quora's Q&A structure, which encourages diverse perspectives on a given topic, fosters in-depth discussions on controversial subjects, including politics [35, 64]. Similarly, as described in the previous section, BnQuora provides a unique space for in-depth analysis of Bengali political discourse without significant concerns about the contextual or broader topical relevance. Hence, instead of searching with keywords on the broad topic (e.g., politics), Das et al. [20] recommended using more specific terms related to the broader topic to identify Q&A threads to collect data from BnQuora. We conducted a Qualtrics survey to know what specific topics are crucial to contemporary Bengali political discourse. The survey presented common political discussion points [46] as options while allowing participants to add unlisted responses. We circulated the survey through our social networks as members of Bengali communities in Bangladesh and West Bengal and through Bangladeshi, Indian, and South Asian student organizations at two North American universities. We thematically consolidated the 74 responses received between October 5 and 21, 2024, into a list of key topics/themes in Bengali political discussions and used the ten most prominent ones to collect our dataset. Besides these topics, we used other related keywords (see Table 3) to search for Q&A threads on BnQuora. We periodically ran a Python script using Selenium from November 1, 2024, to February 15, 2025, to automate browser interactions (e.g., refresh, scroll) to manage dynamic page content, which collected Q&A threads containing those keywords.

3.3 Data Preprocessing and Organization

Our collected data from PoliticsSE and BnQuora are primarily in English and Bengali, respectively, with occasional use of the other language for certain terms or phrases. However, the languages of Reddit data vary significantly, including Bengali and English, with occasional code-switching and Romanized Bengali, i.e., phonetic Bengali using English fonts. We translated all collected posts in Bengali and English using OpenAI's API with the GPT-4 engine, which is comparable to commercial translation products [37], using the following prompts: *"You are a translator who can translate* `{Bengali/English}` *and Banglish (Bengali in romanized fonts) to* `{English/Bengali}`*."* In our capacity as natively Bengali-speaking researchers, we also manually verified and fixed the translations if needed. Hence, for each unique post ID, besides the original post, which may have used a mix of languages, we have its translations in Bengali and English. This makes our

Table 3. Key topics of Bengali political discussions, related keywords, and number of Q&A threads collected.

Topical Themes	**Keywords**	**#Q&A threads**
foreign policy	পররাষ্ট্র নীতি	140
constitution	সংবিধান	246
secularism	ধর্মনিরপেক্ষতা	57
public education	সরকারি শিক্ষাক্রম	18
cultural identity	বাঙালি, বাংলাদেশি, বাংলাদেশী, ভারতীয়, ইন্ডিয়ান	87
LGBTQ+ rights	সমকামী/ট্রান্সজেন্ডার অধিকার	42
political parties	আওয়ামী লীগ, জামায়াতে ইসলাম, তৃণমূল কংগ্রেস, বিএনপি, বিজেপি	119
religion	ধর্ম, মুসলিম, ইসলাম, হিন্দু, ধর্মীয় সংখ্যালঘু	29
women's rights	নারী অধিকার	131
ethnic minorities	আদিবাসী, ক্ষুদ্র নৃতাত্ত্বিক জনগোষ্ঠী	7

dataset uniform and multilingual, including a total of 2235 posts' original titles and bodies, their translations in Bengali and English, answers and comments, posting time, and tags, if available. Though the concern of misinformation is often intensified in political discussions and our dataset comes from online communities with varied information quality, where StackExchange platforms like PoliticsSE are seen as reliable [55] but Reddit has documented misinformation issues [5, 59], not screening for misinformation while including a post in our dataset allowed BTPD to stay true to the dynamics and reflect the nature of political discussions in Bengali communities online.

4 Dataset Content

In this section, we provide a brief descriptive overview of our developed dataset. After preprocessing (e.g., excluding stopwords, stemming), Table 4 shows that lengths (average and median) and timestamps of the earliest and the latest posts in our dataset varied significantly across platforms.

Table 4. Overview of the collected data by platforms.

Platform	**#Sentences**	**#Words**	**Earliest and latest posts**
Reddit	37.1, 15.0	198.0, 99.5	2023/01/02, 2025/02/17
PoliticsSE	16.2, 10.0	130.3, 90.0	2012/12/13, 2024/12/22
BnQuora	23.7, 12.0	213.4, 109.0	5 years ago, 2025/02/15

As described earlier, our multilingual dataset includes the Bengali and English versions of each post. To compare the variances in the Bengali and English versions of the posts, we used principal component analysis (PCA) on their TF-IDF (Term Frequency-Inverse Dense Frequency) vectors. Examining the differing elbows in the scree plot (see Figure 1(a)), we can see that the number of principal components needed to retain a fixed proportion of variance (e.g., 80%) varies across languages–for instance, approximately the first 500 for Bengali and 1000 for English.

Figure 1(b) shows the common words appearing in our dataset using a wordcloud. Though we addressed language-specific characteristics (e.g., Bengali's bidirectional structure) and provided Unicode fonts, the existing NLP tools could not visualize the Bengali wordcloud properly. We conducted topic modeling of the titles and bodies of the posts to get an overview of the broad topics included in our dataset. Given the lack of enough evidence of how common topic modeling approaches like latent Dirichlet allocation (LDA) and non-negative matrix factorization (NMF) work in Bengali, we tried to identify topics through clustering of the posts based on their sentence embeddings but did not find this approach informative. Since NMF works better than other common approaches like LDA for topic modeling of short texts [24, 29], in Table 5, we report ten topics identified by NMF on the posts' English translations with each topic's corresponding top five words.

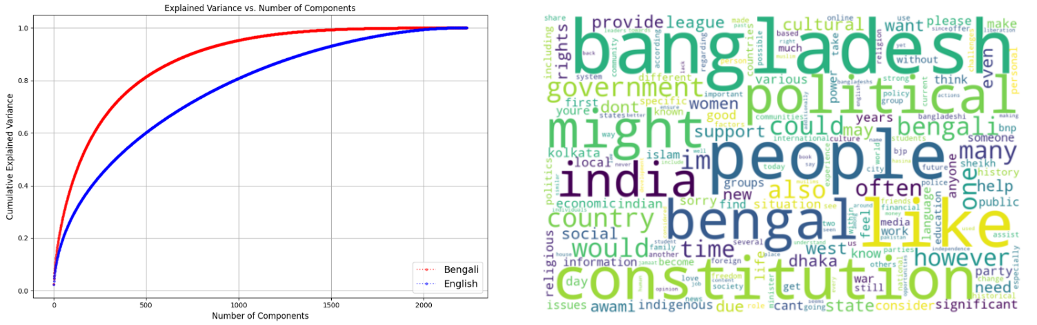

Fig. 1. (a) Scree plots of principal component analysis of Bengali (in red) and English (in blue) versions of the posts in our dataset (b) Common words in the English translations of the posts

Table 5. Topics identified in the English versions of the posts by NMF with common words.

Topic	Words	Topic	Words
0	assist, sorry, request, information, content	1	country, like, people, Awami-League, time
2	constitution, according, written, country, Indian	3	West-Bengal, chief-minister, BJP, Mamata-Banerjee, state
4	India, foreign-policy, Dr-Ambedkar, Hindu, draft	5	Indigenous, people, communities, tribes, Bengalis
6	provide, text, translation, information, need	7	women-rights, men, Islam, equal, freedom
8	Bengali, Trinamool, Congress, BJP, parties	9	Bangladesh, secularism, Pakistan, war, prime-minister

While some of these topics (e.g., 0, 6) are generic, some topics closely relate to particular domains of Bengali political discourse. For example, topics 3 and 8 focus on West Bengal's state-level politics in India, whereas topic 9 covers Bangladesh's historical political issues, and topic 5 deals with the politics around settler Bengalis and the Indigenous and ethnic tribes in Bangladesh. Interestingly, topic 7 seems to engage closely with equality of rights and freedom across different genders in Islam. Overall, topics 4, 7, and 9 highlight the centrality of religion and caste to politics in Bengal by mentioning words like Islam, secularism, Hindu, and Dr. Ambedkar (an Indian social reformer with great contributions in alleviating underprivileged castes, who, being elected from the Bengal region, chaired the Indian constitution drafting committee [57]). While some of the top words shown in Table 5 are identical to words used for keyword-based search, NMF surfaced more important keywords and identified connections among the words that were not implied during data collection.

5 Conclusion

This poster follows traditional NLP strategies while being informed by CSCW and social computing scholarship in considering different prototypical examples of online communities. It develops a textual corpus of transnational Bengali political discussions, which would address a resource need in one of the major global languages and be useful for examining cross-platform information dynamics and cultural and longitudinal shifts in political discourse in one of the largest global ethnolinguistic communities. While future research should contribute more data instances to BTPD from other online communities, include additional metadata like fact-checking labels, and link the online discussions with reliable sources, this artifact would facilitate political deliberation among Bengali communities and critical algorithmic audits of political biases of Bengali NLP systems, such as large language models, automated content moderation, and recommendation systems.

References

[1] Morsheda Akhter and Philip Q Yang. 2023. The Bangladeshi Diaspora in the United States: History and Portrait. *Genealogy* 7, 4 (2023), 81.

[2] Ashlyn Anderson and Alyssa Ayres. 2015. Economics of Influence: China and India in South Asia. https://www.cfr.org/expert-brief/economics-influence-china-and-india-south-asia. Last accessed: 27-03-2025.

[3] Cecilia Aragon, Shion Guha, Marina Kogan, Michael Muller, and Gina Neff. 2022. *Human-centered data science: an introduction*. MIT Press.

[4] World Atlas. 2022. Who are the Bengali People? https://www.worldatlas.com/society/the-10-most-spoken-languages-in-the-world.html. Last accessed: March 21, 2025.

[5] Lia Bozarth, Jane Im, Christopher Quarles, and Ceren Budak. 2023. Wisdom of Two Crowds: Misinformation Moderation on Reddit and How to Improve this Process—A Case Study of COVID-19. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–33.

[6] Amy Bruckman. 2006. A new perspective on" community" and its implications for computer-mediated communication systems. In *CHI'06 extended abstracts on Human factors in computing systems*. 616–621.

[7] Pete Burnap, Rachel Gibson, Luke Sloan, Rosalynd Southern, and Matthew Williams. 2016. 140 characters to victory?: Using Twitter to predict the UK 2015 General Election. *Electoral Studies* 41 (2016), 230–233.

[8] Statistics Canada. 2021. Census of Population. https://www12.statcan.gc.ca/census-recensement/2021/dp-pd/prof/details/page.cfm. Accessed: 17-02-2025.

[9] Dipesh Chakrabarty. 2009. Provincializing Europe: postcolonial thought and historical difference-New edition. (2009).

[10] Partha Chatterjee. 1993. *The nation and its fragments: Colonial and postcolonial histories*. Vol. 4. Princeton University Press.

[11] Joya Chatterji. 2002. *Bengal divided: Hindu communalism and partition, 1932-1947*. Number 57. Cambridge University Press.

[12] Emily Chen, Ashok Deb, and Emilio Ferrara. 2022. # Election2020: the first public Twitter dataset on the 2020 US Presidential election. *Journal of Computational Social Science* (2022), 1–18.

[13] Lomat Haider Chowdhury, Salekul Islam, and Swakkhar Shatabda. 2024. A Bengali news and public opinion dataset from YouTube. *Data in Brief* 52 (2024), 109938.

[14] Isobelle Clarke and Jack Grieve. 2019. Stylistic variation on the Donald Trump Twitter account: A linguistic analysis of tweets posted between 2009 and 2018. *PloS one* 14, 9 (2019), e0222062.

[15] Dipto Das, Dhwani Gandhi, and Bryan Semaan. 2024. Reimagining Communities through Transnational Bengali Decolonial Discourse with YouTube Content Creators. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW2 (2024), 1–36.

[16] Dipto Das, Shion Guha, Jed R Brubaker, and Bryan Semaan. 2024. The"Colonial Impulse" of Natural Language Processing: An Audit of Bengali Sentiment Analysis Tools and Their Identity-based Biases. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–18.

[17] Dipto Das, Shion Guha, and Bryan Semaan. 2023. Toward cultural bias evaluation datasets: The case of Bengali gender, religious, and national identity. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*. 68–83.

[18] Dipto Das, AKM Najmul Islam, SM Taiabul Haque, Jukka Vuorinen, and Syed Ishtiaque Ahmed. 2022. Understanding the Strategies and Practices of Facebook Microcelebrities for Engaging in Sociopolitical Discourses. In *Proceedings of the 2022 International Conference on Information and Communication Technologies and Development*. 1–19.

[19] Dipto Das, Carsten Østerlund, and Bryan Semaan. 2021. " Jol" or" Pani"?: How Does Governance Shape a Platform's Identity? *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–25.

[20] Dipto Das, Arpon Podder, and Bryan Semaan. 2022. Note: A sociomaterial perspective on trace data collection: Strategies for democratizing and limiting bias. In *Proceedings of the 5th ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies*. 569–573.

[21] Dipto Das and Bryan Semaan. 2022. Collaborative identity decolonization as reclaiming narrative agency: Identity work of Bengali communities on Quora. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–23.

[22] Mithun Das and Animesh Mukherjee. 2023. Banglaabusememe: A dataset for bengali abusive meme classification. *arXiv preprint arXiv:2310.11748* (2023).

[23] Maryam Davoodi, Eric Waltenburg, and Dan Goldwasser. 2020. Understanding the language of political agreement and disagreement in legislative texts. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 5358–5368.

[24] Roman Egger and Joanne Yu. 2022. A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. *Frontiers in sociology* 7 (2022), 886498.

[25] Asno Azzawagama Firdaus, Anton Yudhana, Imam Riadi, et al. 2024. Indonesian presidential election sentiment: Dataset of response public before 2024. *Data in Brief* 52 (2024), 109993.

[26] Adabriand Furtado, Nazareno Andrade, Nigini Oliveira, and Francisco Brasileiro. 2013. Contributor profiles, their dynamics, and their importance in five q&a sites. In *Proceedings of the 2013 conference on Computer supported cooperative work*. 1237–1252.

[27] Kristalina Georgieva. 2023. Bangladesh and its Partners are Launching the Bangladesh Climate and Development Platform to Leverage Adaptation and Mitigation Investments. https://www.imf.org/en/News/Articles/2023/12/03/bangladesh-launch-climate-development-platform-to-leverage-adaptation-and-mitigation-investments. Last accessed: 27-03-2025.

[28] Sarah A Gilbert. 2020. " I run the world's largest historical outreach project and it's on a cesspool of a website." Moderating a Public Scholarship Site on Reddit: A Case Study of r/AskHistorians. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–27.

[29] Astha Goyal and Indu Kashyap. 2023. Comprehensive Analysis of Topic Models for Short and Long Text Data. *International Journal of Advanced Computer Science & Applications* 14, 12 (2023).

[30] Mahmud Hasan, Labiba Islam, Ismat Jahan, Sabrina Mannan Meem, and Rashedur M Rahman. 2023. Natural language processing and sentiment analysis on bangla social media comments on russia–ukraine war using transformers. *Vietnam Journal of Computer Science* 10, 03 (2023), 329–356.

[31] Tahmid Hasan, Abhik Bhattacharjee, Kazi Samin, Masum Hasan, Madhusudan Basak, M Sohel Rahman, and Rifat Shahriyar. 2020. Not low-resource anymore: Aligner ensembling, batch filtering, and new datasets for Bengali-English machine translation. *arXiv preprint arXiv:2009.09359* (2020).

[32] Libby Hemphill, Jahna Otterbacher, and Matthew Shapiro. 2013. What's congress doing on twitter?. In *Proceedings of the 2013 conference on Computer supported cooperative work*. 877–886.

[33] Brian Heredia, Joseph D Prusa, and Taghi M Khoshgoftaar. 2018. Location-based twitter sentiment analysis for predicting the US 2016 presidential election. In *The Thirty-First International Flairs Conference*.

[34] Tunazzina Islam, Shamik Roy, and Dan Goldwasser. 2023. Weakly supervised learning for analyzing political campaigns on facebook. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 17. 411–422.

[35] Sou Hyun Jang, Sangpil Youm, and Yong Jeong Yi. 2023. Anti-Asian discourse in Quora: Comparison of before and during the COVID-19 pandemic with machine-and deep-learning approaches. *Race and Justice* 13, 1 (2023), 55–79.

[36] Mohammad Hossein Jarrahi, Ali Memariani, and Shion Guha. 2023. The principles of data-centric ai. *Commun. ACM* 66, 8 (2023), 84–92.

[37] Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Is ChatGPT a good translator? Yes with GPT-4 as the engine. *arXiv preprint arXiv:2301.08745* (2023).

[38] Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, Yu Wang, and Jiebo Luo. 2017. Detection and analysis of 2016 us presidential election related rumors on twitter. In *Social, Cultural, and Behavioral Modeling: 10th International Conference, SBP-BRiMS 2017, Washington, DC, USA, July 5-8, 2017, Proceedings 10*. Springer, 14–24.

[39] Kristen Johnson and Dan Goldwasser. 2016. "all I know about politics is what I read in twitter": Weakly supervised models for extracting politicians' stances from twitter. In *Proceedings of COLING 2016, the 26th international conference on computational linguistics: technical papers*. 2966–2977.

[40] Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 6282–6293. doi:10.18653/v1/2020.acl-main.560

[41] Jaehong Kim, Chaeyoon Jeong, Seongchan Park, Meeyoung Cha, and Wonjae Lee. 2024. How Do Moral Emotions Shape Political Participation? A Cross-Cultural Analysis of Online Petitions Using Language Models. In *Findings of the Association for Computational Linguistics ACL 2024*. 16274–16289.

[42] James Lane. 2023. The 10 Most Spoken Languages In The World. https://www.babbel.com/en/magazine/the-10-most-spoken-languages-in-the-world. Last accessed: Feb 26, 2023.

[43] Chang Li and Dan Goldwasser. 2021. Using social and linguistic information to adapt pretrained representations for political perspective identification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 4569–4579.

[44] Sebastian Linxen, Christian Sturm, Florian Brühlmann, Vincent Cassau, Klaus Opwis, and Katharina Reinecke. 2021. How weird is CHI?. In *Proceedings of the 2021 chi conference on human factors in computing systems*. 1–14.

[45] Michael Liu and Kim Geron. 2008. Changing neighborhood: Ethnic enclaves and the struggle for social justice. *Social Justice* 35, 2 (112 (2008), 18–35.

[46] Ruibo Liu, Chenyan Jia, Jason Wei, Guangxuan Xu, and Soroush Vosoughi. 2022. Quantifying and alleviating political bias in language models. *Artificial Intelligence* 304 (2022), 103654.

[47] Adenike Tosin Odegbile and Olufemi Moses Oyelami. 2024. A dataset of the 2023 presidential election in Nigeria. *Data in Brief* 57 (2024), 110847.

[48] Lucas Oliveira, Pedro Vaz de Melo, Marcelo Amaral, and José Antônio Pinho. 2018. When politicians talk about politics: Identifying political tweets of Brazilian congressmen. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 12.

[49] Helene Olsen, Étienne Simon, Erik Velldal, and Lilja Øvrelid. 2024. Socio-political events of conflict and unrest: A survey of available datasets. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024)*. 40–53.

[50] Joyojeet Pal and Anmol Panda. 2019. Twitter in the 2019 Indian general elections: Trends of use across states and parties. *Economic and Political Weekly* 54, 51 (2019), 1–17.

[51] Anmol Panda, A'ndre Gonawela, Sreangsu Acharyya, Dibyendu Mishra, Mugdha Mohapatra, Ramgopal Chandrasekaran, and Joyojeet Pal. 2020. Nivaduck-a scalable pipeline to build a database of political twitter handles for india and the united states. In *International Conference on Social Media and Society*. 200–209.

[52] Sumanth Patil and Kyumin Lee. 2016. Detecting experts on Quora: by their activity, quality of answers, linguistic characteristics and temporal behaviors. *Social network analysis and mining* 6 (2016), 1–11.

[53] Sharoda A Paul, Lichan Hong, and Ed H Chi. 2012. Who is authoritative? understanding reputation mechanisms in quora. *arXiv preprint arXiv:1204.3724* (2012).

[54] Nauros Romim, Mosahed Ahmed, Hriteshwar Talukder, and Md Saiful Islam. 2021. Hate speech detection in the bengali language: A dataset and its baseline evaluation. In *Proceedings of International Joint Conference on Advances in Computational Intelligence: IJCACI 2020*. Springer, 457–468.

[55] Tiago Santos, Simon Walk, Roman Kern, Markus Strohmaier, and Denis Helic. 2019. Activity archetypes in question-and-answer (q8a) websites—a study of 50 stack exchange instances. *ACM Transactions on Social Computing* 2, 1 (2019), 1–23.

[56] Salim Sazzed. 2020. Cross-lingual sentiment classification in low-resource Bengali language. In *Proceedings of the sixth workshop on noisy user-generated text (W-NUT 2020)*. 50–60.

[57] Dwaipayan Sen. 2018. *The decline of the caste question: Jogendranath Mandal and the defeat of Dalit politics in Bengal.* Cambridge University Press.

[58] Ali Akbar Septiandri, Marios Constantinides, Mohammad Tahaei, and Daniele Quercia. 2023. WEIRD FAccTs: How Western, Educated, Industrialized, Rich, and Democratic is FAccT?. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 160–171.

[59] Vinay Setty and Erlend Rekve. 2020. Truth be told: Fake news detection using user reactions on reddit. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 3325–3328.

[60] Md Shihab Shahriar, Ahmad Al Fayad Chowdhury, Md Amimul Ehsan, and Abu Raihan Kamal. 2023. Question Answer Generation in Bengali: Mitigating the scarcity of QA datasets in a low-resource language. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. 430–441.

[61] Yuya Shibuya, Andrea Hamm, and Teresa Cerratto Pargman. 2022. Mapping HCI research methods for studying social media interaction: A systematic literature review. *Computers in Human Behavior* 129 (2022), 107131.

[62] Kate Starbird and Leysia Palen. 2012. (How) will the revolution be retweeted? Information diffusion and the 2011 Egyptian uprising. In *Proceedings of the acm 2012 conference on computer supported cooperative work*. 7–16.

[63] Moshahida Sultana. 2005. *Do migrants transfer tacit knowledge?: the case of highly skilled Bangladeshi immigrants in the United States.* Ph. D. Dissertation. Massachusetts Institute of Technology.

[64] David Q Sun, Artem Abzaliev, Hadas Kotek, Zidi Xiu, Christopher Klein, and Jason D Williams. 2023. DELPHI: Data for Evaluating LLMs' Performance in Handling Controversial Issues. *arXiv preprint arXiv:2310.18130* (2023).

[65] Joshua A Tucker, Andrew Guess, Pablo Barberá, Cristian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal, and Brendan Nyhan. 2018. Social media, political polarization, and political disinformation: A review of the scientific literature. *Political polarization, and political disinformation: a review of the scientific literature (March 19, 2018)* (2018).

[66] Morgan Vigil-Hayes, Marisa Duarte, Nicholet Deschine Parkhurst, and Elizabeth Belding. 2017. # indigenous: tracking the connective actions of native American advocates on twitter. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*. 1387–1399.

[67] Gang Wang, Konark Gill, Manish Mohanlal, Haitao Zheng, and Ben Y Zhao. 2013. Wisdom in the social crowd: an analysis of quora. In *Proceedings of the 22nd international conference on World Wide Web*. 1341–1352.

[68] Galen Weld, Amy X Zhang, and Tim Althoff. 2024. Making online communities 'better': a taxonomy of community values on reddit. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 18. 1611–1633.