Narrative Building in Propaganda Networks on Indian Twitter

Saloni Dash
Microsoft Research
Bangalore, Karnataka, India
salonidash77@gmail.com

Arshia Arya
Microsoft Research
Bangalore, Karnataka, India
arshia1012@gmail.com

Sukhnidh Kaur
Microsoft Research
Bangalore, Karnataka, India
kaursukhnidh@gmail.com

Joyojeet Pal
University of Michigan
Ann Arbor, Michigan, USA
joyojeet@umich.edu

ABSTRACT

The misuse of social media platforms to influence public opinion through propaganda campaigns are a cause of rising concern globally. Particularly, countries like India, where politicians communicate with the public through unmediated, curated twitter feeds, have witnessed a significant surge in strategic online manipulation. In this paper, we study propaganda messaging on Indian Twitter during two politically polarizing events. We collect over 80M Hindi and English tweets from over 26k politicians and 6k influencers. Using a mixed-methods approach, we identify major propaganda narratives across all events. We further use a network causal inference based approach to isolate influential actors who play a significant role in propagating the identified narratives. We conclude by discussing how these opinion leaders and their information dissemination, are central to instigating and building propaganda campaigns on Twitter.

KEYWORDS

propaganda, influencers, causal impact estimation, narratives

ACM Reference Format:
Saloni Dash, Arshia Arya, Sukhnidh Kaur, and Joyojeet Pal. 2022. Narrative Building in Propaganda Networks on Indian Twitter. In *14th ACM Web Science Conference 2022 (WebSci '22), June 26–29, 2022, Barcelona, Spain.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3501247.3531581

1 INTRODUCTION

The central role of social media platforms in news consumption and interpersonal engagement globally has allowed political organizations to pointedly build and proselytize narratives to their constituents, a process that was earlier typically inaccessible in a mainstream media mediated news environment. We see this particularly in countries like India, where politicians eschew the inconvenience of independent-media scrutinized debate in favor of unhindered outreach, often aided by large retinues of supporters who help promote convenient narratives as doctrinal fact [13]. This has resulted in what may seem at first glance to be a more participatory democratic discourse where citizens converse directly with representatives, but in practice it enables the weaponization of platforms for propaganda and misinformation campaigns [8].

Although propaganda and incitement campaigns have always been part of political outreach [7], social media allows a mix of discursive, network, and temporal techniques to build peripheral and sectarian beliefs in scale and tone, and normalize them as acceptable to the mainstream. In spite of recent efforts by social media platforms like Twitter [1] and Facebook[2] to curb the misuse of their platform to manipulate information online, hostile influence operations continue to operate undetected on social media.

A crucial aspect of detecting propaganda campaigns is understanding the overarching narratives that propagandists attempt to build [18]. A narrative is a collection of stories with coherent themes [6]. According to the narrative paradigm theory [5], for a narrative to be successful, it must have coherence and fidelity, i.e. it must create a shared meaning among its audience. Discovering propaganda narratives and the channels of influence that instigate them may provide insights into the strategic goals and ideological foundations of propaganda campaigns [18].

Studying the contours of narrative building thus requires an examination of the prominent social media accounts that engage and impact the threads shape the public imaginary. To do this, we used a dataset of over 6k "influencers" accounts – individuals who derive their legitimacy either solely or partly through online operation, or have an offline sphere of expertise such as entertainers, journalists [4] and over 26k politicians from from two major national parties in India – the national incumbent Bharatiya Janata Party (BJP) and the opposition party Indian National Congress (INC). We consider two polarizing issues that have seen spikes in online traffic from influential individuals in India from 2019-2021. In the order of their onset, they are the Citizenship Amendment Act (CAA) and the COVID-19 pandemic.

The Citizenship Amendment Act (CAA), which disqualified Muslims from qualifying for refugee status when entering from India's neighboring countries, was enacted alongside the government's plans to conduct a nationwide exercise to identify "illegal foreigners" living in India by requiring historical documentation of residency. This stoked fears of losing citizenship among India's Muslims and led to widespread protests.

WebSci '22, June 26–29, 2022, Barcelona, Spain

ACM ISBN 978-1-4503-9191-7/22/06…$15.00
https://doi.org/10.1145/3501247.3531581

[1] https://help.twitter.com/en/rules-and-policies/platform-manipulation
[2] https://about.fb.com/news/2018/12/inside-feed-coordinated-inauthentic-behavior/

A key event, during the early phase of the COVID-19 pandemic in India, was that of the Tablighi Jamat incident. This referred to a religious conference of members of the Muslim sect that began in early March 2020, and had thousands of visitors from various parts of the world. The attendees had only partially dispersed by the time a nationwide lockdown was announced, and several were infected with coronavirus. The gathering would eventually be projected as a super-spreader event in the media, one to which thousands of coronavirus cases were attributed. The event triggered a social media frenzy, building resentment towards the attendees, and by extension, towards the Muslim community in India.

In the context of these contentious issues, we study narrative building and propagation in propaganda networks by investigating the following research questions –
(i) *What were the predominant narratives in propaganda networks of influencers and politicians?* (ii) *How can we characterize actors that were central to instigating and propagating such narratives?*

Our contributions include systematically studying narrative building in propaganda networks in a non-western context using: (i) multilingual narrative extraction using a mixed methods approach of clustering multilingual sentence embeddings and manual annotation (ii) causal impact estimation of influential actors in narrative building and propagation.

2 RELATED WORK

The definition of propaganda used in this paper, as described in [3] is *"any expression of opinion or action by individuals or groups deliberately designed to influence opinions or actions of other individuals or groups with reference to predetermined ends"*. Bolsover and Howard [2] identify two main elements of propaganda, (i) trying to influence opinion, and (ii) doing so on purpose. Martino et al. [9] curate a list of propaganda techniques used to disseminate propaganda messaging, including *name calling*, *repetition* and *slogans*. These findings inform our dataset and propaganda detection technique, where we ensure (i) by curating hashtags that were most likely used in propagandistic settings and (ii) by considering a dataset of politically influential accounts that potentially benefit from engaging in propaganda campaigns.

In order to model narratives, we extend a stance detection technique described in [14], which clusters sentence embeddings of tweets in the embedding space. Blackburn et al. [1] use a similar technique, using agglomerative clustering of word embeddings of ngrams of the tweets. Other common modelling techniques use generative probabilistic models like LDA [16, 17].

3 METHOD

3.1 Dataset

The dataset of influential figures in the Indian context is from Dash et al. [4]. We briefly describe their data collection method here.

3.1.1 Politicians. As outlined in Dash et al. [4], the politicians' dataset is built using a publicly available dataset [12] of 36k Twitter accounts of Indian politicians, which include several elected representatives from prominent parties in India, as well as party members like treasurer or youth wing leader. A Machine Learning classification pipeline, followed by validation from human annotators called is used by Panda et al. [12] to curate the dataset. Furthermore, the politicians are also manually labelled by the state and party they belong to. We use the party labels to filter 14,094 BJP politicians and 12,341 INC politicians.

3.1.2 Influencers. In order to construct the dataset of influencers, Dash et al. [4] iteratively fetch the friends of the 26k BJP and INC politicians. From the resultant list, they remove all the politicians from Panda et al. [12], non-Indian global figures etc., such that they are left with a total of 10k Twitter accounts that are highly followed by Indian politicians. They then manually annotate the filtered accounts to remove the false positives, and categorize each influencer into 1 of 12 categories by which they are primarily known. This results in a total of 6626 influencer accounts. Refer to Dash et al. [4] for more details on the complete list of categories.

3.1.3 Data Collection. Finally, over 80M English and Hindi tweets are collected[3] from these 26k politicians and 6k influencers accounts between June 2019 - March 2021. The tweets are preprocessed in a manner similar to Rashed et al. [14], including case-folding and removal of links, emojis, punctuations and non alphanumeric characters.

3.2 Multilingual Tweet Classification

To capture event specific tweets, we use the pipeline in Dash et al. [4], where they use a Word2Vec bag-of-words based technique to classify the collected tweets. Briefly, they first define a set of high precision keywords that are indicative of the event and then train a Word2Vec model based on the tweets that contain at least one of the keywords. The word embeddings are then used to iteratively expand the seed set according to the cosine similarity based criteria. Refer to Dash et al. [4] for more details and the specific keywords used to classify the English tweets into the different events. We extend this method to classify tweets in the Hindi language. Some of the resultant keywords and hashtags are in Table 1.

	Keywords & Hashtags
CAA/NRC	सीएए, एनआरसी, #नागरिकतासंशोधनविधेयक, राष्ट्रीय रजिस्ट
COVID-19	कोविड19, #कोरोनावाइरस, सोशल डिस्टन्सिंग, महामारी

Table 1: Hindi Keywords & Hashtags by Event.

The number of tweets for each event, identified by the classification pipeline, including the number of users who have tweeted about that particular event are in Table 2

	Users		**Tweets**	
	English	Hindi	English	Hindi
CAA/NRC	13,744	14,250	527,702	619,758
COVID-19	20,871	18,496	3,069,851	2,861,439

Table 2: Number of Users and Tweets by Event

[3]Using Tweepy - https://www.tweepy.org/

3.3 Propaganda Detection

In order to detect tweets that are potentially part of propaganda campaigns, we use our contextual knowledge of the events and curate a list of hashtags that were most likely used in propagandistic messaging. We display the top propaganda hashtags for each event in Table 3. The distribution of propganda tweets and users for each event is in Table 4.

	Hashtags
CAA/NRC	#shaheenbaghcracks, #tukdefundedcaastir, #शाहिनबाग_का_भंडाफोड़, #शाहीन_बाग_की_बिकाऊ_औरते
COVID-19	#coronajihad, #tabhlighijamaatvirus, #कोरोना_जिहाद,#तबलीगी_जमात_जिहाद

Table 3: Propaganda Hashtags by Event

To obtain an estimate of the performance of this propaganda detection technique, we randomly select a representative sample (i.e. we calculate sample size using 95% confidence level and 3% confidence interval) and have two annotators manually annotate whether each tweet is part of a propaganda campaign or not. We report an overall accuracy of 90.66% across both events with an inter-annotator agreement of 84.03%.

Event	# Users	# Propaganda Tweets
CAA/NRC	1,435	7,850
COVID-19	1,576	6,783

Table 4: Distribution of Number of Propaganda Tweets and Users by Event

3.4 Narrative Extraction

In order to remove redundant tweets that are not exact duplicates, but convey the same information, we use fuzzy string matching with the token sort ratio[4] metric to remove all tweets that are close duplicates of each other. For instance, *" big names from congress and its cronies is there any surprises, pfi funded anticaa demonstrations..."* and *" pfi funded anticaa demonstrations..."* have a token sort ratio of 83% since they contain significant overlaps in their content. Therefore, we remove all tweets which have token sort ratio score of more than 80%.

All tweets are then mapped to an embedding space using an extended version [15] of the multilingual Universal Sentence Encoder (mUSE) [19] which extends the base model to 50 languages. The vector space generated by mUSE is aligned, i.e. tweets with high semantic similarity in different languages lie in the same neighbourhood in the embedding space. The tweet embeddings are further projected to a two dimensional space using UMAP [11]. We then use hierarchical density based clustering (HDB-SCAN) [10] to cluster semantically similar tweets. We then manually identify the dominant messaging in each cluster and construct narrative maps for each event.

[4]https://github.com/seatgeek/thefuzz

3.5 Impact Estimation

To quantify the effect of individual users in building and propagating narratives, we measure each account's unique causal contribution to a narrative by using the Impact Estimation metric in Smith et al. [17]. The measure accounts for social confounders (e.g., community membership, popularity) and disentangles their effects from the causal estimation.

Briefly, the network potential outcome of a vertex v_i, denoted by $Y_i(\mathbf{Z}, \mathbf{A})$, is the number of tweets of that vertex under exposure to the narrative from the source vector $\mathbf{Z}$ and influence matrix $\mathbf{A}$. Precisely, $\mathbf{Z}$ is a binary vector where each Z_i indicates whether vertex v_i is a source in the network and $\mathbf{A}$ is the transpose of the adjacency matrix of the retweet network of the narrative (since influence flows in the opposite direction of a retweet). The unique causal contribution ζ_i of each vertex v_j to the narrative propagation over the entire network is defined as:

$$\zeta_j(z) = \frac{1}{N}\sum_{i=1}^{N}(Y_i(\mathbf{Z} = z_j+, \mathbf{A}) - Y_i(\mathbf{Z} = z_j-, \mathbf{A})), \tag{1}$$

where N is the is the total number of vertices, and ζ_j is the average difference in the number of tweets in the narrative with v_j as the source $z_j+ := (z_1, z_2, ...z_j = 1, ..z_N)$ versus with v_j as not the source $z_j- := (z_1, z_2, ...z_j = 0, ..z_N)$. Therefore this metric measures the average number of additional tweets generated in the narrative network by an individual's participation in the narrative. Since we observe $Y_i(\mathbf{Z} = z_j+, \mathbf{A})$ in the network, we estimate $Y_i(\mathbf{Z} = z_j-, \mathbf{A})$ by modelling Y_i as a Poisson Generalized Linear Mixed Model (GLMM) $Y_i \sim Poisson(\lambda_i)$ with parameters $(\tau, \gamma, \beta, \mu)$:

$$log\lambda_i = \tau Z_i + (\sum_{n=1}^{N_{hop}} s_i^{(n)} \tau \prod_{k=1}^{n} \gamma_k) + \beta^T \mathbf{x}_i + \mu + \epsilon_i, \tag{2}$$

where τZ_i is the effect of the primary source, $s_i^{(n)}$ is the amount of social exposure at the n-hop (for simplicity we only consider 1-hop neighbourhoods from the source) and $\mathbf{x}_i$ is the covariate vector for vertex v_i, indicating its popularity in the network (represented by the vertex's degree) and μ is the baseline effect on the entire population. The model parameters are estimated using the python package statsmodels[5] which estimates the posterior distribution using Laplacian approximation.

4 RESULTS

In this section, we first extract the dominant narratives during the two events, which we map semantically, illustrating the sequential building of narratives in increasing order of nuance, and then characterizing the influential actors who instigated and propagated the narratives using Impact Estimation from Section 3.5.

4.1 Mapping Propaganda Narratives

4.1.1 CAA/NRC. In Figure 1, we see that the discourse around protests against the CAA bill was dominated by narratives that were framed within nationalistic sentiment, ranging from notions of the protests being funded by organizations alleged to have terror links, having political goals driven by opposition parties, or simply being anti-Hindu in nature. In addition to the overall attacks on

[5]https://www.statsmodels.org/stable/index.html

Figure 1: CAA/NRC Narrative Map

the drivers of the movement, there were sub-narratives that vilified groups that took part in the demonstrations – such as claims that the Muslim women who sat in protest were being paid Rs.500 (USD 7) per day, and given a plate of biryani rice for their efforts. The efficacy of the multilingual narrative extraction method can be seen in Figure 2, where each cluster is mapped to a narrative.

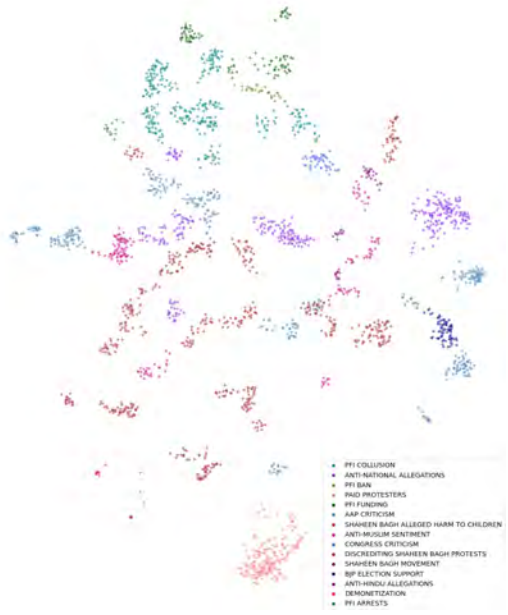

Figure 2: CAA/NRC Clusters

4.1.2 COVID-19. In the COVID-19 period, we see three broad narratives in Figure 3. First, it is critical that COVID-19 came at the heels of the CAA/NRC protests in India, thus elements of anti-Muslim social media narratives already had momentum. While the early social media narratives around COVID were focused on China or fake cures, they quickly devolved into anti-Muslim messaging, just as the first wave of the disease took hold in India.

The Tablighi Jamat story has several sub-strands that tie into a larger narrative of othering Muslims. First, the congregation is blamed for skirting the rules on gatherings (later shown to be untrue), a notion extended to suggest disdain for rules in the Muslim

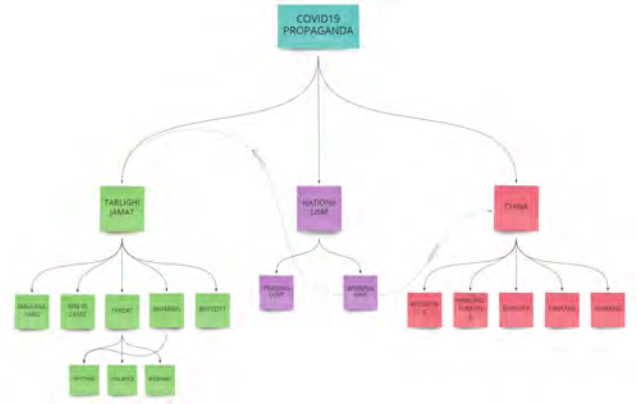

Figure 3: COVID-19 Narrative Map

community. Second, the congregants were presented as bringing the virus, then spreading it through its members, throughout India - suggesting a conspiracy. Third, the spread narratives were accompanied by imagery or stories of Muslim individuals as spitting in public or vendors licking food meant to be purchased by Hindu patrons, suggesting a lack of hygeine in Muslims and a deliberate intent to spread the virus, images of confrontations with doctors, suggesting a lack of respect for authority. The messaging also used derivatives of two terms "bomb" and "jihad" thus "coronabomb" or "biojihad" referring to terrorism. These eventually led to tweets calling for boycotts of Muslims businesses. Eventually, the reemergence of China narratives piggybacked on the anti-Muslim messaging, proposing that China manufactured the virus, and colluded with Pakistan. The positive-themed narratives align with nationalism – mainly praising the government for its good handling of the pandemic. Similar to CAA/NRC, we see in Figure 4 that each cluster is mapped to a narrative, thereby indicating the generalizability of our narrative extraction methodology.

4.2 Narrative Building and Propagation

Moreover, in order to identify the influential actors who instigate and propagate the dominant narratives studied previously, we use the Impact Estimation scores described in Section 3.5 to quantify the causal effect of each individual in the propagation of a narrative in the network, for all major narratives across both events.

An analysis of the top 20 accounts for each narrative reveals that a group of individuals are consistently significant sources of multiple narratives across all events. This is highlighted in Figure 5 and 6 where individuals with the highest impact are shown per narrative. The categories appearing on the top are politicians, media houses, journalists and writers, all of who are either members of the ruling party, or pro-government based on their profile descriptions. The impact values show that a few accounts are common conduits for the propagation and building of narratives. Due to this, they are able to craft and propagate narratives starting from everyday language and snowballing into extreme speech and violence spearheading instigation and building propaganda campaigns.

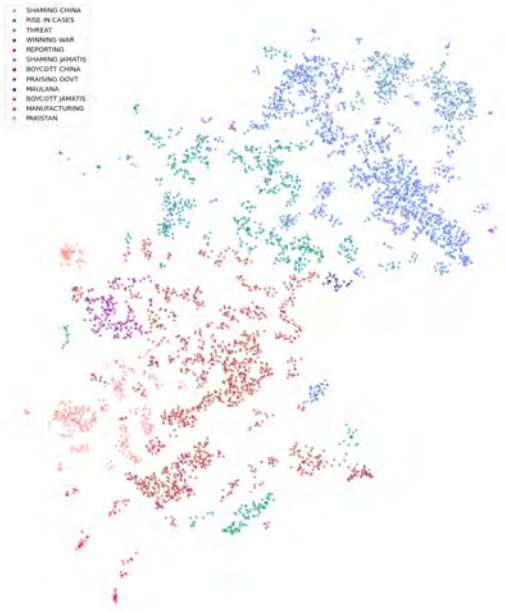

Figure 4: COVID-19 Clusters

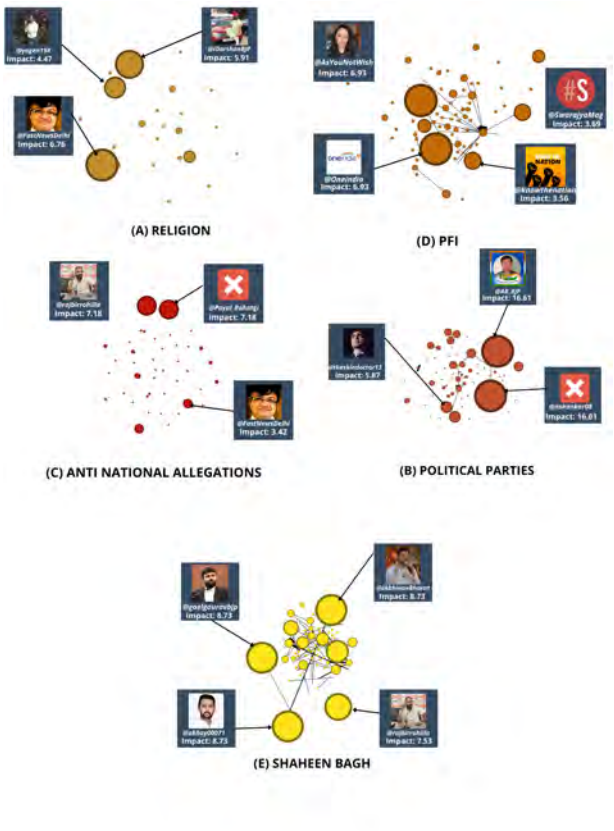

Figure 5: Impact Estimation CAA/NRC

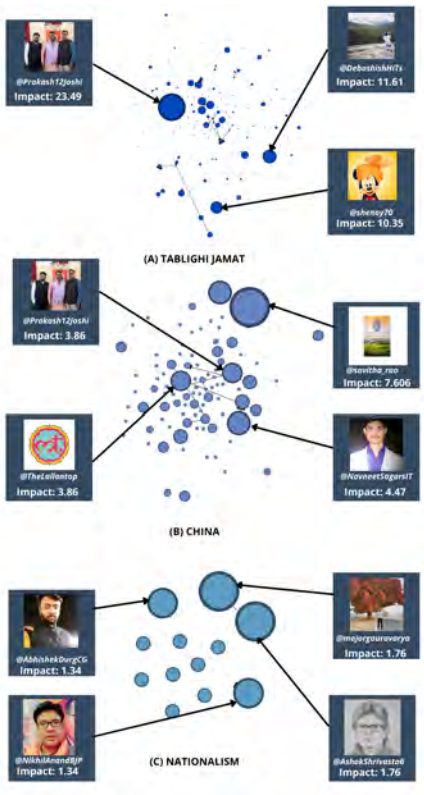

Figure 6: Impact Estimation COVID 19

5 DISCUSSION AND FUTURE WORK

The nodal role of a group of individual Twitter accounts across all the narratives show that a small number of effective and networked accounts can have a significant impact on propagating a specific discourse on social media. We see that despite precipitating events of viral social media significance being fundamentally different – such as a policy event related to immigration, and a public health emergency, the same set of accounts can act as connectors to turn a narrative towards an accepted dogma, in this case the notion of Muslims as a political enemy.

The findings and methodologies in our paper lay the foundation for exploring future directions of propaganda research. A complete deconstruction of propaganda propagation would require an analysis of temporal elements in narrative building and studying coordinated messaging to understand narrative evolution and intent to manipulate online discourse. A nuanced analysis of linguistic elements in propagandistic messaging, for instance the use of humour etc. is required to understand which narratives get amplified in the social networks. Moreover, influential actors that emerge through Impact Estimation can be studied in conjunction with Twitter metadata like following, location etc. in order to regulate such malicious accounts.

REFERENCES

[1] Mack Blackburn, Ning Yu, John Berrie, Brian Gordon, David Longfellow, William Tirrell, and Mark Williams. 2020. Corpus development for studying online disinformation campaign: a narrative+ stance approach. In *Proceedings for the First*

International Workshop on Social Threats in Online Conversations: Understanding and Management. 41–47.

[2] Gillian Bolsover and Philip Howard. 2017. Computational propaganda and political big data: Moving toward a more critical research agenda. , 273–276 pages.

[3] Hadley Cantril. 1938. Propaganda analysis. *The English Journal* 27, 3 (1938), 217–221.

[4] Saloni Dash, Dibyendu Mishra, Gazal Shekhawat, and Joyojeet Pal. 2021. Divided We Rule: Influencer Polarization on Twitter During Political Crises in India. *arXiv preprint arXiv:2105.08361* (2021).

[5] Walter R Fisher. 1985. The narrative paradigm: An elaboration. *Communications Monographs* 52, 4 (1985), 347–367.

[6] Jeffry Halverson, Steven Corman, and H Lloyd Goodall. 2011. *Master narratives of Islamist extremism*. Springer.

[7] Garth S Jowett and Victoria O'donnell. 2018. *Propaganda & persuasion*. Sage publications.

[8] Sangeeta Mahapatra and Johannes Plagemann. 2019. Polarisation and politicisation: the social media strategies of Indian political parties. (2019).

[9] Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeno, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. 2020. A survey on computational propaganda detection. *arXiv preprint arXiv:2007.08024* (2020).

[10] Leland McInnes and John Healy. 2017. Accelerated hierarchical density based clustering. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, 33–42.

[11] Leland McInnes, John Healy, and James Melville. 2020. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv:1802.03426 [stat.ML]

[12] 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.

[13] Anmol Panda, Ramaravind Kommiya Mothilal, Monojit Choudhury, Kalika Bali, and Joyojeet Pal. 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 (2020), 1–14.

[14] Ammar Rashed, Mucahid Kutlu, Kareem Darwish, Tamer Elsayed, and Cansın Bayrak. 2020. Embeddings-Based Clustering for Target Specific Stances: The Case of a Polarized Turkey. *arXiv preprint arXiv:2005.09649* (2020).

[15] Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. *arXiv preprint arXiv:2004.09813* (2020).

[16] Karishma Sharma, Emilio Ferrara, and Yan Liu. 2021. Characterizing Online Engagement with Disinformation and Conspiracies in the 2020 US Presidential Election. *arXiv preprint arXiv:2107.08319* (2021).

[17] Steven T. Smith, Edward K. Kao, Erika D. Mackin, Danelle C. Shah, Olga Simek, and Donald B. Rubin. 2021. Automatic detection of influential actors in disinformation networks. *Proceedings of the National Academy of Sciences* 118, 4 (2021). https://doi.org/10.1073/pnas.2011216118 arXiv:https://www.pnas.org/content/118/4/e2011216118.full.pdf

[18] Douglas Wilbur. 2019. *Understanding the Propagandist's Narrative: Applying the Narrative Paradigm Theory to Peer-to-Peer Propaganda*. SAGE Publications Ltd.

[19] Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, et al. 2019. Multilingual universal sentence encoder for semantic retrieval. *arXiv preprint arXiv:1907.04307* (2019).