

The Koo Dataset: An Indian Microblogging Platform with Global Ambitions

Amin Mekacher¹, Max Falkenberg¹, Andrea Baronchelli^{1, 2}

¹City University of London, Department of Mathematics, London EC1V 0HB, (UK)

²The Alan Turing Institute, British Library, London NW1 2DB, (UK)

amin.mekacher@city.ac.uk, max.falkenberg@city.ac.uk, abaronchelli@turing.ac.uk

Abstract

Increasingly, alternative platforms are playing a key role in the social media ecosystem. Koo, a microblogging platform based in India, has emerged as a major new social network hosting high profile politicians from several countries (India, Brazil, Nigeria) and many internationally renowned celebrities. This paper presents the largest publicly available Koo dataset, spanning from the platform's founding in early 2020 to September 2023, providing detailed metadata for 72M posts, 75M comments, 40M shares, 284M likes and 1.4M user profiles. Along with the release of the dataset, we provide an overview of the platform including a discussion of the news ecosystem on the platform, hashtag usage, and user engagement. Our results highlight the pivotal role that new platforms play in shaping online communities in emerging economies and the Global South, connecting local politicians and public figures with their followers. With Koo's ambition to become the town hall for diverse non-English speaking communities, our dataset offers new opportunities for studying social media beyond a Western context.

Introduction

For many years, online social interactions have taken place primarily on well-established, US-based platforms such as Facebook, Twitter or Instagram. Consequently, significant research efforts have gone towards studying these platforms, for example considering their role in political campaigns (Punia 2016; McGregor 2020), and the digital news ecosystem (Picard 2014). However, in recent years the social media landscape has evolved with the advent of the alt-tech ecosystem. These emerging platforms are becoming increasingly influential venues in the political communication ecosystem, attracting key right-wing politicians such as former US president Donald Trump and former Brazilian president Jair Bolsonaro (Allyn 2022; Ribeiro and Mendes 2022), hosting major debates (Gilbert 2023) and functioning as a political campaign ground prior to presidential elections (Leingang 2023).

Despite the growing importance of alt-tech platforms, research remains in its infancy in part due to a lack of publicly available data. This has motivated the release of datasets for many platforms including Gab (Fair and Wesslen 2019) ,

BitChute (Trujillo et al. 2022), Truth Social (Gerard, Botzer, and Weninger 2023) and Voat (Mekacher and Papasavva 2022). However, these datasets overwhelmingly focus on platforms based in the West, catering largely to fringe, right-wing audiences. Launched in India in 2020, Koo has grown to become the second largest microblogging platform worldwide (Singh 2022), and has attracted a number of political communities who have been critical of Twitter for censoring their discourse (Bhat 2021). Most notably, Koo has become a leading social platform used by the Bharatiya Janata Party (BJP), the governing political party in India.

Given the dominance of the BJP on Koo, the platform appears to be similar to many other alt-tech platforms with a focus on right-wing politics. However, beyond India, Koo has successfully attracted a politically diverse set of users, including both supporters and opponents of former Nigerian president Buhari, and of Brazilian president Lula. This success highlights how Koo has the potential to move beyond the politically homogeneous model of most alt-tech platforms to become a major politically heterogeneous venue for both political and apolitical discussions across multiple countries, challenging the US social media hegemony.

Data Release. In this work, we present the most extensive Koo dataset currently available, including 71.7M posts, 74.6M comments, 283.5M likes, 40.0M shares, and 1.4M user profiles. Our dataset spans from Koo's launch in early 2020 up until September 2023. Due to the ability to paginate through a user's complete timeline to access historical activity via the API, we are confident that our dataset provides a near-complete overview of the interactions carried out by users who are actively using the platform.

Relevance. Our dataset provides researchers with the opportunity to study an alt-tech platform based in India which has attracted an international community of high-profile politicians and celebrities, primarily from India, Brazil and Nigeria. With BJP politicians heavily endorsing Koo, our dataset enables the study of political rhetoric through the analysis of the political content shared online. Previous research has shown that the BJP promote islamophobic (Werleman 2021) and populist (Ammassari, Fossati, and McDonnell 2022) rhetoric. It is therefore of interest to consider whether their claims on Koo aim to polarize opinions and stoke hate in their audience. Moreover, with many news outlets shared on Koo, the dataset will enable the study of po-

litical discourses through an analysis of the alignment between news media outlets’ editorial line and the national political parties’ ideologies (Marques, Mont’ Alverne, and Mitozo 2019; Barclay, Pichandy, and Venkat 2014). Finally, the dataset can contribute to cross-platform comparisons of political rhetoric.

Paper organization. In this paper, we first explain Koo’s structure before detailing its impact on the social media ecosystem. Second, we describe the method used to query posts, interactions and user profiles from Koo’s public API before describing the structure of our dataset. Finally, we present results which provide an overview of the Koo platform, before discussing our results and the importance of our dataset in the conclusion.

What is Koo?

Koo is a multi-lingual microblogging platform launched in early 2020 by Bombinate Technologies, a company based in Bangalore, India. Social interactions on Koo work in a very similar way as interactions on other microblogging platforms (e.g., Twitter): Users can create an account and then follow, or be followed by, other users. User profiles can be personalized with a profile-picture and additional information including a user description, personal title, and links to other social platforms. User profiles also display the account username and creation date.

Once logged in, users can submit public posts, called *koos*. A user can like a post, comment on it, or share it (a *reko*). Koo also provides the ability to translate a koo to many languages. Since March 2023, the platform has integrated ChatGPT in a bid to increase the number of users actively creating content for the platform (Dang 2023).

Figure 1 provides an example of a post on Koo. Posts have a 500 character limit and may include images and videos. The panel under each post indicates how many times the koo has been commented, liked and shared by other users, and allows the viewer to share the koo on other platforms. The profile panel, located above the koo, displays the user’s username and their self-indicated title. The yellow tick at the right of the username, known as the yellow tick of eminence, is granted by Koo administrators to accounts considered “significant representative[s] of the Voices of the World” (Tech 2022b). Other users can earn a green badge by self-verifying their account.

The Demographics of Koo

Koo aims at becoming a central outlet for non-English speaking communities online by targeting countries where the uptake of traditional social media has been low (Phartiyal 2021). The platform has experienced three key surges in user registrations. First, in February 2021 Indian users joined Koo following an open conflict between Narendra Modi’s cabinet and Twitter, after the government ordered Twitter to remove more than 1,000 accounts they alleged were responsible for spreading misinformation around the Indian farmers’ protests (Phartiyal 2021). Many government officials and celebrities endorsed the platform in the wake of the conflict, including the chief minister of Uttar

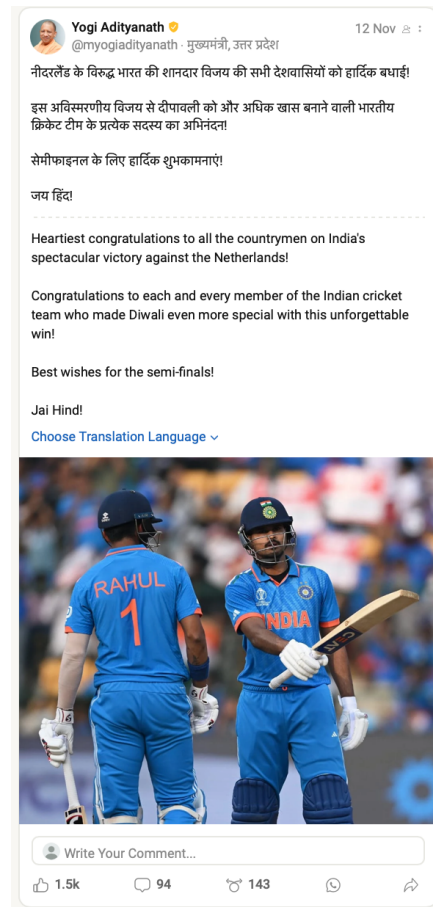


Figure 1: An example of a koo. The main panel includes the original post in Hindi, its translation in English and an image. The top panel provides information about the poster, including their user handle, profile picture, their self-declared title, the yellow tick of eminence (if applicable), and the post creation date. The bottom panel allows logged-in users to comment, share or like the post. Koo provides additional icons to share posts on other platforms.

Pradesh Yogi Adityanath and cricket player Virat Kohli. Indian Prime Minister Narendra Modi has expressed his support for Koo but does not currently use the platform (Majumdar 2021). Previous research has highlighted that about 80% of the politicians who signed up on Koo are members of the BJP (Group 2022). Consequently, Koo has been accused of supporting nationalist politics in India, despite its attempts to attract other political parties (Cornish 2022).

Second, in June 2021, Twitter temporarily suspended then Nigerian president Muhammadu Buhari from the platform after he threatened his political opponents with violence. This led Buhari’s government to instate a nationwide Twitter ban on June 5, 2021 (Paquette 2021). Afterwards, Koo experienced a significant uptake among Nigerian government officials and civilians, both supporters and opponents of the Buhari regime, leading the platform to roll out accessibility for the main vernacular languages spoken in Nigeria

(Abubakar and Nilesh 2021). Koo quickly gained legitimacy within the country, becoming an official channel of communication for the government alongside Facebook and Instagram. Twitter was reinstated in Nigeria in January 2022, after they agreed to conditions set by the Nigerian government (Akinwotu 2022), leading to a reduction in Koo’s usage from 289,000 monthly active Nigerian users to 40,000 by September 2022 (Dosunmu 2023).

Finally, in November 2022, shortly after Twitter was purchased by Elon Musk, Koo attracted a large Brazilian community following a Koo-related linguistic pun posted by Felipe Neto, a Brazilian YouTuber and online influencer (González Ormerod 2022). Consequently, Brazilian president Luiz Inácio Lula da Silva and many of his supporters joined Koo; Lula gained over 50,000 followers in less than 4 hours (Mihindukulasuriya 2022) with Koo becoming the most downloaded app on the Apple App Store and the Google Play Store for a few days in Brazil (D’Cruze 2022). Koo is still on the lookout for new markets to expand into. The Philippines, Malaysia and Indonesia in particular are countries the platform is targeting (Tech 2022a).

Table 1 shows the total number of posts, comments, likes and shares made in the top 10 languages on the platform, as well as the percentage of posts made in each language. Hindi makes up about half of the total messages, highlighting the dominance of the Indian community on Koo. English, Portuguese and Nigerian English are the only non-Indian languages which have significant use on the platform.

Focusing on high-profile Koo users, Table 2 shows the fraction of users with a green (verified) or yellow (account of eminence) badge for the 10 largest linguistic communities on Koo. Over a quarter of Portuguese-speaking users have received a green badge, the largest percentage for any linguistic community, indicating heavy use of Koo’s self-verification feature. In contrast, only 0.33% of Nigerian English accounts are self-verified. Hindi-speaking users also display a high level of adoption for the self-verification process. For each linguistic community, only a small fraction of accounts are awarded the yellow tick of eminence, with a slight bias favouring Indian communities.

Studies have shown that two accounts who share a common set of social media interactions typically share similar political or ideological views (Falkenberg et al. 2022, 2023). To map these connections on Koo, we construct the co-occurrence network of Koo users with the account of eminence badge following the methodology outlined in (Falkenberg et al. 2023), see Fig. 2. Each node in the co-occurrence network corresponds to a single eminent user, coloured according to their modal post language. Two eminent users are connected by an edge if their posts are liked by at least 50 common users. For visual clarity, we only show the giant connected component, and eminent users with edges to at least two other users.

The network shows that clusters are strongly influenced by linguistic factors: prominent users are more likely to be part of a similar social circle on Koo if they predominantly use the same language on the platform. Aside from the isolated Portuguese-speaking cluster and the Nigerian English speaking users, eminent users speaking Kannada,

Telugu and Gujarati are also slightly disconnected from the English/Hindi dominant community. However, there is little evidence of structural polarization within linguistic communities on Koo, unlike on Twitter (Falkenberg et al. 2023).

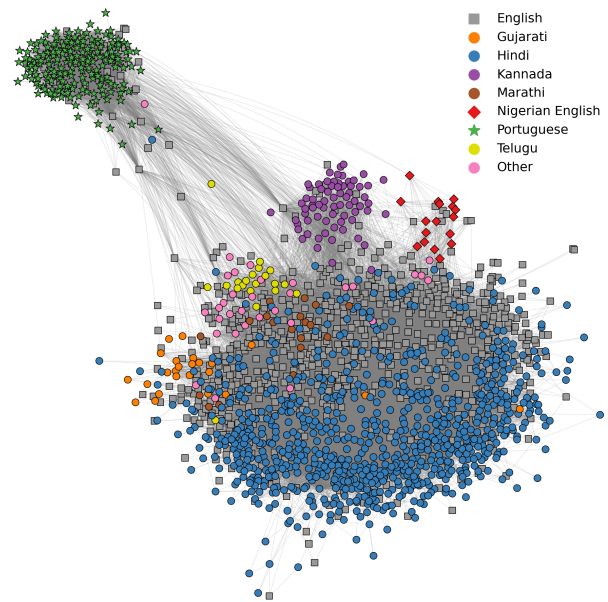


Figure 2: Co-occurrence network of accounts of eminence. Two eminent users are connected by an edge if at least 50 accounts on Koo interact with both of them. Nodes are coloured according to modal account language. Node shapes differentiate Indian and non-Indian languages.

Data Collection

Koo’s API. Koo provides a public API, which can be accessed by sending a GET request to <https://www.kooapp.com/api/>. All data included in this dataset was retrieved by querying the user profile endpoint: knowing a user ID or account handle, the API provides specific endpoints to retrieve their posts, comments, shares and likes.

User Profile Collection. In order to collect a user’s activity, we first need to retrieve their profile to get their user ID. We started with a short list of manually collected high-profile users, including prominent politicians, who are active on Koo. Through their interactions on Koo (see below) we identify any user they have interacted with. We query the user profile of these accounts, and repeat this process iteratively, collecting more users through a snowball sampling method until we retrieve no new user profiles. A limitation of this data acquisition strategy is that we do not capture lurkers on Koo (users who use Koo passively). However, we do capture all the interactions for every user who interacted with any account in the giant connected component on Koo. The user profiles provided by the API include information on a user’s username, display name, profile description and manually entered title.

Language	Posts	Ratio (%)	Comments	Likes	Shares
Hindi	35,074,082	48.9	33,762,199	141,851,555	22,710,161
English	21,341,502	29.8	15,364,430	59,405,972	10,745,495
Portuguese	4,933,641	6.88	9,747,183	51,323,645	1,521,073
Telugu	2,127,279	2.97	2,218,213	1,975,845	104,001
Kannada	1,889,473	2.63	3,700,669	5,972,733	859,069
Marathi	1,286,933	1.79	913,997	5,237,631	128,886
Tamil	1,268,948	1.77	376,721	1,262,414	82,940
Bengali	1,194,315	1.67	1,031,905	2,912,268	203,597
Gujarati	1,014,200	1.41	504,975	2,991,499	103,317
Nigerian English	610,568	0.85	1,076,604	1,893,175	571,813

Table 1: The top 10 languages used on Koo. The ratio indicates the percentage of the total number of posts written in each language. Columns indicate, for each language, the number of comments, likes and shares associated with the language.

Language	Self-Verified (%)	Accounts of Eminence (%)	User Profiles
Hindi	10.8	0.79	552,941
English	7.33	1.37	212,244
Portuguese	27.96	0.5	178,471
Telugu	4.05	0.35	16,265
Kannada	2.95	0.93	25,547
Marathi	5.81	0.63	11,226
Tamil	5.3	1.15	5,021
Bengali	5.66	0.34	14,619
Gujarati	5.51	0.58	9,931
Nigerian English	0.33	0.31	17,184

Table 2: Ratio of self-verified accounts and those with a yellow tick of eminence for the top 10 linguistic communities on Koo. The table also indicates the number of user profiles provided for each linguistic community.

User Activity. Given a user ID, Koo’s API provides specialized endpoints to query the user’s posts, comments, shares and likes. The endpoints are respectively **profile/createdKus**, **profile/commentsKus**, **profile/rekookKus** and **profile/likedKus**, with the user ID provided as a query parameter. We paginate through the results to ensure that all of a user’s posts and interactions are collected.

Fair Principles. The data released in this paper align with the FAIR guiding principles for scientific data:

- *Findable:* We assign a unique constant digital object identifier (DOI) to our dataset.
- *Accessible:* Our dataset is openly accessible.
- *Interoperable:* Releasing the data with JSON files ensures that it can easily be read with various programming languages and operating systems.
- *Reusable:* We provide the data schema in the *Data Description* Section, to ensure that the purpose of each data field is clear.

Ethical Considerations. The data released with this paper was collected using Koo’s public API, in accordance with the platforms’ Terms of Service. Koo does not offer the possibility for users to make their account private, therefore the data we are releasing contains publicly accessible information only. We do not release metadata related to geolocation. All data released are available publicly and results are aggregated across accounts. To ensure that our data collection pipeline and data sharing complies with the relevant

regulations, we have completed a Data Protection Impact Assessment which has been approved by our institution.

Data Description

We now present the structure of the dataset, available on a public repository (Mekacher 2024). We release the metadata related to each type of user activity, as well as the user profiles, in separate JSON files. Table 3 indicates, for each user activity, the fields included in the JSON schema, including the data type of each field and a short description. For user convenience, we provide language-specific files for each interaction type. In total, we release 34 files for comments, shares and likes, and 43 files for the posts. This difference is due to small linguistic communities with fewer than 5 posts and no other interaction types.

Results

In this section, we provide a quantitative overview of Koo including (1) an analysis of user engagement over time (2) an analysis of Koo’s news media ecosystems, and (3) an analysis of user content. These results offer a more in-depth description of Koo’s growth within the digital platform ecosystem and its potential to harbour a unique online community.

User retention. Previous studies have shown that, although alternative platforms are successful at attracting a large number of user registrations, many newcomers do not remain active on the platform and become idle within

Key	Type	Description
language_code_posts.json (43 files)		
id	string	ID of the post
creatorId	string	ID of the user who posted
title	string	Content of the post
createdAt	int	UNIX timestamp of the post
handle	string	Handle of the user who posted
language_code_likes.json (34 files)		
id	string	ID of the liked post
creatorId	string	ID of the user who posted
createdAt	int	UNIX timestamp of the like
handle	string	Handle of the user who posted
liker_id	string	ID of the user who liked
language_code_comments.json (34 files)		
id	string	ID of the commented post
creatorId	string	ID of the user who posted
title	string	Content of the comment
createdAt	int	UNIX timestamp of the comment
handle	string	Handle of the user who posted
commenter_id	string	ID of the user who commented
language_code_shares.json (34 files)		
id	string	ID of the shared post
creatorId	string	ID of the user who posted
createdAt	int	UNIX timestamp of the share
handle	string	Handle of the user who posted
sharer_id	string	ID of the user who shared
koo_users.json (1 file)		
id	string	ID of the user
handle	string	Handle of the user
title	string	Self-given title
description	string	Profile description
createdAt	int	Timestamp of the account creation

Table 3: Description of the metadata in each data file.

days of registering (Mekacher, Falkenberg, and Baronchelli 2023). Figure 3A shows the number of posts, comments, shares and likes, recorded on a daily basis on Koo. Figure 3B shows the daily number of active users over time.

After steady growth in 2020, the platform experienced a jump in activity in February 2021, attributed to BJP officials signing up on Koo with their followers. The daily activity count starts plateauing afterwards, but does not experience any substantial decrease. We see a second major increase in activity in November 2022 after Brazilian celebrities endorsed Koo, although activity levels quickly return to levels observed before the Brazilian migration. User engagement and activity does not fall substantially after any of the spikes in user registrations. This result may indicate low retention amongst users who joined Koo during the major collective migrations, a pattern that has been observed in other alt-tech outlets (Mekacher, Falkenberg, and Baronchelli 2023; Mekacher and Pappasavva 2022).

News ecosystem. Next, we look at news media URLs

shared on Koo. Previous studies have highlighted the emergence of online ecosystems where users are only exposed to a limited selection of sources, often due to information-filtering mechanisms (Sullivan et al. 2020). This phenomenon is commonly referred to as an *epistemic bubble* and has been implicated in online radicalisation processes (Nguyen 2018) and populist discourses (Anderson 2021). The question of news diversity is of particular interest in the case of an Indian social platform that aims to host a wide range of cultural backgrounds. Previous studies have underlined the ethical issues that arise when religious events are covered in a controversial way by mainstream outlets that do not promote media secularism (Ramesh 2022).

To measure the news ecosystem for each linguistic community, we separate posts into languages for the news media analysis. This approach allows us to map the news ecosystem shared among users who predominantly use the same language on the platform. Comparing the results across languages demonstrates an outlet’s ability to seep through language barriers. Figure 4 shows the top 20 online domains shared on Koo and their respective prevalence within the major linguistic communities on the platform. We notice a strong tendency for Hindi news outlets to be part of the top domains on Koo, as Hindi speakers represent almost half of Koo’s user base. However, some websites are also broadly shared within other linguistic communities. This is the case of *ETV Bharat*, an Indian news channel that is available in 11 major Indian languages. Social media platforms, such as YouTube and Facebook, are also widely shared across linguistic communities.

Some news outlets are found to be very popular amongst smaller Indian linguistic communities. This is the case for *Eenadu*, the most prominent daily newspaper in Telugu. Along with *Sakshi*, it reaches more than 70% of the Telugu-speaking audience (Monitor 2019). For the Tamil-speaking community, the lesser-known news media *OneIndia* is widely shared. *OneIndia* has been very prolific on Koo and has garnered a large following by systematically sharing links to its articles, thereby outranking some of the better known Tamil newspapers such as *Dinakaran*.

Interestingly, when looking at Nigerian English speakers, the most shared outlet is *Peoples Gazette*, an online newspaper launched in 2020 that has led several investigations about cases of corruption in Muhammadu Buhari’s government (Adenekan 2022). The second most shared news outlet among Nigerian English speakers (exc. YouTube and Telegram) is *The Punch*, the most widely read newspaper in Nigeria, which is also highly critical of Buhari’s politics (Times 2018). These findings suggest that the Nigerian government was also followed by its dissenters when it migrated to Koo in June 2021 and has a less prominent influence on the relevant news ecosystem than their political opponents.

Portuguese speakers widely share links to social media platforms, with Instagram, YouTube and Koo links being the most shared websites. These are followed by *G1 Globo*, one of the most popular news outlets in Brazil. Similar to the other major newspapers in Brazil, *Grupo Globo*, the conglomerate that owns *G1 Globo*, was highly critical of Jair Bolsonaro’s political stances (Newman et al. 2021). How-

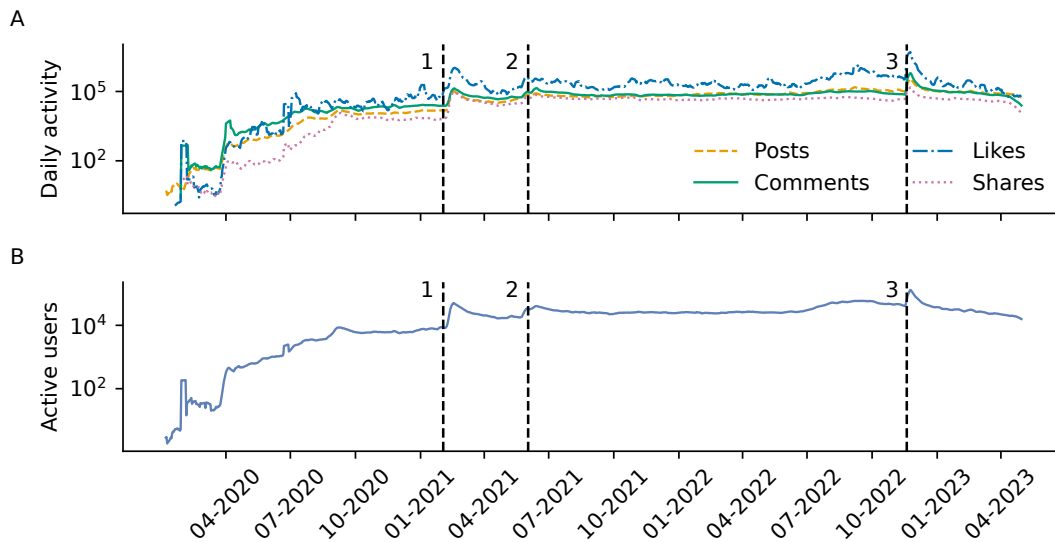


Figure 3: Daily activity and number of active users. A) 7-day moving window average of the amount of content (posts, comments, likes and shares) posted on Koo. B) 7-day moving average of the number of active users on a given day. A user is considered active if they created a new post or if they commented, shared or liked an existing post. The dashed lines indicate the events that led to the major collective migrations on Koo, namely 1) the farmers’ protest in India; 2) Twitter getting banned in Nigeria and 3) Elon Musk’s purchasing Twitter and the subsequent Brazilian migration.

ever, the network has also been accused of delegitimizing Lula and other leaders of the Workers’ Party, such as former president Dilma Rousseff (van Dijk 2017). In a similar fashion as for Nigeria, these results suggest that the Brazilian community on Koo share news media content which is primarily antagonistic to the current political regime. However, the Brazilian migration also included media actors that are supportive of Lula’s regime. The second most shared news platform within the Brazilian community is *Diario do Centro do Mundo*, a left-leaning digital news outlet criticised as “charlatans hired by the Workers’ Party” (Paiva 2023).

The dominance of specific news outlets within linguistic communities, particularly in India, suggests that Koo may not harbour, or be conducive to, a social environment that cultivates media pluralism. Media pluralism has often been defined as a hallmark of a healthy democracy (Keller 2011), while also being an important tool for tackling misleading news (Joris et al. 2020). To measure news diversity, we compute the Gini coefficient for the set of web domains shared by each linguistic community on Koo. A Gini coefficient of 1 indicates that only one news source is being shared within the community, whereas a coefficient of 0 would represent perfect equality across all news sources. Figure 5 shows the Gini coefficient for each linguistic community, plotted against its population size. The figure shows that larger linguist communities on Koo tend to have a Gini coefficient close to 1. This further highlights the dominance of individual news media outlets in each of these communities.

Hashtags analysis. Next, we look at the conversations being hosted on Koo. Previous studies focused on the US alt-tech ecosystem have shown that populist themes are

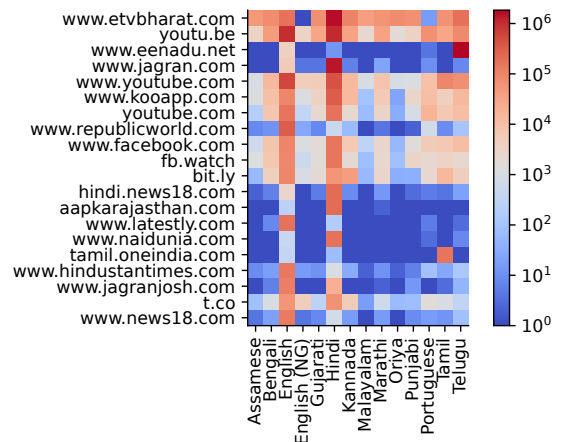


Figure 4: Top-shared web domains and their prevalence in the dominant linguistic communities. Number of links leading to a web domain shared by the top-10 linguistic communities on Koo. The top-20 shared domains are shown.

widely discussed in culturally homogeneous communities (Allchorn 2023). To identify the most salient topics users discuss, we perform an analysis of the hashtags shared by the major linguistic communities on Koo. Often used to bring visibility to a specific topic, hashtags have been shown to play an instrumental role on several social platforms, both in political campaigning and to rally citizens to activist movements. These social phenomena have been extensively stud-

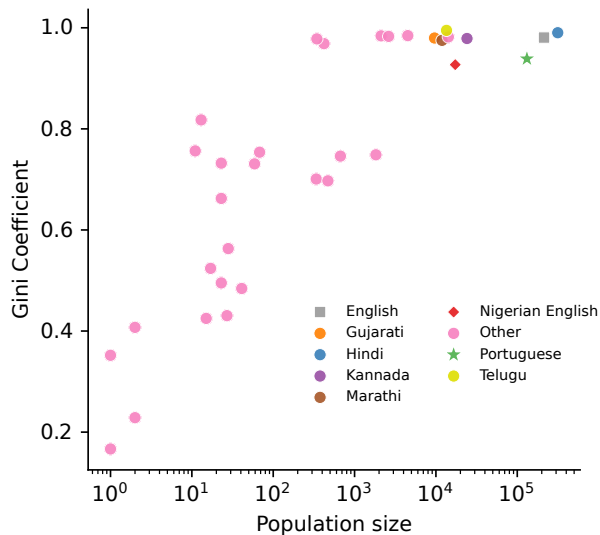


Figure 5: Media plurality across linguistic communities on Koo. The Gini coefficient for the news media web domains shared by each linguistic community, plotted against the population size of each linguistic community. A Gini coefficient close to 1 highlights a monopoly held by a single news source, whereas a Gini coefficient close to 0 indicates a more diverse link-sharing ecosystem.

ied in India (Losh 2014; Panda et al. 2020), Nigeria (Opeibi 2019; Akpojivi 2019) and Brazil (Almeida 2020; Soares and Recuero 2021).

Table 4 shows the 20 most shared hashtags for the Hindi, English, Nigerian English and Portuguese speaking Koo communities. For Hindi and English speakers, we note that the most popular hashtags gravitate around Indian matters (`#kooforindia`, `#india`), suggesting that the English-speaking community are largely based in India. Most Hindi and English hashtags relate to national politics and economic matters (`#narendramodi`, `#HindiNews`, `#Budget2022`), which further highlights that Koo has succeeded in hosting a sustainable conversation within India. However, we also find apolitical hashtags, particularly related to Cricket in India (`#Cricket`, `#iplauction2022`). Finally, this analysis highlights interest in Rampal Singh Jatain (`#Sat_Bhakti_Sandesh`, `#SantRampalJiMaharaj`), an Indian religious cult leader who has been previously arrested for murder (Chandra Pandey 2018). He signed up on Koo in August 2020 and garnered more than 330k followers.

The Nigerian English-speaking community prominently uses hashtags related to the Twitter ban and the migration to Koo in June 2021 (`#letskoo`, `#bantwitter`, `#nigeriatwitterban`). Furthermore, many hashtags also mention locations in the country, such as cities (`#kano`, `#kaduna`, `#lagos`) or regions (`#arewa`). These results show that the Twitter ban remained a major topic of debate, and that Nigerian users attempted to rebuild their social network on Koo by connecting with peers located in the same region of the country. National politics are also discussed (`#Buhari`, `#buhari-`

`train`, `#OsinbajoDay`) and highlight the presence of both pro- and anti-Buhari cohorts on the platform. The `#OsinbajoDay` hashtag refers to then Vice President Yemi Osinbajo, hailed by grassroots campaigners as the right candidate to take over “the good works already started by Muhammadu Buhari” (Oluwafemi 2021). In contrast, the `#buharitrain` hashtag is used on social media to criticize president Buhari’s railway project between the capital city of Abuja and its airport, with the investment considered a failure (Day 2023).

The hashtags used by the Brazilian community highlight a desire from users to connect with their peers on Koo (`#koo-brasil` and `#sdv`, Portuguese slang for “follow me back” on social platforms). This is further suggested by the prominence of hashtags related to Elon Musk’s purchase of Twitter (`#elonmusktwitter`, `#layoffs`, `#riptwitter`). The initial migration indeed occupies a significant space in the Brazilian conversation, with other prominent hashtags related to the linguistic pun that triggered the migration to Koo (`#brasil-nokoo`, `#tomandonokoo`) (González Ormerod 2022). Unlike the Indian and Nigerian communities, none of the most prevalent hashtags in the Brazilian community are related to contemporary political topics. Instead, mentions of the 2022 FIFA World Cup (`#copadomundo2022`, `#fifaworldcup`, `#vaibrasil`) and general conversations (`#memes`, `#humor`, `#art`) are more common. These results highlight that the Brazilian migration, unlike the Indian and the Nigerian ones, was not initially politically motivated.

Overall, the hashtags indicate that the main communities on Koo are all involved in conversations relating to the events which triggered their communities’ respective migrations to Koo. This highlights the impact that social media deplatforming can have on online discourses (Mekacher, Falkenberg, and Baronchelli 2023). The stark difference in the prominence of political hashtags between the Brazilian community and the Nigerian / Indian communities underlines that the initial motivations leading to a platform migration can be both political and apolitical, and can heavily impact the dominant discourses that subsequently emerge on alt-tech platforms.

Conclusion

In this work, we release and present a Koo dataset totalling over 72M posts, 399M user interactions, and 1.4M user profiles. The whole dataset was collected via Koo’s undocumented public API.

Our analysis highlights the growth of a multi-lingual, cross-country online ecosystem, with collective migrations to the platform often spearheaded by influential politicians. Many narratives on the platform relate to the events which triggered the initial migrations to Koo, resulting in prominent debates on the national political landscape of India and Nigeria, and to a lesser extent Brazil. For India, Koo content is heavily biased in support of the ruling BJP party. In contrast, both the Nigerian and Brazilian communities on Koo share pro- and anti-Government content, although the Brazilian Koo discussion is less political than the others.

These collective migrations highlight how the success of an alt-tech fringe platform can be intimately linked to the decisions made by mainstream platforms and to platform- or

Hindi		English		Nigerian English		Portuguese	
Hashtag	Ratio (%)	Hashtag	Ratio (%)	Hashtag	Ratio (%)	Hashtag	Ratio (%)
kooftheday	0.77	kooftheday	0.62	nigeria	5.41	koobrasil	4.46
comedy	0.72	india	0.37	letskoo	3.38	brasilnokoo	3.78
memes	0.71	memes	0.32	Nigeria	2.34	koo	3.77
memeoftheday	0.71	India	0.31	bantwitter	1.42	copadomundo2022	3.61
dailymemes	0.42	koo	0.31	kano	1.32	brasil	2.94
Budgetmemes	0.37	comedy	0.30	kanconnect	1.14	tomandonokoo	2.56
SaintRampalJi	0.31	kooforindia	0.29	koo	1.04	meme	1.57
HindiNews	0.29	memeoftheday	0.28	kaduna	1.01	koonobrasil	1.42
koo	0.25	narendramodi	0.27	lagos	0.95	riptwitter	1.30
SantRampalJiMaharaj	0.25	motivation	0.25	arewapeeps	0.87	memes	1.13
Sat_Bhakti_Sandesh*	0.24	dailymemes	0.25	nigeriatwitterban	0.82	elonmusktwitter	0.79
india	0.24	trending	0.23	kooyouropinion	0.81	humor	0.78
MpNews	0.20	Budget2022	0.20	june12	0.80	layoffs	0.77
entertainment	0.20	Budgetmemes	0.19	arewakooconnect	0.78	koobr	0.75
Navabharat	0.19	Cricket	0.18	ifollowback	0.67	fifaworldcup	0.75
Nan	0.19	news	0.18	OsinbajoDay	0.67	copadomundo	0.64
kooforindia	0.18	COVID19	0.18	covid19	0.66	vaibrasil	0.56
iplauction2022	0.18	SaintRampalJi	0.17	Buhari	0.64	Brasil	0.54
narendramodi	0.18	SantRampalJiMaharaj	0.17	Arewa	0.64	sdv	0.54
kookiyakya	0.18	life	0.15	buharitrain	0.63	art	0.53

Table 4: Top hashtags used on Koo. Prevalence of the top 20 hashtags used by the major linguistic communities on Koo. The percentage indicates the percentage of usage of the respective hashtag within the considered linguistic community. The hashtag indicated with the asterisk (*) is translated from Hindi.

regulator-driven deplatforming events (e.g., because Twitter banned users, or because regulators banned Twitter) (Mekacher, Falkenberg, and Baronchelli 2023). Catering to these communities, Koo has become a prominent venue for political and social debates for several communities. However, despite a lack of news media and content diversity, and unlike most alt-tech fringe platforms, Koo has had success in attracting a politically heterogeneous user-base beyond India in both Nigeria and Brazil. This sets Koo apart from other emerging platforms based outside the US - especially those in China and Russia where politicians have emphasised the need for “Internet sovereignty” (Broderick 2023) - in that their ambition is to expand beyond their home jurisdiction. Consequently, a platform like Koo has the potential to challenge the current dominance of US-based social media platforms, a change which may have important consequences for social media regulation.

We anticipate that this dataset will support further research investigating digital ecosystems based outside the US, and how they impact political campaigning and news sharing. This is particularly important given upcoming elections in several major democracies, including India, in 2024 (Hsu, Thompson, and Lee Myers 2024). Given the well-documented risks that social media can be used to promote political disinformation (Hsu, Thompson, and Lee Myers 2024), and that Koo attracted some users who were deplatformed from its mainstream competitors, the platform may face challenges related to content moderation, balancing its free-speech agenda with its aim to prevent the spread and amplification of hate speech.

Finally, given Koo’s linguistic diversity, we hope that our dataset will motivate researchers to further develop computational tools for studying social media narratives and news media sharing in underrepresented languages. This work will be crucial to enable a robust analysis of the online news media ecosystem, misinformation, and its potential impact on political campaigns across diverse markets.

Acknowledgments

M.F. and A.B. acknowledge support from the IRIS Infodemic Coalition (UK government, grant no. SCH-00001-3391). The authors declare no competing interests.

References

- Abubakar, I.; and Nilesh, C. 2021. Koo is selling itself as a Twitter substitute in Nigeria. <https://restofworld.org/2021/koo-is-selling-itself-as-a-twitter-substitute-in-nigeria/> Accessed 15 August, 2023.
- Adenekan, S. 2022. Police raid Peoples Gazette office, arrest five staffers. <https://www.premiumtimesng.com/news/top-news/544341-police-raid-peoples-gazette-office-arrest-five-staffers.html?tztc=1> Accessed 4 December, 2023.
- Akinwotu, E. 2022. Nigeria lifts Twitter ban seven months after site deleted president’s post. <https://www.theguardian.com/world/2022/jan/13/nigeria-lifts-twitter-ban-seven-months-after-site-deleted-presidents-post> Accessed 2 November, 2023.
- Akpojivi, U. 2019. I won’t be silent anymore: Hashtag ac-

- tivism in Nigeria. *Communicatio: South African Journal of Communication Theory and Research*, 45(4): 19–43.
- Allchorn, W. 2023. Far-Right Extremist Exploitation of AI and Alt-Tech: The Need for P/CVE Responses to an Emerging Technological Trend. <https://gnet-research.org/2023/10/09/far-right-extremist-exploitation-of-ai-and-alt-tech-the-need-for-p-cve-responses-to-an-emerging-technological-trend/> Accessed 14 December, 2023.
- Allyn, B. 2022. Google is now distributing Truth Social, Trump’s Twitter alternative. <https://www.npr.org/2022/10/14/1129024148/google-truth-social-trump> Accessed 30 October, 2023.
- Almeida, H. B. d. 2020. From shame to visibility: hashtag feminism and sexual violence in Brazil. *Sexualidad, Salud y Sociedad (Rio de Janeiro)*, 19–41.
- Ammassari, S.; Fossati, D.; and McDonnell, D. 2022. Supporters of India’s BJP: Distinctly Populist and Nativist. *Government and Opposition*, 58(4): 807–823.
- Anderson, E. 2021. Epistemic bubbles and authoritarian politics. *Political epistemology*, 11–30.
- Barclay, F. P.; Pichandy, C.; and Venkat, A. 2014. Indian Elections, 2014: Political Orientation of English Newspapers. *Asia Pacific Media Educator*, 24(1): 7–22.
- Bhat, P. 2021. Platform Politics: The Emergence of Alternative Social Media in India. *Asia Pacific Media Educator*, 31(2): 269–276.
- Broderick, T. 2023. Russia Is Trying to Leave the Internet and Build Its Own. <https://www.scientificamerican.com/article/russia-is-trying-to-leave-the-internet-and-build-its-own/> Accessed 11 January, 2024.
- Chandra Pandey, M. 2018. Godman Rampal sentenced to life in prison till death in murder case. <https://www.indiatoday.in/india/story/sant-rampal-life-imprisonment-murder-1368904-2018-10-16> Accessed 6 December, 2023.
- Cornish, C. 2022. India’s homegrown ‘Twitter’ courts political parties to shed nationalist tag. <https://www.ft.com/content/bd296702-18ad-4303-8192-bbb1268e3b56> Accessed 14 September, 2023.
- Dang, S. 2023. Twitter rival Koo integrates ChatGPT to help users create content. <https://www.reuters.com/technology/twitter-rival-koo-integrates-chatgpt-help-users-create-content-2023-03-13/> Accessed 6 November, 2023.
- Day, B. 2023. How Buhari’s train to nowhere worsens Nigeria’s debt woes. <https://businessday.ng/business-economy/article/how-buharis-train-to-nowhere-worsens-nigerias-debt-woes/> Accessed 4 December, 2023.
- D’Cruze, D. 2022. Made in India Twitter-rival Koo enters Brazil with support for Portuguese language. <https://www.businesstoday.in/technology/news/story/made-in-india-twitter-rival-koo-enters-brazil-with-support-for-portuguese-language-353635-2022-11-21> Accessed 29 November, 2023.
- Dosunmu, D. 2023. When Nigeria banned X, Koo had a golden opportunity, and squandered it. <https://restofworld.org/2023/nigeria-koo-twitter-rival-flop/> Accessed 2 November, 2023.
- Fair, G.; and Wesslen, R. 2019. Shouting into the Void: A Database of the Alternative Social Media Platform Gab. *Proceedings of the International AAAI Conference on Web and Social Media*, 13: 608–610.
- Falkenberg, M.; Galeazzi, A.; Torricelli, M.; Di Marco, N.; Larosa, F.; Sas, M.; Mekacher, A.; Pearce, W.; Zollo, F.; Quattrociocchi, W.; et al. 2022. Growing polarization around climate change on social media. *Nature Climate Change*, 12: 1114–1121.
- Falkenberg, M.; Zollo, F.; Quattrociocchi, W.; Pfeffer, J.; and Baronchelli, A. 2023. Affective and interactional polarization align across countries. *ArXiv Preprint*, 2311.18535.
- FORCE11. 2020. The FAIR Data principles. <https://force11.org/info/the-fair-data-principles/>.
- Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.
- Gerard, P.; Botzer, N.; and Weninger, T. 2023. Truth Social Dataset. *Proceedings of the International AAAI Conference on Web and Social Media*, 17: 1034–1040.
- Gilbert, D. 2023. The GOP Presidential Debate Is Livestreaming on Rumble, Home to Notorious White Nationalist Nick Fuentes. <https://www.wired.com/story/gop-debate-rumble-livestream-nick-fuentes/> Accessed 9 January, 2024.
- González Ormerod, A. 2022. Twitter drama has Brazilians flocking to Indian platform Koo. <https://restofworld.org/2022/twitter-brazil-koo/> Accessed 16 August, 2023.
- Group, P. R. 2022. Challenging big social media, falling to partisanship: The Koo Conundrum. <https://precog.iiit.ac.in/blog/2021/08/30/the-koo-conundrum/> Accessed 15 November, 2023.
- Hsu, T.; Thompson, S.; and Lee Myers, S. 2024. ‘Stakes are really high’: misinformation researcher changes tack for 2024 US election. <https://www.nytimes.com/2024/01/09/business/media/election-disinformation-2024.html> Accessed 11 January, 2024.
- Joris, G.; De Grove, F.; Van Damme, K.; and De Marez, L. 2020. News Diversity Reconsidered: A Systematic Literature Review Unraveling the Diversity in Conceptualizations. *Journalism Studies*, 21(13): 1893–1912.
- Keller, P. 2011. *Democracy, Pluralism, and the Media*, 405–448. Oxford University Press.
- Koo App. 2022. Community Guidelines. <https://info.kooapp.com/community-guidelines/> Accessed 12 December, 2023.
- Leingang, R. 2023. Biden campaign joins Trump’s Truth Social platform: ‘Converts welcome!’. <https://www.theguardian.com/us-news/2023/oct/16/biden-campaign-joins-truth-social-trump> Accessed 9 January, 2024.
- Losh, E. 2014. Hashtag feminism and Twitter activism in India. *Social Epistemology Review and Reply Collective*, 3(3): 11–22.

- Majumdar, R. 2021. What is the Koo app and why are government officials joining it? <https://www.indiatoday.in/india-today-insight/story/what-is-the-koo-app-and-why-are-government-officials-joining-it-1769895-2021-02-16> Accessed 6 December, 2023.
- Marques, F. P. J.; Mont’Alverne, C.; and Mitozo, I. 2019. Editorial journalism and political interests: Comparing the coverage of Dilma Rousseff’s impeachment in Brazilian newspapers. *Journalism*, 22(11): 2816–2835.
- McGregor, S. C. 2020. “Taking the Temperature of the Room”. *Public Opinion Quarterly*, 84(S1): 236–256.
- Mekacher, A. 2024. The Koo Dataset: An Indian Microblogging Platform With Global Ambitions. <https://doi.org/10.5281/zenodo.10476212> Accessed 10 January, 2024.
- Mekacher, A.; Falkenberg, M.; and Baronchelli, A. 2023. The systemic impact of deplatforming on social media. *PNAS Nexus*, 2(11).
- Mekacher, A.; and Papasavva, A. 2022. “I Can’t Keep It Up.” A Dataset from the Defunct Voat.co News Aggregator. *Proceedings of the International AAAI Conference on Web and Social Media*, 16(1): 1302–1311.
- Mihindukulasuriya, R. 2022. Brazilian President Lula joins India’s Twitter rival Koo, gets over 50,000 followers in 4 hours. <https://theprint.in/world/brazilian-president-lula-joins-indias-twitter-rival-koo-gets-over-50000-followers-in-4-hours/1245516/> Accessed 23 October, 2023.
- Monitor, M. O. 2019. Media Ownership Matters. <https://www.mom-gmr.org/en/> Accessed 22 November, 2023.
- Newman, N.; Fletcher, R.; Schulz, A.; Andi, S.; Robertson, C. T.; and Nielsen, R. K. 2021. Reuters Institute digital news report 2021. *Reuters Institute for the study of Journalism*.
- Nguyen, C. T. 2018. Echo chambers and epistemic bubbles. *Episteme*, 17(2): 141–161.
- Oluwafemi, A. 2021. Group to hold “Osinbajo Day” on Sunday, describes him as “Nigeria’s future for 2023”. <https://www.thecable.ng/group-to-hold-osinbajo-day-on-sunday-describes-him-as-nigerias-future-for-2023> Accessed 12 December, 2023.
- Opeibi, T. 2019. The Twittersphere as political engagement space: A study of social media usage in election campaigns in Nigeria. *Digital Studies/Le champ numérique*, 9(1).
- Paiva, L. 2023. Ciro Gomes deverá indenizar DCM por dizer que é ‘tudo picareta’ contratado pelo PT. <https://www.jota.info/justica/ciro-gomes-devera-indenizar-dcm-por-dizer-que-e-tudo-picareta-contratado-pelo-pt-06022023> Accessed 11 December, 2023.
- Panda, A.; Kommiya Mothilal, R.; Choudhury, M.; Bali, K.; and Pal, J. 2020. Topical Focus of Political Campaigns and its Impact: Findings from Politicians’ Hashtag Use during the 2019 Indian Elections. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1): 1–14.
- Paquette, D. 2021. Nigeria suspends Twitter after the social media platform freezes president’s account. <https://www.washingtonpost.com/world/2021/06/04/nigeria-suspends-twitter-buhari/> Accessed 30 October, 2023.
- Phartiyal, S. 2021. Explainer: The Indian Twitter rival staging a Koo. <https://www.reuters.com/business/media-telecom/indian-twitter-rival-staging-koo-2021-02-10/> Accessed 30 October, 2023.
- Picard, R. G. 2014. Twilight or New Dawn of Journalism?: Evidence from the changing news ecosystem. *Journalism Practice*, 8(5): 488–498.
- Punia, S. K. 2016. Social Media: Political Campaigns and Elections. *Madhya Pradesh Journal of Social Sciences*, 21(2): 65.
- Ramesh, A. 2022. Media and religion: Media framing of significant religious issues in English Newspapers of India. *Mass Communicator: International Journal of Communication Studies*, 16(1): 12–20.
- Ribeiro, G.; and Mendes, G. 2022. Bolsonaro’s echo chambers. <https://brazilian.report/newsletters/brazil-weekly/2022/04/18/social-media-gettr-airline-tickets/> Accessed 30 October, 2023.
- Singh, S. 2022. Koo said to be the second largest microblogging platform in the world now. <https://www.cnbctv18.com/business/companies/koo-becomes-second-largest-microblogging-platform-in-the-world-after-twitter-15179881.htm> Accessed 1 December, 2023.
- Soares, F. B.; and Recuero, R. 2021. Hashtag wars: Political disinformation and discursive struggles on twitter conversations during the 2018 Brazilian presidential campaign. *Social Media+ Society*, 7(2): 20563051211009073.
- Sullivan, E.; Sondag, M.; Rutter, I.; Meulemans, W.; Cunningham, S.; Speckmann, B.; and Alfano, M. 2020. Vulnerability in Social Epistemic Networks. *International Journal of Philosophical Studies*, 28(5): 731–753.
- Tech, H. 2022a. Koo app user base hits 15-mn mark, looks to fly to foreign lands. <https://tech.hindustantimes.com/tech/news/koo-app-user-base-hits-15-mn-mark-looks-to-fly-to-foreign-lands-71635089172833.html> Accessed 30 October, 2023.
- Tech, H. 2022b. Koo app yellow tick of verification: Here is how to get it. <https://tech.hindustantimes.com/how-to/koo-app-yellow-tick-of-verification-here-is-how-to-get-it-71637228876715.html> Accessed 2 November, 2023.
- Times, P. 2018. Presidency attacks Punch Newspaper. <https://www.premiumtimesng.com/news/more-news/282609-presidency-attacks-punch-newspaper.html> Accessed 4 December, 2023.
- Trujillo, M. Z.; Gruppi, M.; Buntain, C.; and Horne, B. D. 2022. The MeLa BitChute Dataset. *Proceedings of the International AAAI Conference on Web and Social Media*, 16: 1342–1351.
- van Dijk, T. A. 2017. How Globo media manipulated the impeachment of Brazilian President Dilma Rousseff. *Discourse & Communication*, 11(2): 199–229.
- Werleman, C. 2021. Rising Violence Against Muslims in India Under Modi and BJP Rule. *Resurgence of Anti Islam in the World*, 23(Spring 2021): 39–49.

Paper Checklist

1. For most authors...
 - (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? **Yes. Our dataset advances our understanding of social media use in emerging economies and discusses how political campaigns make use of social media. The paper respects privacy norms by only presenting data on aggregate. We do not comment on the choices or views of specific cultures of socio-economic groups. To ensure compliance with relevant regulations, we have completed a data protection and impact assessment which has been approved by our institution.**
 - (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? **Yes. Our abstract primarily discusses the dataset released and its potential use cases.**
 - (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? **Yes. The analytical tools we use support the claims made regarding the potential use cases of our dataset.**
 - (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? **Yes. We comment on the limitations of the data collection pipeline, specifically that it is less likely to capture passive users on the platform.**
 - (e) Did you describe the limitations of your work? **Yes. The limitations of our data collection pipeline are discussed in the *Data Description* Section.**
 - (f) Did you discuss any potential negative societal impacts of your work? **No, our aim is to provide additional resources to study online communities in emerging economies. As such, we do not see any direct potential negative societal impact our work could have.**
 - (g) Did you discuss any potential misuse of your work? **No, the main focus of our work is to make a dataset publicly available. We have ensured not to release any personally identifiable information to protect users' anonymity. All data released is already publicly available and all our analysis presents results aggregated across social media accounts.**
 - (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? **Yes. We have noted that metadata which may reveal personal information, including geolocation data, has been removed from our dataset.**
 - (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? **Yes, our data collection conforms with Koo's Terms of Service and regulatory requirements. We have also completed a Data Protection Impact Assessment with our institution, to ensure that our research protocol complies with ethical regulations.**
2. Additionally, if your study involves hypotheses testing...
 - (a) Did you clearly state the assumptions underlying all theoretical results? **NA**
 - (b) Have you provided justifications for all theoretical results? **NA**
 - (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? **NA**
 - (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? **NA**
 - (e) Did you address potential biases or limitations in your theoretical framework? **NA**
 - (f) Have you related your theoretical results to the existing literature in social science? **NA**
 - (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? **NA**
3. Additionally, if you are including theoretical proofs...
 - (a) Did you state the full set of assumptions of all theoretical results? **NA**
 - (b) Did you include complete proofs of all theoretical results? **NA**
4. Additionally, if you ran machine learning experiments...
 - (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? **NA**
 - (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? **NA**
 - (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? **NA**
 - (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? **NA**
 - (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? **NA**
 - (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? **NA**
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity...**
 - (a) If your work uses existing assets, did you cite the creators? **NA**
 - (b) Did you mention the license of the assets? **NA**
 - (c) Did you include any new assets in the supplemental material or as a URL? **Yes, we provide the URL at which our data are publicly available. See <https://doi.org/10.5281/zenodo.10476212>**
 - (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? **Yes. We note that in Koo's privacy policy users are informed that their data will be publicly available to anyone. We note that our data collection pipeline does not infringe the platform's Terms of Service.**

- (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? **Yes.** The data includes user profiles as described in the *Data Description* Section. We have removed sensitive personal data such as geolocation data, but have not removed user handles. Our dataset provides the textual content of the posts and comments. Koo's community guidelines prohibit hateful and discriminatory content (Koo App 2022), therefore we do not expect our data to contain offensive content.
 - (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? **Yes,** see the *Data Collection* Section.
 - (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? **No,** but we provide a detailed description of the metadata included in the data files in the *Data Description* Section and we extensively describe our data collection process in the *Data Collection* Section.
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity...**
- (a) Did you include the full text of instructions given to participants and screenshots? **NA**
 - (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? **NA**
 - (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? **NA**
 - (d) Did you discuss how data is stored, shared, and de-identified? **NA**