Understanding Online Protests: Unveiling Strategies, Collective Narratives, and Harmful Behaviors

BY

KUMARI NEHA

Under the supervision of

Dr. Arun Balaji Buduru

Prof. Ponnurangam Kumaraguru (“PK”)

INDRAPRASTHA INSTITUTE OF INFORMATION TECHNOLOGY DELHI

February 19, 2024

New Delhi, 2024

Understanding Online Protests: Unveiling Strategies, Collective Narratives, and Harmful Behaviors

A THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

BY

KUMARI NEHA

COMPUTER SCIENCE AND ENGINEERING
INDRAPRASTHA INSTITUTE OF INFORMATION TECHNOLOGY DELHI

NEW DELHI- 110020

February 19, 2024

THESIS CERTIFICATE

This is to certify that the thesis titled **UNDERSTANDING ONLINE PROTESTS: UNVEILING STRATEGIES, COLLECTIVE NARRATIVES, AND HARMFUL BEHAVIORS**, submitted by **Kumari Neha**, to the Indraprastha Institute of Information Technology, Delhi, for the award of the degree of **Doctor of Philosophy**, is a bona fide record of the research work done by her under our supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Dr. Arun Balaji Buduru
Thesis Supervisor
Assistant Professor
Dept. of Computer Science & Engineering
IIIT Delhi, 110 020

Prof. Ponnurangam Kumaraguru (“PK”)
Thesis Supervisor
Professor
Dept. of Computer Science & Engineering
IIIT Hyderabad, 500 032

Place: New Delhi

Date: February 19, 2024

Dedicated to my mother, who stood by my side, through the storms and sunshine!

ACKNOWLEDGEMENTS

This thesis represents the culmination of years of rigorous research, analysis, and dedication. It is the result of countless hours of study, reflection, and collaboration toward a journey that seemed like an uphill battle when I first started. As my journey as a Ph.D. research scholar comes toward completion, there are a number of people who I would like to extend my gratitude and thanks, without whom the journey would be impossible.

I would like to extend my deepest appreciation to Prof. Ponnurangam Kumaraguru, my thesis co-advisor, and Dr. Arun Balaji Buduru my thesis advisor for their unwavering support and guidance during my Ph.D. journey. Their invaluable insights and constructive feedback have been instrumental in shaping the direction and quality of my research. I am truly fortunate to have had mentors who not only believed in my potential but also challenged me to push the boundaries of my knowledge and capabilities. I will forever remain indebted to my advisors for their relentless dedication and passion that has shaped me as a researcher and individual.

I would like to extend my gratitude to the faculty members and experts who served on my thesis committee. Their valuable input and critical assessment enriched my work and motivated me to strive for excellence. Each one of them contributed unique perspectives that have undoubtedly strengthened the credibility and significance of my findings.

I would also take this moment to appreciate the IT team of IIIT Delhi. The dedication and support the IT team extended during the COVID-19 lockdown was phenomenal. It is the IT support team, who made it possible for me to chase deadlines, without worrying about any server-related issues. Additionally, I am thankful to the administrative staff, and everyone else who played a role in facilitating my research and providing logistical support. I would like to acknowledge the academic institutions, libraries, and online resources that provided access to a wealth of knowledge, making my research both comprehensive and well-informed.

The students that I have worked with throughout my Ph.D. life, remain the most important part of the journey. I have been blessed throughout my journey to have worked with the smartest of kids and their role in the thesis are irreplaceable, to say the least. More than they have learned working with me, I have learned from the students, and I am extremely proud of each one of them. I extend my gratitude especially to Vibhu Argawal, Tushar Mohan, Vishwesh Kumar, Abhishek Gupta, Saurav Chattani, Garvit Gupta, Shreyans Mongia, Shubham Singh, Sonali Singhal, Shashank Srikanth, and Shwetanshu Singh.

My sincere appreciation goes to my family members who have been my pillars of strength throughout this academic endeavor. Their unwavering support, encouragement, and belief in my abilities have been the driving force behind my success. From the countless late nights to the moments of self-doubt, their love and encouragement never wavered, and I am forever indebted to them for their sacrifices and encouragement. My mother's support for me throughout the journey played a very vital role in keeping me sane.

Last, but not least, I would like to thank my husband, for his patience while dealing with me, especially around paper rejections. I am fortunate to have found a gem of a partner. I extend my deepest appreciation and love to you for supporting me in the journey.

Completing this thesis would not have been possible without the collective efforts and support of all these individuals, and I am sincerely grateful for their contributions.

With heartfelt appreciation,

Kumari Neha

KEYWORDS: Social Media Protest, Online Campaign, Twitter Activism, Unsupervised Stance Detection, Hate Speech Detection, Coordinated Behavior, Inauthentic Activity, Citizenship Amendment Act, Black Lives Matter, Kill The Bill, Farmers Bill Protest, Social Network Analysis.

ABSTRACT

Protests (or movements) are a form of collective sociopolitical action in which members with similar beliefs express their objections to a cause or situation. Often, a heated debate during protests on social media, such as Twitter, may lead to divided users and divergent discourse. On the bright side, studying divergent discourse on contentious topics can help infer the collective perceptions of people in terms of their collective narratives. On the dark side, narratives shared during protests may become susceptible to various harmful influence operations, disrupting society's peaceful fabric. This thesis aims to understand digital strategies to organize protests, identify collective narratives shared during protests, and identify harmful behaviors with potential online and offline consequences. We focus on hate speech and coordinated inauthentic behavior as proxies for harmful conduct during online protests. We divide the thesis into four parts: (i) Understanding strategies used for conducting online protests, (ii) Detecting and analyzing collective narratives shared during protests, (iii) Detecting and analyzing opposing stances during the protest, inclusive of authentic and inauthentic actors, (iv) Detecting and analyzing harmful behavior during protest.

To focus on the strategies used for conducting protests on social media, we examine the protest over the cause of the death of Sushant Singh Rajput (#SSR) [1] on Twitter. Study of shared hashtags and retweets during #SSR protests reveals a combination of centralized and decentralized information aggregation strategies in retweet networks, suggesting a mix of self-motivated individuals and organized entities. Next, we propose an unsupervised clustering-based framework to focus on the collective narratives shared during protests. Our findings suggest clusters of call-to-action and on-ground activities across protests under study. Next, we delve into the opposing stances formed during an online protest, using #CAA protest on Twitter as a case study. We build an unsupervised stance detection technique to identify users' stances and analyze their

[1] An Indian actor whose untimely death by suicide led to an online movement towards his cause of death.

content, follower networks, and inauthentic behavior (i.e, bots, suspended users, and deleted users). Our findings suggest homophily (i.e., users of the same stance follow each other on Twitter) in follower network and presence of edges between authentic and inauthentic users, suggesting their connectedness. Finally, we focus on hate speech and coordinated inauthentic behavior as proxies for harmful conduct and study their contribution during the divergent discourse on #CAA. To this end, we built a multi-task classification model with hate speech detection as the primary task and stance detection as an auxiliary task and obtained an F1 score of 0.92. Our findings suggest that more hateful users produced more tweets, received faster retweets, and held a central position in the retweet network. Regarding coordinated inauthentic behavior, our findings suggest that coordinated communities, which were highly inauthentic, showed the highest clustering coefficient towards a greater extent of coordination.

In conclusion, this thesis examines strategies, collective narratives, and harmful behavior within protests, comprehensively exploring the intricate facets of online activism. To summarize, the research contributions of the thesis are: - (i) Analyze protest hashtags and retweet communities to provide insights into protest strategies, (ii) Build an unsupervised collective narrative detection technique, (iii) Build an unsupervised stance detection technique for user-level stance detection for multi-lingual Twitter data, (iii) Build automated hate speech detection method for opposing stances, (iv) Build a framework for coordinated communities in opposing stances. Through our research, we aim to foster a more secure digital environment for participants in online protests.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

OSM Online Social Media

OSN Online Social Networks

ML Machine Learning

DL Deep Learning

SNA Social Network Analysis

SSR Shushant Singh Rajput

CAA Citizenship Amendment Act

AI Artificial Intelligence

FP Farmer Protest

KTB Kill the Bill protest

URL Uniform Resource Locator

BLM Black Lives Matter

CP Counter-protesters **P** Protesters

MTL Multi-Task Learning

FRT First Retweet Time

CTA Call-to-action

OGA On-ground activities

SKEP Skepticism

GRV Grievances

CHAPTER 1

INTRODUCTION

Recent technological advancement has transformed Online Social Media (OSM) platforms into a significant place for debate around socio-political phenomena, helping users express their opinions and mediate social interactions [154]. The abundance of communication capabilities on social media has produced a resilient network of people responsible for a continuous flow of information between the offline and the online ecosystem [47]. With the help of provided communication capabilities, users can identify like-minded people who boost their belief system on social media [72]. While identifying like-minded people on social media gives a sense of belongingness and helps fight for a cause [32], it sometimes leads to a polarized information flow between users who are ignorant of the other side [91]. Moreover, the tendency of users to adjust interests, opinions, and actions according to ongoing developments (e.g., call to action on the online ecosystem for offline protest) introduces a feedback effect, where the offline and the online ecosystem might affect each other [157]. This feedback loop becomes particularly impactful during protests, as information circulating about the protest can influence people's judgments and actions.

Protests and social movements are scarce; however, they may lead to dramatic outcomes when they occur [76]. Social media, such as Twitter, has become central to organizing and developing collective action, such as online protests worldwide. Manifestation of collective identity (for example, #wearethe99percent launched by the Occupy Wall Street movement) is accompanied by a set of goals that provides users with a collective sense of self and what they stand for [74]. Often a heated debate between participants on Twitter during protests may lead to divided users and divergent discourse around the topic [70]. In particular, social-technical convergence has paved way for social sensing, where humans act as data sensors that continuously post about ongoing phenomena [206]. On the bright side, studying divergent discourse on contentious topics can help infer the collective perceptions of people in terms of collective narratives. Information from the collective narratives produced by human sensors can provide data-driven decision support to policy-makers and stakeholders [1] for making an informed decision and adjusting any interventions according to the needs of people [12].

Initially hailed as a powerful tool for promoting diverse perspectives, critical thinking, and democratic discussions, social media's global reach and information-sharing capabilities presented great opportunities [190]. However, as it became the primary source of information for many, the convergence of social and technological factors began to pose significant risks to society [38]. For example, due to the innate human tendency of "confirmation bias", individuals often gravitate towards consuming information that aligns with their existing beliefs. As a result, the potential benefit of exposure to diverse perspectives is greatly restricted [221]. Moreover, the algorithms governing news feed curation and the dynamics of social networks contribute to the reinforcement of selective exposure mechanisms [42]. This transformation has resulted in the formation of echo chambers, where users tend to reinforce their own opinions and biases on a given topic instead of engaging in a truly democratic discussion [55].

Apart from "confirmation bias", selective exposure mechanisms, and formation of echo chambers, online environment created by social media also becomes a perfect breeding ground for malicious activities ranging from promoting terrorist activities [23], disruption of foreign campaigns [22], and inducing fear among fragile audience [9]. Threats to secure society can range from genuine users involved in occasional harassing fragile people [62] to more profound inauthentic actors who purposefully become part of an online discussion with ill-intention to create polarization [78], spread propaganda [189], among other intentions. Despite the efforts of the platform to remove malicious content, posts made by malicious accounts may reach a wider audience before being suspended from the platform [172]. Malicious accounts constantly adapt their content to evade platform regulations, camouflaging themselves within benign online social media (OSM) content to achieve viral reach before any intervention occurs. The feedback loop between online and offline ecosystems amplifies the impact of malicious content. For instance, vaccine debate on social media sparked an infodemic online and hindered the progress of vaccinations offline [75].

Our primary objectives are to understand how protests are conducted digitally, identify collective narratives shared during protests, and identify harmful behaviors that

[1] Examples of stakeholders are political parties, advocacy groups, religious and ethnic minorities, etc.

may have potential online and offline consequences. Specifically, we focus on hate speech and coordinated inauthentic behavior as proxies for harmful conduct during online protests. We investigate these factors to uncover underlying elements contributing to negative outcomes and help develop effective strategies to mitigate such risks.

1.1 Understanding Protest Strategies

In the past decade, the prevailing approach to studying social movements emphasized the significance of shared grievances [183] and potential routes for addressing them as fundamental prerequisites for collective actions [66]. However, more recently, the previously strong hypothesis regarding the centrality of grievances has shifted to a weaker one [200]. The current assumption suggests that any society always has enough discontent to foster a movement as long as the campaign is effectively organized [131]. Protests, as observed during the Arab Spring at the beginning of the decade, involve a dynamic interplay between two distinct logics. First is the formal association of organizational resources, known as the logic of collective action. Second, is the users' inclination to share personalized content on social media, referred to as the logic of connective action [29]. Social media has become a prime site where protests are created, channeled, and contested [74]. Social media has emerged as a central platform for organizations and individuals to share protest-related content, and its potential contributions are manifold. For example, the study of protesters' posts on social media on the 'no ban, no wall' protest was done to reduce prejudice towards a given section of society [212]. In another instance, the study of social media posts was used to understand the dogmatic mindset of the users of a marginalized community [62]. The new direction of social movement research has attracted much attention in two directions: the movement-media relationship [223] and social movement strategy [49].

1.1.1 Challenges

Understanding the major strategies adopted for a socio-technical protest requires (i) extracting actionable and concise knowledge from the online ecosystem; (ii) identifying and characterizing prime advocates involved in the online social movement; and (iii) designing suitable techniques to demystify online strategies used by activist for the movement online. Research on social movements in non-western countries is limited compared to the extensive body of work focused on protests in the Western context. Understanding and analyzing protests outside Western societies poses significant challenges [159]. Gathering user-generated content from OSM comes with its fair share of challenges, including incompleteness, information overload, and multidimensional information (text, images, videos). Another major challenge concerning study protests in non-western contexts is the barriers posed by content shared in low-resource languages [85]. Online protests may be single-sided [208] and rich in discourse [70]. We address the two challenges mentioned above: (i) broader the scope of studying the protest in non-western content; (ii) addressing the low resource languages for protest study.

1.1.2 Solutions

Social Network Analysis, when combined with AI techniques, facilitates the modeling of information flow between users in online social networks (OSNs). It allows for identifying key actors and provides valuable insights into people's perceptions and behaviors within networks [117]. Using AI and SNA to understand protest-related activities, we can delve into the user's stance on a particular debate [70], discontent with a political change [208], among other knowledgeable insights. By focusing on the retweet mechanism as a form of protest participation [36], we can study strategies for protests. Specifically, we can categorize participants as generators (creating content related to the protest) and drivers (propagating the protest through retweets) to comprehensively analyze protest strategies.

1.2 Understanding Protest Narratives

Narratives are verbal, graphic, or written interpretations of related events and participating actors, evolving through a given duration [158]. Using a hashtag to build a collective narrative makes Twitter one of the prime spots for conducting protest [208]. Conducting a detailed analysis of the diverse narratives within a protest can help us to gain knowledgeable insights into the protest, understand people's perceptions, and shed light on the main focus of the protest. Gaining concise knowledge from the shared narratives is especially important as highly influential users can disseminate information to a wider audience, potentially shaping the course of the protest [208].

1.2.1 Challenges

Extracting concise knowledge from social media protests encompasses several challenges. One of the major challenges is that protests might occur due to different factors, such as racial and religious discrimination, some political outcomes, etc. Protest's uniqueness and subjectivity pose a significant challenge to finding common topics across protests [144]. Another major challenge for understanding common narratives across protests is the barriers posed by content shared in low-resource languages in non-western countries [85].

1.2.2 Solutions

Extracting relevant topics from large OSM discussions requires a combination of AI and SNA. We can classify a huge collection of data, understand hidden patterns of information in the data, and use the network of users and content to understand the user's perception and beliefs of the topic of discussion. Catering to the subjective nature of protests, we propose to use AI and SNA to develop an unsupervised framework for narrative detection.

1.3 Understanding Online Threats and Harmful Behavior

Posts on OSM platforms are susceptible to the spread of hate, propaganda, manipulation, and disinformation campaign, among other harmful behavior [78]. Unreliable content shared by various inauthentic users gets attention from unaware users, who can fall prey to harmful activities [191]. Hate and fear speech on the platform might affect the debate on social media [169]. When a protest unfolds into discourse, there is a risk of inauthentic or disrespectful content on either side of the discourse [70]. To fully understand the interplay of inauthentic activity within the discourse, we need to focus on both sides of the discourse. Hence, we focus on the threats from inauthentic and harmful actors and their shared content on protest. Since the co-existence of harmful users online can lead to targeted hate being deliberated during the protest [169], we use hateful content as the first proxy for harmful behavior. Since protests are often coordinated in nature [118], we use the coordination in the protest as a medium and use inauthentic and hateful coordinated activity as the second proxy for harmful behavior.

1.3.1 Challenges

While protests are prevalent worldwide, those in non-western countries have been understudied due to a lack of diverse language representations. Recently, we have seen a slight rise in the study of harmful behavior in non-western countries [150]. The major challenge with studying online harmful behavior continues to be the barrier of low-resource language. Understanding the political and moral values of a country or community poses another significant challenge in studying protests [86; 164], particularly when it comes to examining harmful behaviors like hate. It is crucial to navigate and comprehend the diverse political and moral awareness within a specific context. In terms of inauthenticity, the detection of bots and platform-aided suspension of malicious accounts are fostered by the OSM platform. However, malicious accounts continually adjust their content to evade platform regulations, disguising themselves within harmless OSM content to quickly gain widespread attention before detection or intervention occurs.

1.3.2 Solutions

Since the first proxy for harmful behavior is taken as hateful content, we devise a framework to understand the hateful content during the #CAA discourse. To address the subjectivity of political and moral perspectives, we propose annotating hateful tweets during discourse and developing a classification model for hate speech, including opposing stances, within the context of protests. By studying hate from opposing viewpoints during protests, we hypothesize that valuable insights can be gained, which may have broader applicability to similar protests. We account for inauthenticity through Twitter suspension [2] and bot detection techniques [218]. We suggest investigating coordinated behavior within a discourse to tackle the adaptability of harmful users evading platform regulations. By identifying coordinated communities through shared mechanisms like hashtags, retweets, and mentions, we aim to demystify their presence. We hypothesize that studying the interplay of different inauthentic and harmful behaviors on Twitter can shed light on the complexity of vulnerabilities during protest participation.

1.4 Legal and Ethical Concerns

Users' profile data in OSM may be publicly available, but it is important to recognize its inherent sensitivity. When users participate in a protest or express their opinions, they might not anticipate the potential use of their data by anyone, especially when sensitive topics are involved. In our study, we have only utilized publicly available information and have not attempted to explore user-level demographic data. We aim to focus on understanding public perception of various protests rather than individual perspectives, thus safeguarding user privacy. In typical studies like ours, sharing tweet IDs is a common practice. However, due to the sensitive nature of the campaign, there is a risk associated with sharing tweet IDs. Sharing tweet IDs makes it possible to extract user-level information from these tweets, posing a threat to privacy. Consequently, we have not disclosed the tweet IDs used in our studies. Instead, we share tweets and user-level features that do not reveal personal information (such as profile names, descriptions, usernames, and other identifiable details). This approach ensures that users' privacy is maintained while allowing us to analyze relevant data for our research.

1.5 System Requirements

For our experiments, we used a Linux-based system with Xeon(R), an x86 microprocessor developed by Intel with a system memory of 62GB. We ran our machine learning and deep learning models using NVIDIA-SMI GPU with a driver version of 440.33.01 and installed Cuda version 10.2. Another server used for training our deep learning models was the Nvidia RTX 3090 GPU system with an installed Cuda version of 11.3.

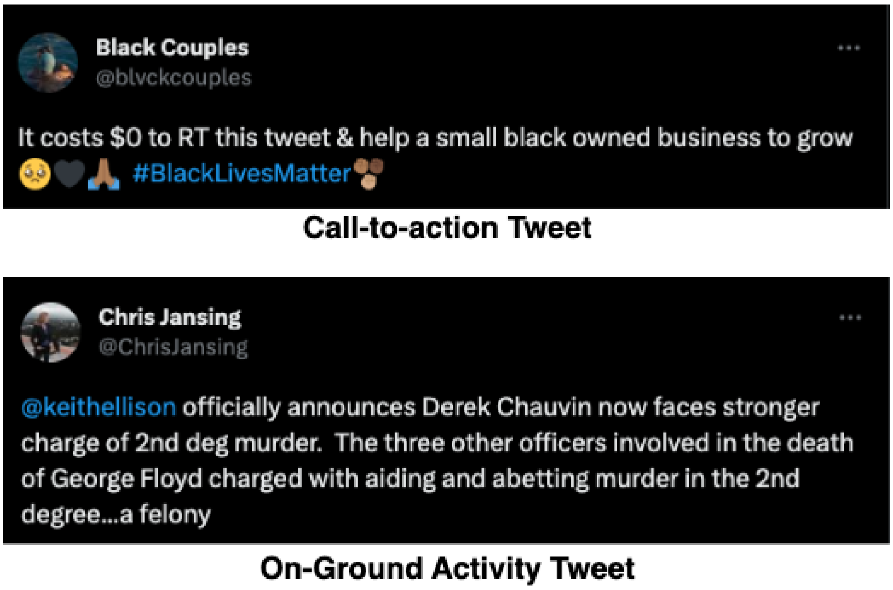
[2] https://help.twitter.com/en/managing-your- account/suspended-twitter-accounts

1.6 Contributions

We divide the thesis into four parts: (i) Understanding the strategies used for online protest, (ii) Detecting and analyzing collective narratives shared during protests, (iii) Detecting and analyzing the opposing stances during the protest, inclusive of authentic and inauthentic actors, (iv) Detecting and analyzing harmful behavior during protest.

1.6.1 Understanding Strategies Used For Online Protest

First, we examine how Twitter activists build a diverse global support network and challenge the dominant narrative during an online protest. As a case study, we analyze strategies of the protest surrounding the cause of the death of Indian actor Sushant Singh Rajput (#SSR). Despite the cause of death being reported as a suicide by the officials, a counterpublic movement emerged, discussing alternative theories such as nepotism and murder, leading to an online protest on various social media, including Twitter. Counterpublics [104] are defined as marginalized communities that distribute messages to diverse social groups, raise awareness, and challenge dominant narratives. Counterpublics leverage hashtags to build a diverse support network and share content on a global platform that counters the dominant narrative. Our first work applies the framework of connective action to the counter-narrative campaign over the cause of death of #SushantSinghRajput. We combine descriptive network, modularity, and hashtag-based topical analysis to identify the campaign's three major mechanisms: generative role-taking, hashtag-based narratives, and forming an alignment network toward a common cause. Using the case study of #SushantSinghRajput, we highlight how the connective action framework can be used to identify different strategies adopted by counterpublics for the emergence of connective action. Our findings indicate a connected community of counterpublics, combining centralized and decentralized information aggregation. Our results suggest a mix of self-motivated individuals and organized entities, raising concerns about the potential partial manipulation of the online campaign. We also found that different communities formed shared similar topics, showing that the shared content united users divided by communities.



Black Couples
@blvckcouples

It costs $0 to RT this tweet & help a small black owned business to grow 🥺🖤🙏🏽 #BlackLivesMatter🏾🏿🏽

Call-to-action Tweet

Chris Jansing
@ChrisJansing

@keithellison officially announces Derek Chauvin now faces stronger charge of 2nd deg murder. The three other officers involved in the death of George Floyd charged with aiding and abetting murder in the 2nd degree...a felony

On-Ground Activity Tweet

Figure 1.1: Example tweet showing call-to-action and on-ground activity reporting for protests.

1.6.2 Detecting And Analyzing Different Collective Narratives Present During Online Protes

Next, we study and examine collective narratives shared during protests and their role in shaping and advancing collective opinions. To this end, we propose an unsupervised clustering-based framework to examine collective narratives shared during a protest. We focus on four protests: #CitizenshipAmendmentAct and #FarmersProtest in India, #KillTheBill protest in the U.K, and #BlackLivesMatter in the U.S. to study the collective narratives across protests. Next, we investigate the evolution of identified converging narratives across the protests. We further identify the most influential participants in a protest and study their contribution to spreading various narratives. Our results suggest that clusters with call-to-action tweets [3] and on-ground activity reporting tweets [4] are common narratives across all protests. Figure 1.1 shows the example tweets for call-to-action and on-ground activity used in protest. Analysis of the evolution of narrative suggests that the call-to-action narrative is the most consistent during the protest. Community detection over the retweet network across protests suggests narrative-centric community formation.

[3] Call-to-action category represent the tweets that urged the users to participate in the protest.

[4] tweets that narrate the current and ongoing development of the protest in real time

1.6.3 Detecting And Characterizing Opposing Stances During Online Protest, Concerning Authentic And Inauthentic Actors

Since contentious topics are prone to divergent discourse, we delve into the opposing stances formed during an online protest in the next part of the thesis. We use India's #CitizenshipAmendmentAct protest as a case study to investigate the opposing stances and the content they shared during the discourse. Keeping the campaign participants as the prime focus, we study 9,947,814 tweets produced by 275,111 users during the starting 3 months of protest. Our investigation of the opposing stances accounts for different authentic and inauthentic actors on the platform and compares their shared content and network structure. Among the opposing stances, users who opposed the Act were identified as protesters, while users who supported the Act were identified as counter-protesters. The opposing stances were further divided into authentic and inauthentic users based on whether they were genuine users (Authentic users) or were identified as bots, suspended, or deleted by Twitter (Inauthentic users). We contribute to being the first study to perform a fine-grain analysis of the contention around the #CitizenshipAmendmentAct on Twitter regarding opposing stances and authenticity vs. inauthenticity combined. Our findings show different themes in shared tweets among opposing stance users, while the following network of users suggests homophily among users on the same side of the discourse. Our findings also suggest the presence of inauthentic activities on both sides of the discourse and the presence of a path between authentic and inauthentic users in the following network suggesting the connectedness of inauthentic users to their authentic counterparts.

1.6.4 Detecting And Analyzing Harmful Behavior During Protest

Among the harmful behaviors, we first focus on disseminating hateful content during online protests. To this end, we study how hateful users exploited the elements of protest mobilization (i.e., *resources* defined as the engagement methods on Twitter such as tweeting, retweeting, etc. and *ability to use them*) during the divergent discourse on #CitizenshipAmendmentAct in India. We define hate speech as "any content that promotes violence against the opposing stance cohort, directly or indirectly threatening the people based on their race, ethnicity, national origin, religious affiliation, political ideology, and political affiliation" [174]. Since the user's stance plays a vital role in hateful tweet detection, we build a multi-task classification model with hate speech detection as the primary task and stance detection as an auxiliary task. Our model outperforms previous models catered towards Indian tweets [173], with an F1-score of 0.92. We further use our model to analyze the hateful users and tweets during the protest mobilization. Our key findings suggest that more hateful users produced more tweets and received faster retweets during the protest than non-hateful users. Across the opposing stances, hateful users held a more central position in the retweet network, indicating their reachability to genuine users. To delve deeper into the harmful activities in play during a protest, we combine different forms of harmful behavior with inauthenticity in our final part of the thesis. We use #CitizenshipAmendmentAct as a case study and decipher the various forms of inauthentic activities (bots, suspended users) and harmful behavior (hate speech and coordinated inauthentic behavior) exerted by the opposing stances during the online discourse. To this end, we identify the coordinated communities in the opposing stances, marked by the exceptional similarity between two users through different mechanisms such as hashtags, retweets, and mentions. Our key findings reveal that the most hateful, strongly coordinated communities of opposing stances also showed highly inauthentic behavior. Another key finding reveals that strongly coordinated communities that produced hate may not necessarily show a high degree of inauthentic behavior.

This thesis stands at the forefront of advancing our understanding of online protests by introducing methodologies and conducting a comprehensive study on protest related posts on Twitter. The study's novelty lies in its in-depth analysis of how Twitter activists construct diverse global support networks, challenge dominant narratives, and engage in harmful behaviors and opposing stances during online protests. The selection of Twitter as the platform for the study was done due to Twitter's influential role in shaping real-time conversations, global reach, and information dissemination during protests. [5] By specifically focusing on diverse protests such as the #SushantSinghRajput, #CitizenshipAmendmentAct, #FarmersProtest, #KillTheBill, and #BlackLivesMatter, the research offers a nuanced understanding of varied protest contexts. The selection of protests for the study, facilitates a comparative analysis, revealing commonalities and distinctions in collective narratives and protest dynamics across different socio-political

[5] https://www.pewresearch.org/journalism/2009/06/25/iran-and-twitter-revolution/

landscapes. The research in this thesis provides valuable insights for future studies, policymakers, and stakeholders seeking to navigate the complexities of online activism and create a more secure digital environment during protests.

1.7 Thesis Outline

The rest of the thesis is organized as follows. Chapter 2 discusses the related literature. Chapter 3 discusses the strategies used in online protest. Chapter 4 describes our approach and findings of collective narrative detection techniques around protests. Chapter 5 focuses on detecting and characterizing opposing stances during online protests. Chapter 6 and Chapter 7 discuss harmful behavior during online protests, considering hate speech detection and coordinated inauthentic activity during the protest, respectively. Finally, Chapter 8 concludes the thesis, discusses limitations, and proposes future directions.

1.8 Publications Based On Thesis

- **Neha, Kumari**, Tushar Mohan, Arun Balaji Buduru, and Ponnurangam Kumaraguru. "Truth and travesty intertwined: a case study of #ssr counterpublic campaign." In Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 643-648. 2021.

- **Neha, Kumari**, Vibhu Agrawal, Arun Balaji Buduru, and Ponnurangam Kumaraguru. "The pursuit of being heard: An unsupervised approach to narrative detection in online protest." In 2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 256-260. IEEE, 2022.

- **Neha, Kumari**, Vibhu Agrawal, Vishwesh Kumar, Tushar Mohan, Abhishek Chopra, Arun Balaji Buduru, Rajesh Sharma, and Ponnurangam Kumaraguru. "A tale of two sides: Study of protesters and counter-protesters on #citizenshipamendmentact campaign on Twitter." In Proceedings of the 14th ACM Web Science Conference 2022, pp. 279-289. 2022.

- **Kumari, Neha**, Mrinal Anand, Tushar Mohan, Ponnurangam Kumaraguru, and Arun Balaji Buduru. "CamPros at CASE 2022 Task 1: Transformer-based Multilingual Protest News Detection." In Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE), pp. 169-174. 2022.

- **Neha, Kumari**, Vibhu Agrawal, Saurav Chhatani, Arun Balaji Buduru, and Ponnurangam Kumaraguru. "An Unsupervised Narrative Detection Framework to Demystify Online Protest Composition." (2023). (*Under Review*)

- **Neha, Kumari**, Vibhu Agrawal, Tushar Mohan, Anmol Goyal, Arun Balaji Buduru, Rajesh Sharma, and Ponnurangam Kumaraguru. (2023). Detection and Characterization of Hate Seeping in During Protest Mobilization. (*Under Review*)

- **Neha, Kumari**, Vibhu Agrawal, Saurav Chhatani, Rajesh Sharma, Arun Balaji Buduru, and Ponnurangam Kumaraguru. (2023). Understanding Coordinated Communities Through The Lens Of Protest-centric Narratives: A Case Study on #CAA Protest. (Accepted at The 18th International AAAI Conference on Web and Social Media 2024)

CHAPTER 2

LITERATURE REVIEW

Protests are a form of collective sociopolitical action in which members with similar beliefs express their objection to a cause or situation [15]. Time and again, the world witnesses protests over a government policy, bill [160; 212], or the government itself [190]. Social media use to conduct protests has been evident from the start of the past decade [76]. Previous researchers have used social media as a social sensor to predict when the protest will take place, i.e., protest prediction [137; 77], and whether and how social media contributes to the explosion during protests, i.e., protest recruitment [76]. More recently, researchers have focused on understanding emotional dynamics during protest [49] and understanding the underlying opinion of participants to counter societal challenges such as dogmatism [62] and prejudice against immigrants [212]. Other recent focus includes opinion modeling, which reflects and justifies the belief or judgment of a person towards a target entity, irrespective of having the same stance [84].

Twitter is one of the prime social media platforms for conducting online protests. The role of Twitter in mass mobilization against fraudulent elections in Moldova and Iran led to the coined phrase "Twitter Revolution" in 2009.[1] Although "Twitter Revolution" was celebrated at first for shifting power to people, research over the past decade shows the double-edged sword of using social media for protest, ranging from influence operations [22], manipulation [24], coordination [95] among others. The involvement of inauthentic users has become more prevalent on the platform [64]. The various forms of manipulation of the debate are studied with regards to bots [179; 30; 197; 40], pre-defined campaign toolkit users [106], co-ordinated accounts [148; 150; 180], or trolls [120; 78]. Hence, the presence of inauthentic activities and harmful behavior on the platform during protest conduct cannot be denied.

In this chapter, we first discuss the previous protest-related social media studies catered toward understanding strategies adopted during protests. Next, we study related

[1] https://www.pewresearch.org/journalism/2009/06/25/iran-and-twitter-revolution/

work with respect to collective narratives, online threats, and harmful behavior during online social media protests.

2.1 Understanding Protest Strategies

This section delves into strategies adopted by different social media-mediated protests over the past decade.

Financial Crisis Protest of Spain, 2011: In 2011, Spain witnessed a mass mobilization in response to the political management of the financial crisis, fueling a call for renewed democratic representation. The primary aim was to orchestrate a significant protest on May 15, uniting people in 59 cities. Participants established camps in city squares from May 15 to May 22, strategically coinciding with regional and national elections. Despite its initial vigor, the movement gradually waned and dissipated.

González-Bailón *et al.* [76] examined the recruitment dynamics of the protest, aiming to determine whether mass mobilization relies on weaker connections (broadcasting links) or stronger connections. Over 30 days, they analyzed 581,750 protest-related tweets from 87,569 users. By constructing a network based on followers and retweets among the most active users, they applied threshold-based metrics to uncover recruitment patterns. Their findings indicated that multiple exposures, rather than repeated exposure from a single individual, played a vital role in fostering social contagion and motivating users to join the protest.

Egyptian Uprising, 2011: A wave of political uprisings swept the Arab world, including Egypt, triggered by Tunisia's successful demonstrations in 2011. In Egypt, the protest aimed to topple the authoritarian regime. The uprising commenced on January 25, 2011, and persisted for 18 days until President Mubarak resigned on February 11, 2011. Initially, the protest unfolded as a peaceful demonstration, but on February 2nd, a significant shift occurred as clashes emerged between pro-Mubarak and anti-Mubarak groups. This escalation led to episodes of violence, with pro-Mubarak individuals acting as "thugs" and assaulting anti-Mubarak activists.

Starbird and Palen [190] studied the interplay of users involved in online activism and those on the ground during the protest. Starbird and Palen [190] collected hashtag-based tweets and user information from Twitter API and studied the diffusion of the most popular tweets during the protest. The most popular tweet during the protest included the text “Uninstalling dictator” with a progress bar and reporting of on-ground activity of the users at the protest’s location. The “Uninstalling dictator” variations with progress bar tweets appear 19,836 in the dataset. The retweet was found to be the most prominent feature for the propagation of the tweet during the protest.

Brazil Summer Protest, 2011: During the summer of 2011, protests in Brazil began with the disruption caused by increased public transport fares. However, as the protest progressed, it expanded its focus to include political corruption and police brutality against the demonstrators.

Costa *et al.* [49] analyzed tweets shared during the protests in Brazil to find the emotional dynamics of the posts. They found that the peak in the tweets coincided with days with substantial online activity. Costa *et al.* [49] used an SVM classifier on the initially collected tweets to identify protest-relevant tweets. On the protest-related tweets, a multi-nominal naive Bayes classifier was trained with 9,003 tweets manually annotated as positive, neutral, and negative emotions. The authors found the presence of both negative and positive emotions in protest tweets.

Gezi Park protest, 2013: Gezi Park protest commenced peacefully in Turkey, a country already marked by political divisions. On May 28, 2013, a group of approximately 0-100 environmental activists gathered for a sit-in at Gezi Park in Taksim Square, Istanbul. They aimed to protest against the demolition of one of the city’s last remaining public green spaces. The government’s plan to replace it with commercial establishments and upscale residences fueled the outrage. Unfortunately, the peaceful demonstration was met with harsh police response, including tear gas and water cannons, which ignited clashes between the authorities and protesters. The occupation of the park lasted until June 15, marking the end of the protest.

Varol *et al.* [202] focused on extracting topics of conversation about the social uprising and identified the trending topics. Varol *et al.* [202] also studied the spatio-temporal characteristics of the conversation, including where tweets about protests started and what locations shared the most identical trends and topics. They also reported that online content shared was highly affected by on-ground activities.

Brexit Referendum, UK, 2016: Brexit (UK EU membership referendum) occurred on June 23, 2016, in the UK and Gibraltar. The referendum aimed to determine whether the country should remain a member of the European Union or leave. In October 2015, the pro-Remain campaign group called "Britain Stronger in Europe" was formed, while two groups advocating for Brexit, namely “Leave EU" and “Vote Leave" competed to become the official Leave campaign. On April 13, 2016, the Electoral Commission declared Vote Leave the official campaign representing the Leave option. The UK government officially supported the Remain option. The voter turnout for the referendum was 71.8%, with over 30 million people casting their votes. Ultimately, the Leave campaign secured 51.9% of the votes, while the Remain campaign received 48.1%.

Grčar *et al.* [79] focused on two main questions regarding the Brexit Referendum, i.e., the mood of users on the Brexit referendum and who are the most influential users in the pro/anti-stances. The authors collected geo-tagged tweets related to the Referendum, and the results of their opinion mining from the Twitter data matched well with the opinion polls on the topic. This becomes a significant result, as it sheds light on the importance of sharing on social media, such as Twitter can be equated to people’s views on a given opinion piece. Howard and Kollanyi [92] showed that the two most important accounts in the referendum were indeed bots, i.e., $@iVoteLeave$, $@ivotestay$, whose purpose was to amplify the source simply by aggregating the content and then retweeting it. Grčar *et al.* [79] collected 4.5 million tweets from almost 1 million users about Brexit from May 12, 2016, to June 24, 2016. For determining stance, 35,000 tweets were randomly selected for manual annotation by Grčar *et al.* [79]. The study uses a score metric that considers users’ leave, remain, and neutral tweet counts to judge the user’s stance on the topic. The analysis of users who joined the leave vs. remain discourse shows that leave users gradually increased compared to remaining users who were persistently present and contributing to the debate. As for the influence, Grčar *et al.* [79] used retweets and the number of posts a user created to measure influence. The leave group was found to be considerably more active in the generation and retweeting of content, while the Remain side was found to be less active.

Day Without Immigrants & No Ban, No Wall Protest, 2017: The “Day Without Immigrants” and the “No Ban, No Wall” protests were the most recent nationwide protests in the US that aimed to show the important contributions of immigration and to resist punitive immigration policies. The “Day Without Immigrants” was held on February

16, 2017, in response to Donald Trump's plans to build a border wall, deport potentially millions of undocumented immigrants, and strip sanctuary cities of federal funding. The main aim of the protest was to show the importance of immigrants in the US economy. The "No Ban, No Wall" protest took place on January 28, 2017, in response to President Donald Trump's plan to ban citizens of certain Muslim countries from entering the US and suspend the admission of all refugees. Both protests used social media to disseminate information and aided the online protests that were going on at the time.

Wei *et al.* [212] performed a control focus group-based study to identify and reduce online prejudice towards a given part of the community. The work focuses on identifying a focal event that impacts people's behavior. Prejudices are a very mild form of hate or predefined mindset that a person has towards another community or people. The authors used the two protests as an intervention to reduce online prejudice. The results show positive and negative changes in people's prejudice after the protest. The authors also identified features of users who are more likely to change (or resist) their mindset after a protest. The findings of the work can be used to design targeted interventions during a protest-like situation.

Venezuela Political Crisis, 2019: The past decade has witnessed sociopolitical fragmentation in Venezuela due to differences in interests, identities, and politics. There are two ideologies in Venezuela, i.e., Chavism, embraced by supporters of the political ideology of the late president Hugo Chavez, and Anti-Chavism, embraced by people who strongly oppose Chavez's legacy. Chavism, however, still controls the Venezuelan political system with Nicolas Maduro as the state's head. The re-election of Nicolas Maduro as the country's president on January 10, 2019, led to the beginning of a presidential crisis driven by claims of illegitimacy and reports of coercion and fraud. The crisis continued for a while and slowly faded after March 25 when the Russian aircraft were seen arriving at the Caracas airport guarded by the Venezuelan military.

Horawalavithana *et al.* [91] focused on the content being shared on social media during the crisis as a response to external and internal factors. The authors used Venezuela's political crisis in early 2019 as a case study to gauge how the external and internal factors drive the related activities on social media. The external data for the analysis was taken from ACLED (Armed Conflict Location and Event Dataset) [156] and GDELT (Global Database of Events, Language, and Tone) database [114]. The study analyzed Twitter activity related to the political situation in Venezuela, specifically focusing on tweets either supporting or opposing President Maduro. The researchers categorized the tweets into pro-Maduro and anti-Maduro groups, with internal drivers including politicians, media outlets, and regular users. By identifying the 200 most influential users from both sides, they examined the influence within each community. Clustering analysis revealed that the anti-Maduro community's clustering coefficient decreased significantly when media accounts were removed, indicating media involvement in the anti-Maduro campaign. Conversely, the clustering coefficient decreased for the pro-Maduro community when political accounts were removed. Removing random users had minimal impact on the clustering coefficient. The study also investigated external drivers by correlating the volume of anti-Maduro and pro-Maduro Twitter activities with offline events reported in ACLED and GDELT databases. It was found that the anti-Maduro community displayed a stronger correlation with ACLED, suggesting that online discussions from the anti-side aligned with reports of protests and violent clashes documented by ACLED.

2.2 Understanding Collective Narratives

Social media protests often bring social justice and help marginalized social groups [214]. Researchers have studied protest tweets to help reduce online prejudice around certain social groups [212]. The study of anti-vaccine infodemic helped to understand human perception around the topic [75]. The analysis of textual features for understanding the sentiment of protest tweets shows the prevalence of negative sentiment [49] and specific psycho-linguistic lexicons over the others [56]. A study of tweeting activity during a protest indicates that social media activists plan the protest and share relevant tweets with a future date and time of offline protest conduct (call-to-action) to gain critical mass [139; 220]. Besides understanding the objectives, the call-to-action tweets have also helped successfully predict future protests [137]. With the help of shared grievances, people's will and hardships can be understood during protests [49]. Objectives of a protest can also be understood through the study of narratives shared during the protest. Understanding of narratives shared during the protest lies at the intersection of understanding protest participation and protest growth. Narratives are verbal, graphic, or written arguments of interconnected actors and events, developing through time [158]. Social media posts provide a fragmented narrative structure through chained social media posts. Chained social media posts create stories through events spread across multiple sources. Researchers have focused on the politician's use of social media to create a "us vs. them" narrative, leading to marginalization and polarization among the public in Turkey [108]. Identifying evolutionary trends that connect the narrative components temporally is known as Story Evolution Detection [158]. Previous work on narrative evolution has discussed the shifts in the narratives shared across social media blogs [99]. The evolution of narratives has also been studied to analyze the themes in the misinformation spread during COVID-19 [126]. The narratives shared during the protest can be loosely divided into grievance [184], call-to-action [167; 76], and reporting of on-ground activity [118].

2.3 Understanding Online Harmful Behaviour

With the rise in the use of social media use for conducting protest activity, social media started becoming the target of various radicalization groups [188], inauthentic actors [119] who started to use social media for nefarious reasons such as influence operations. Influence operations for online manipulation of users can include one or many of the strategic tools, including fake news [34], propaganda [78], hate speech [196], paid trolls [38; 120] and bots [63]. Although a vast body of work is centered around building strategic tools to identify and mitigate different perpetrators, the efficacy of proposed methods is still under scrutiny, while information operations on social media are far from being solved. More recently, researchers have started focusing on the interplay of different inauthentic activities. In this section, we focus on the harmful behavior adopted by various inauthentic and malicious actors on social media in play during the protest.

2.3.1 Accounts Identified As ISIS Groups

Spiro and Ahn [188] used the pre-identified 25,538 ISIS accounts and conducted a forecasting task to identify extremist users, estimating whether regular users would adopt their content and whether users would reciprocate contacts created by the extremists. The authors detected extremist users with 93% AUC, while the adoption of extremist content was forecasted with 80% AUC. The users were predicted to reciprocate interaction with extremist users with 72% AUC. The datasets the authors collected included 3,395,901 tweets by ISIS group accounts 9,193,267 tweets generated by users exposed to the ISIS content from the ISIS account followers data, which was taken for 25,538 random users from the set of followers. The authors curated several feature sets for their prediction purpose and implemented several machine learning models for the classification task. The models included Logistic Regression with LASSO regularization and Random Forest with k-fold cross-validation with the value of k set as 5. The authors used the greedy method to select the best features for their prediction problem. Exposure to the content of the ISIS account is determined by the Retweet mechanism on Twitter while reciprocating the user's reply to the tweet as an alibi. As for the static prediction task, the model doesn't take advantage of the timeline of the activity sequence, while a dynamic model looks into the time while making the prediction. Random Forest takes advantage of the temporal data dependency for real-time prediction. Spiro and Ahn [188] shed light on the beginning of a new era of social media, where extremist groups and content manipulators started co-existing in the digital ecosystem along with the other naive users.

2.3.2 Bots

While some accounts are purposefully created to deceive humans on social media, the automated accounts, i.e., bots, have drawn a lot of traction [197; 40]. Bots try to create content that may be polarized [119], talking highly of one side or even helping spread propaganda on social media [92]. The involvement of bots has led to discourse and tension in the online world, which are prevalent in Elections [182]. However, the bots have most recently invaded every discussion space on the social media platform [63]. The threat of automated and semi-automated accounts has been rising in social media and needs to be tackled for a safer society.

2.3.3 Russian Trolls on Twitter

By 2016, researchers warned about trolls and other forms of online manipulations. Researchers defined the trolls used in the 2016 US elections as semi-automated accounts with humans in their blackened [24]. The authors could accurately identify the Russian trolls with AUC 96% using 10-fold cross-validation. The most important features for their classification task were bot-like activity, account-level features, and political ideology. The authors collected 43 million tweets from 5.7 million users between September 16, 2016, and November 9, 2016. The dataset also contained 221 Russian trolls-produced tweets. The best algorithm for their case came as the Gradient boosting algorithm, whereas, in features, political ideology came as the most important feature in the task. The work analyzes how the users on social media are susceptible to the content they are exposed to and how easily target people can be made.

2.3.4 Hateful Users

While hate speech on social media is rising in general [129], the study of political discourse reveals that party affiliation, gender, and ethnicity are reasons for individuals resorting to posting hate speech for political leaders [186]. Online hate speech tends to pollute online discussions. Hate speech is *any content that promotes violence against the opposing stance cohort, directly or indirectly threatens the people based on their race, ethnicity, national origin, religious affiliation, political ideology, and political affiliation.* [174]. Recent studies have highlighted the growing prevalence of hate speech during protests [169]. The study of hate speech detection of low-resource languages is still nascent [130]. Early work on hate and offensive tweet detection in code-mixed language argues that translating code-mixed or low-resource language might alter the meaning and context of hate speech [130]. Catering to the multi-dimensional issues associated with hateful content detection, understanding of hate speech can help maintain the peaceful cohesion of society [55].

2.3.5 Co-ordinated Users

One of the first instances of coordination in protest participation was witnessed during the political uprising in Egypt in 2011, where participants used the "Uninstalling dictator" with progress bar tweet with different variations towards a common goal [190]. Studying individual perpetrators may overlook collective influence operations and fail to identify their inauthentic or problematic nature [95]. The study of inauthentic coordinated activity also brings challenges of the distinction of authentic activity from inauthentic activity, grassroots initiatives, or deliberate hate posting activity, and the narratives they share. Pacheco *et al.* [148] proposed to use a binary distinction for coordinated inauthentic groups through retweets and narrative duplication as a metric. More precisely, the definition of coordination adopted by Pacheco *et al.* [148] was the immediacy of systematic retweets by accounts. In their approach, Pacheco *et al.* [150] first created suspicious behavior traces from content (hashtag, n-gram, etc.), activity (timestamp, geolocation), identity (username, description), or a combination of multiple dimensions. Nizzoli *et al.* [146] proposed the definition of coordination as an exceptional similarity between a group of users and chose a network-based approach for inauthentic coordination detection. Vargas *et al.* [201] evaluated the effectiveness of the existing coordination detection approaches by building a binary classifier based on the statistical features extracted from the network for disinformation campaigns and legitimate Twitter communities. The major takeaway from the binary classifier-based approach was that the type of coordination and behavior based on it differ from campaign to campaign [201]. Sharma *et al.* [180] proposed a generative model to capture inherent coordination characteristics, leveraging Russia's Internet Research Agency dataset that targeted the 2016 U.S. Presidential Elections. Hristakieva *et al.* [95] pursued identifying coordination activity combined with propaganda detection and found that the combined analysis revealed harmful coordinated communities that were previously not noticeable. Most previous literature focused on elections to study coordination activity [180; 146; 201]. However, the study of coordination activity for protest is scarce [150].

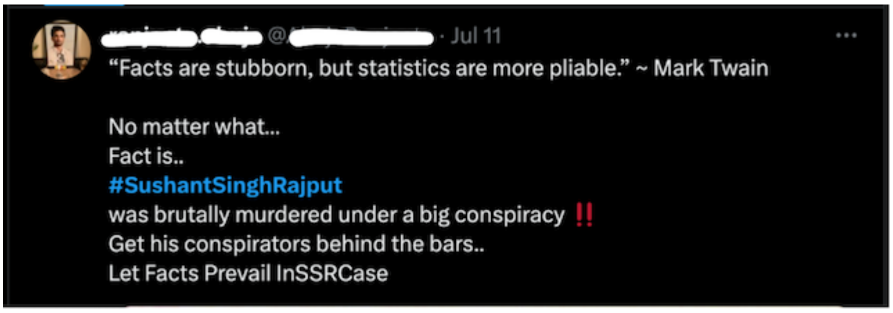
2.4 Discussion

In conclusion, previous literature on social media-mediated protests have shed light on diverse strategies employed, including recruitment dynamics [76], emotional expressions [49], sentiment analysis [65], and the interplay between online and offline activities [66]. However, the interplay of authentic and inauthentic actors during protests is under-explored [60]. This thesis bridges the gap of studying the interplay of authentic and inauthentic actors during protests. The evolution of narratives during protests, is another research gap addressed in this thesis. We also enhance the literature of comparative studies across different protests and provide valuable insights into commonalities and differences in strategies, contributing to a more robust framework for analyzing the complexities of online activism. The previous literature on harmful activities during various protests highlights how social media has become a target for radicalization groups [188], inauthentic actors [188], and influence operations with nefarious intentions[24]. The prevalence and impact of hate speech has been recognized as concerning [213], yet the understanding of hate speech detection in low-resource languages and its nuanced manifestations during protests remains an area requiring more in-depth examination [130]. This thesis bridges the gap of exploring hate speech in low-resource languages during protest. Additionally, while coordinated activity during elections has been extensively studied, there exists a scarcity of literature specifically addressing coordination during protests [180; 146; 201]. We also enhance the study on coordinated inauthentic activity during protests, with focus on different shared narratives.

CHAPTER 3

UNDERSTANDING PROTEST STRATEGIES

This chapter focuses on the strategies adopted to conduct protests on Twitter over the cause of death of an Indian actor, #SushantSinghRajput. We apply the framework of connective action on the counter-narrative protest over the cause of death of #SushantSinghRajput. We combine descriptive network, modularity, and hashtag-based topical analysis to identify three major mechanisms underlying the campaign: generative role-taking, hashtag-based narratives, and forming an alignment network toward a common cause. As online protests involve multiple users and their interactions, we focus on heterogeneous user data, including user profile information [116], the network of users involved in protest [209] as well as content shared during protest [70].



[redacted] @[redacted] · Jul 11

"Facts are stubborn, but statistics are more pliable." ~ Mark Twain

No matter what...
Fact is..
#SushantSinghRajput
was brutally murdered under a big conspiracy ‼
Get his conspirators behind the bars..
Let Facts Prevail InSSRCase

Figure 3.1: Example tweet showing the counterpublic narrative regarding the #SushantSinghRajput death row where the activist opposed the dominant narrative of suicide by the actor.

3.1 Introduction

Celebrity suicide deaths produce numerous posts on Twitter [195] and increase searches on the internet over suicide and depression-related terms [147]. Sushant Singh Rajput (SSR), a Bollywood actor and celebrity, was found dead in his Mumbai apartment on June 14, 2020 [103]. The death of the 34-year-old actor was reported as a case of suicide. However, numerous dark conspiracies triggered on social media, including debates of nepotism [193], and possibly being framed [44] or murdered [48]. A combined study of prominent news channels and politicians over the SSR controversy revealed that the commentators over the topic were rewarded with higher retweet rates, which can be attributed to the widespread discourse engagement [8]. This study focuses on the social media users' narratives that followed after the actor's death broke on the news and social media. The narrative included counterpublics [104], defined as marginalized communities that distribute messages to diverse social groups, raise awareness, and challenge dominant narratives. A Twitter user involved in activism activities such as organizing online petitions and building a counterpublic campaign narrative through hashtags is defined as a Twitter activist [207]. Figure 3.1 shows an example of a tweet with counterpublic narrative over the cause of death of #ShushantSinghRajput. This study aims to reveal the strategies adopted by Twitter activists (i.e., counterpublics) to share, spread, and mobilize the support of the counterpublic campaign about the untimely death of the Bollywood actor. Using the case study of #SushantSinghRajput, we highlight how the connective action framework can be used to identify different strategies adopted by counterpublics for the emergence of connective action.

Theory Of Connective Action: The previously defined logic of collective action answers the general question of why people get involved in collaboration with one another by explaining that people act collectively to achieve a common goal [127]. Traditionally, collective action refers to loosely connected groups of individuals, usually led by certain organizers or influential users [31]. In contrast, the logic of connective action is based on the idea of digital media functioning as organizing agents, whereas traditional organizations are either not present or are loosely responsible for providing coordination [29]. In that sense, connective action leverages the weaker ties present in social media, where users are self-motivated to post about the topic or share them. The interpersonal network hence formed can be similar to collective action sans any formal organizations. There are underlining economic and psychological logic driving the connective action, i.e., co-production and personalized sharing of expression, respectively. The two prominent indicators of a connective action are (i) a large number of participants in a movement and (ii) a very small number of users staging the connective action through the creation of content. To enrich the knowledge of how social media is deployed during social movements and how a movement is carried differently in the online world than the offline counterpart, we need to understand (i) who participates in a given movement and (ii) how people create a narrative in the social media around the protest.

Connective action comprises networked and decentralized actions of mobilization in contrast to the traditional collective action characterized by centralized resource mobilization or led by a formal organization [28]. The most crucial aspect of the emergence of connective action is the rise of self-claimed activists who co-ordinate themselves, challenge the formal organization, and conduct a campaign [31]. Counterpublics have been found to form retweet networks on social media to gain legitimacy [118] and recommend relevant messages to the supporters of the campaign [190]. Connective action holds an assumption of a decentralized network since the activists who participate in the campaign are self-motivated to participate [127]. The user retweet network can therefore be used to analyze the organizational structure of the campaign [209].

We adopt a network perspective to unpack the three major mechanisms of the connective action framework. We focus on the activists and their content posted to understand the first mechanism (i.e., *generative role-taking*) underlying the connective action. When users on social media use common hashtags, it creates a context for like-minded people. The connection of like-minded individuals thus gives rise to a networked public [216; 217; 209]. We divide the networked public into two categories, information generators, and information drivers. The information generators work on content creation, while the drivers engage in driving the discussion by retweeting the content. To inspect the second mechanism (i.e., *hashtag-based storytelling*), we perform an evolutionary analysis of hashtags used in the campaign. We divide the hashtags into buckets based on their mutually exclusive appearance in the tweets and use topic modeling on the content shared among the buckets to identify topics focused on in the different buckets. The third mechanism (i.e., *formation of alignment network*) focuses on how the activists use social media for issue alignment and achieve virality. Identifying fellow activists supporting the cause is crucial to achieving a collective goal (i.e., virality) [31]. We thus use community detection to identify sub-communities within the activist community to account for the diversity of users involved in the campaign. We also focus on how the narratives differ among sub-communities and examine any pattern within and among sub-communities. Broadly, we ask the following research questions:

- RQ1: What is the organizational structure of the social media counterpublic campaign around the death of Singh Rajput (SSR)?
- RQ2: How did hashtag-based storytelling evolve during the counterpublic cam-

paign?

- RQ3: How did the campaign activists with different perspectives achieve issue alignment on the topic?

Table 3.1: Hashtag buckets in the counterpublics campaign against the dominant narrative.

Hashtag bucket	**Hashtag variants**	**Tweet count**
#candleforssr	#candle4ssr, #candleforsushant, #candle4sushant, #candles4s	543,897
#justiceforssr	#justiceforsushantsinghrajput, #ssrkoinsaafdo (give justice to SSR), #arrestculpritsofssr	11,622
#sushantsinghrajput	#sushantsinghrajpoot, #sushantinourheartsforever, #ssrians, #sushanthsinghraj, #shushant	20,486
#bollywood / #media	#akshaykumar, #salmankhan, #kanganaranaut, #bollywoodpakisilink, #rheachakraborty, #ankitalokhande, #boycottkhans	4,064
#cbiforssr	#cbienquiryforsushantsinghrajput, #cbiinvestigationforsushant, #cbicantbedeniedforssr, #cbienquiryforssr	1,904

3.2 Data

The time duration of data collection coincided with an increase in media coverage and counterpublic narratives on Twitter. We used the Twitter search API to collect the tweets about the topic through trending hashtags which included #candle4ssr, #justice4ssr, #ssr, #sushantsingrajput. We curated a total of 1,027,213 tweets from 67,822 users using the official Twitter API. The duration of data collection spanned approximately 102 days, from July 17, 2020, to October 24, 2020. Tweets consisted of 76,781 original tweets and 950,432 retweets. Any random tweet, on average, consists of 14.9 words, giving an approximate fair window for analysis of the user's thoughts around the campaign.

3.2.1 Data Pre-Processing

Before performing any analysis on the collected tweets, we converted all the tweets into lower-case, removed stop-words, and removed any occurrence of URL from the tweets.

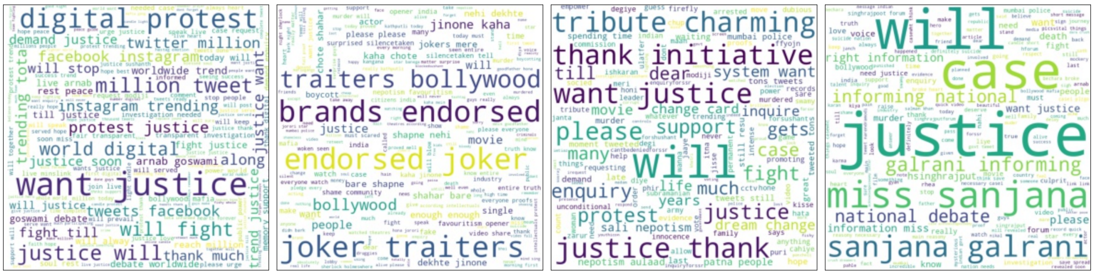
We removed any tweet with less than 3 words to keep informative tweets for further analysis. We also removed tweets with hashtags with a frequency of less than 100 in our dataset. The selection of the most frequent hashtags served to identify the narratives that became popular. Hashtags belonging to a bucket were identified based on a common theme (e.g., Bollywood and media cover hashtags with movie actors or journalists) or a different variation of the same keyword (e.g., candle4SSR written as candleforssr or candle4shushant written as candleforsushant) as shown in Table 3.1. Tweets that used hashtags from more than one bucket were excluded from the analysis due to the limitation of intention understanding that may require looking beyond hashtag usage. For example, a tweet with #JusticeForSSR is - "Sushant did not deserve this..such a beautiful and innocent soul..we demand justice..#justiceforSushanthSinghRajput". While a tweet with #CandleForSSR is - "Please do your contribution. Send your love and peace.#Candle4SSR". An example of a tweet that used hashtags from more than one bucket is - "#Candle4SSR #JusticeforSSR Subramanian Swamy leads #Candle4SSR movement demanding justice for Sushant". Since the last example might fit into either bucket, we excluded such example tweets from further analysis.

We construct a retweet network from the person who posted the message to the user who retweeted the message to capture information-sharing activities for message-motivated communication. The retweet network is directed and weighted, where the direction indicates the flow of information, and the weight indicates the number of retweets between the two users.

3.3 Methodology

We use descriptive network analysis coupled with modularity analysis and hashtag-based topical analysis to examine strategies used by Twitter users to build collective agendas and mobilize attention. We first make a user retweet network that consists of 79,170 nodes and 490,910 directed and weighted edges.

To answer RQ1, we examine the overall network structure and information flow of the tweets among counterpublics. We also identify the most active hashtag activists from the collected dataset, defined by activists' in-degree and out-degree centrality scores. While the in-degree centrality captures the level of user initiative in information



(a) #candleforssr (b) #bollywood (c) #cbiforssr (d) #justiceforssr

Figure 3.2: Word clouds for narrative hashtag bucket from Table 3.1. The word clouds belonging to each bucket show the major topics discussed in the respective bucket tweets were relevant to the bucket.

sharing, the out-degree centrality accounts for the influence and communication power of the activist.

For RQ2, we bucket the hashtags according to their mutually exclusive appearance. Social media users created numerous hashtags relating to the Bollywood actor. Selecting only the popular hashtags was to identify the narratives that went popular during the campaign. The final set of hashtags' buckets used for the study is presented in Table 3.1. We further analyze the content of the tweets from different hashtag buckets to understand the dominant narratives around the hashtags.

To examine RQ3, we apply community detection on the retweet network to discuss how the counterpublic campaign narratives differ among the sub-communities. For community detection, we use CNM (Clauset-Newman-Moore) greedy modularity maximization algorithm [43]. CNM is a bottom-up agglomerative clustering algorithm that maximizes the modularity [143] of the community structure in a greedy manner. Once we have identified the sub-communities, we examine how the topics presented by the sub-communities differ for detecting alignment in the sub-communities.

3.4 Analysis

In this section, we first perform a network descriptive analysis of the retweet network formed during the counterpublic campaign. Next, we study the evolution of the counterpublic campaign narratives. Finally, we discuss how the counterpublics reached issue alignment during the campaign.

3.4.1 Network Descriptive Analysis

Descriptive network analysis of a network can help identify the user dynamics and their clustering patterns during the online campaign. We present the descriptive analysis of the retweet network of the counterpublic campaign in Table 3.2. The retweet network was found to be very sparse, with a network density of 0.000078. The sparseness in the network is expected, given the large number of nodes and edges in the network. Usually, the retweet network tends to cluster rather than be evenly distributed, which can indicate the formation of an echo chamber around a topic [181]. The average in-degree and out-degree centrality for the activists were 7.83, which indicates that the average connection between activists for either retweeting or being retweeted is equal. The average clustering coefficient for the network is 0.060, which is very low. The low clustering coefficient indicates that all the activists are not well connected. Based on the out-degree centrality, a single user's highest number of retweets is 23,210. While based on the in-degree centrality, the user who retweeted the maximum number of times had count of 1,253.

The in-degree centralization of the network is 0.0065, while the out-degree centralization is 0.29. A higher out-degree centralization indicates a set of users who were more frequently retweeted than others. Comparatively, a lower network in-degree suggests that the activists were more or less equally active while retweeting about the campaign. This result indicates the evidence towards slacktivism, defined as actions requiring minimum effort and participation cost, like retweeting since it does not require the user to create their content [36]. Since the counterpublics were mostly slactivists, the campaign's main goal was to obtain momentum and raise awareness about the campaign. To answer RQ1, we divide the activists involved in the counterpublic campaign into two parts based on their in-degree and out-degree centrality measures.

Table 3.2: Descriptive statistics of the overall retweet network for SSR counterpublics campaign.

Metric	**Mean value**
Network Density	0.00078
In-degree Centrality	7.83
Out-degree Centrality	7.83
Clustering Co-efficient	0.060
In-degree Centralization [67]	0.0065
Out-degree Centralization	0.29

Table 3.3: Network descriptive statistics for the top information drivers and generators to understand the organizational structure of the counterpublic campaign. $^{*}p < 0.05, ^{**}p < 0.01, ^{***}p < 0.001$ analyzed using unpaired Mann–Whitney U test. SD stands for Standard Deviation. We find significant differences across all the metrics between the top Information Generators and top Information Drivers.

	Top Information Generator		Top Information Driver		
Metric	Mean	SD	Mean	SD	p
Active Days	7.65	20.19	12.05	24.94	$***$
Number of Followers	8024.8	107137.7	122.084	351.87	$***$
Number of Followees	479.54	3278.9	136.861	336.64	$***$
Number of Tweets	8225.29	22076.6	9204.433	14673.42	$***$
Indegree Centrality	8.37	0.0002	0.0013	0.0052	$*$
Outdegree Centrality	8.37	0.0018	0.0013	0.0042	$*$
Betweenness Centrality	4.86	1.50	1.29	0.00013	$***$
Closeness Centrality	0.003	0.0012	0.015	0.016	$***$
Eigenvector Centrality	0.0012	0.0035	0.0024	0.0097	$*$

We select the top 1000 activists in our dataset based on their in-degree and out-degree centrality. The top 1000 users with high out-degree centrality are referred to as top information generators, and the top 1000 users with the highest in-degree centrality are referred to as top information drivers. We analyze the descriptive network statistics for the top information drivers and generators to understand the organizational structure of the counterpublic campaign. The descriptive network statistics for the top generators and drivers are listed in Table 3.3.

Based on the descriptive statistic analysis summary of the activist's attributes, a typical information generator was active for 7.65 days, had about 8,024 followers, followed 479 users, and tweeted 8,225 times. While on the other hand, a typical information driver was active for 12 days and had a comparable follower-to-followee ratio. Mann–Whitney U tests were performed to examine whether the difference between information generators and information drivers is significant or not. We perform Mann–Whitney U tests since the test does not make any inherent assumption about the population distribution. We found that there is a significant difference between the active days, the number of followers, and the number of followers, as shown in Table 3.3.

Although the average number of days a user participated in the campaign is low for drivers and information generators, we found that the drivers were more active than the generators. From the eigenvector centrality score, we can conclude that since the information driver's score is more than the generator's score, drivers are more actively connected with other active campaign activists. However, the betweenness centrality for a generator is more than the driver, indicating generators are more likely to have a shorter path between two activists. The active retweeting of the campaign hashtags and a mix of centralized information aggregation and decentralized information generation are key to developing connective action.

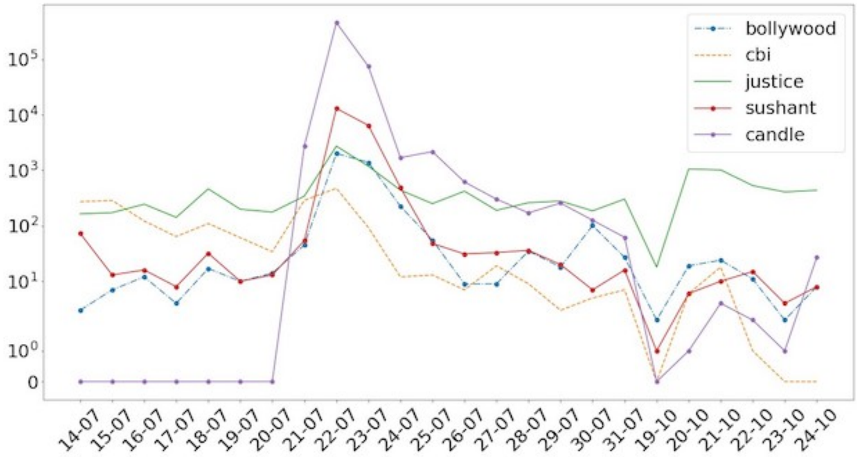


Figure 3.3: Evolution of counterpublic campaign over the period of three months with respect to hashtag buckets as presented in Table 3.1.

3.4.2 Evolution Of Counterpublic Campaign Narratives

To analyze how the counterpublic campaign evolved over the period, we plot the frequency of narratives' buckets identified through hashtags in Figure 3.3. The division of hashtags is presented in Table 3.1. We found that all the hashtags generally saw a spike between July 20, 2020, and July 24, 2020. The tweets with hashtags #cbiforssr and #justiceforssr were initially used more; however, during the period of highest frequency, #candleforssr was used most times. The use of #Bollywood hashtags also rises during the spike. #justiceforssr, however, was the most consistent hashtag bucket throughout the data collection.

To understand what narrative was spread in tweets within the hashtag buckets and how they differ, we plot the word cloud of the tweets from hashtag buckets as shown in Figure 3.2. The dominant narrative from #candleforssr was the declaration of online protest against the debate on the suicide of the late actor. The #candleforssr bucket revolves around demanding justice, mobilization through participation, and mention of debate and journalists (e.g., Arnab Goswami). The #justiceforssr bucket showed some narratives similar to #candleforssr, in addition to mentioning influential people, murder conspiracy, and shades at Mumbai police as shown in Figure 3.2(d). The #bollywood bucket in Figure 3.2(b), mainly included tweets mocking other Bollywood celebrities and despising nepotism. #cbiforssr, which was one of the first spikes in the dataset, consisted of tweets about inquiry, involvement of CBI (Central Bureau of Investigation), and topics of justice, protest, and nepotism as shown in Figure 3.2(c).

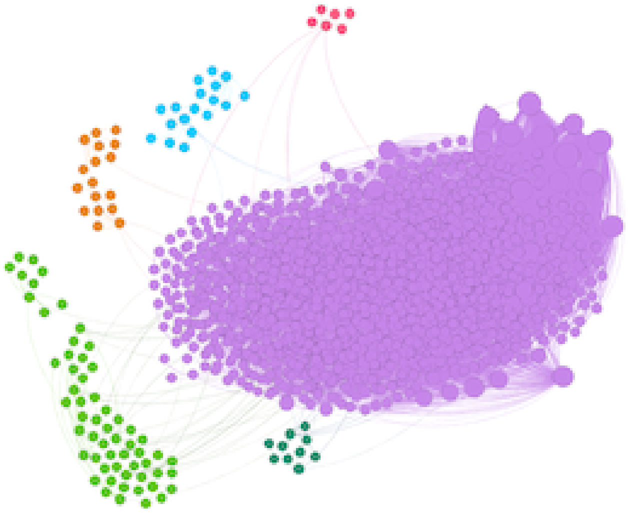


Figure 3.4: Figure showing the community formed among top information generators and their top drivers. Each color uniquely identifies a sub-community. Sub-community 1, shown in purple, constitutes 92.96% of the users. The second sub-community, shown in green, constitutes 4.15% of the users. While the blue sub-community includes 1.27%, orange comprises 1.2%, dark green comprises 0.7%, and pink sub-community comprises 0.42% of the users, respectively.

3.4.3 Issue Alignment Among Counterpublics

We used the top 1000 generators and their top 10 drivers to identify whether there is the formation of any sub-community within the network and whether different sub-communities share different narratives. The reason for selecting the top generators is to account for the most popular content in the campaign. We apply the CNM algorithm for community detection [43] among the counterpublics. The number of iterations for the community detection algorithm was 100. The average clustering coefficient was 0.021, with an average degree of 14.075, modularity of 0.35, and network diameter of 9. We found 6 sub-communities in our user-retweet network as shown in Figure 3.4 with each community represented by a different color. The retweet network of top generators is densely connected, which shows evidence of a connective campaign and a leaderless information-sharing framework. A few nodes with less connection indicate a centralized structure where information is shared from a few generators to many drivers. The formation of the dense cluster is evidence of connective action. We further perform LDA [33] on the combined tweets of top 1000 generators and top 1000 drivers to identify the major topics they share online.

Table 3.4: Table with topics discussed among top 1000 information generators and drivers respectively.

Justice	singh, world, justice, protest, digital
Candle	supporting, hope, smile, many, stand
Support$_T$	tweets, guys, digital, protest, million
Support$_C$	comment, below, million, reach, post
Media	arnab, goswami, debate, worldwide, live
Support	dead, watching, where, living, duty, suicide

Among the top 6 topics from the LDA as shown in Table 3.4, 3 dominant topics revolved around online mobilization represented as Support$_T$, Support$_C$, and Support. In the 3 mobilization topics, the social media users requested SSR fans and fellow social media users to retweet the content for widespread dissemination of information. While Support$_T$ is encouraged to tweet about the campaign, Support$_C$ suggests commenting on the posts to gain momentum on social media. On the topic of Support, the counterpublics used words like duty and watching to encourage fellow campaigners and social media users to participate. The other 3 topics were identical to #justiceforssr and #candleforssr buckets, which were the two most prominent narratives in the overall

Table 3.5: Table with topics discussed among sub-communities.

Protest	protest, want, world, justice, digital, love, tweets
Media	arnab, know, rhea, raha, pagal (mad), aadmi (man), badla (revenge), will
Nepotism	money, huge, production, extract, houses, handle
Candle	light, candle, support, thank, fight, unity, hope, march

campaign. The topic represented as Media included the debate led by news media on the investigation of the suicide.

To answer RQ3, we first run the LDA on the tweets from each sub-communities. Given that the people who were retweeting each other would belong to the same sub-community based on modularity analysis, the same set of tweets is expected from a given sub-community to remain connected. We set the number of topics as 3 with 10 words in each topic. To find the alignment among users from the 6 sub-communities, we identify the common topics in all the sub-communities. Table 3.5 shows the topics in the tweets/retweets of users in different sub-communities were similar, indicating an inter-connected community structure and issue alignment in sub-communities.

3.5 Discussion

This study sheds light on how hashtag activism can evolve into connective action by examining the mechanisms of generative role-taking, hashtag-based storytelling, and issue alignment among diverse activist groups. Applying the connective action framework to analyze the counterpublic campaign surrounding the untimely death of Sushant Singh Rajput (SSR) on online social media provides valuable insights into protest strategies.

Understanding generative role-taking through constructing a user retweet network reveals the importance of influential information generators with a shorter path to reach fellow activists. Additionally, it highlights the active connections maintained by top drivers. This knowledge helps comprehend the dynamics of information dissemination within the campaign and the roles played by key participants.

Identifying the most consistent hashtag, such as #justiceforssr, and the peak usage of #candle4ssr provides valuable insights into the effective mobilization of the counterpublic campaign. Hashtags serve as rallying points, allowing activists to coordinate their efforts, express solidarity, and amplify their message. Analyzing these hashtag patterns informs our understanding of effective communication strategies in online protests.

Moreover, the community detection analysis conducted on the retweet network reveals a combination of centralized and decentralized information aggregation, with densely connected top generators and sparse connections between individuals. Recognizing the patterns of connectivity helps in comprehending the structure and organization of the counterpublic campaign and the interplay between different activist groups.

Our research contributes to understanding protest strategies by demonstrating how the connective action framework can be applied to study online social media campaigns. This chapter provides insights into the dynamics of information diffusion, the role of hashtags in storytelling and mobilization, and the formation of communities within the activist network. Our findings can help future protests and movements toward the development of more effective strategies for online activism. The pattern of connectivity analysis discussed can also help future researchers assess the organization structure of the future protests.

Limitations and Future work: While studying the evolution of hashtag-based storytelling, hashtags were bucketed based on a common theme (e.g., Bollywood and media cover hashtags with movie actors or journalists), including variations of the theme-identified keyword (e.g., candle4SSR written as candleforssr or candle4shushant written as candleforsushant). This approach facilitated a focused analysis of distinct themes, providing deeper insights into the underlying strategies. However, a limitation of the study arises from the decision to exclude tweets incorporating hashtags from multiple thematic buckets. While aiding in clear strategy distinction, this deliberate selection criterion restricts the exploration of more intricate and interconnected themes in the protest. The investigation of tweets featuring a blend of hashtags from different thematic buckets is deferred for future research, presenting a route for a more comprehensive understanding of the multifaceted nature of hashtag usage in protest narratives.

CHAPTER 4

UNDERSTANDING NARRATIVE SHARED DURING PROTEST

In the previous chapter, we delved into the strategies employed by users engaging in online protests. While we have utilized hashtag-based labeling and topic modeling thus far to comprehend the protest strategies, we hypothesize that the shared content during protests may exhibit common narratives and structural similarities. Hence, we aim to investigate shared narratives, analyze their evolution, and explore the surrounding communication patterns around the identified narratives during protests. To achieve this, we put forth an unsupervised clustering-based framework that enables us to comprehend the various narratives in an online protest. Through this analysis, we offer novel insights into the narratives that emerge during online protests by studying four specific protests. Our findings indicate that clusters encompassing call-to-action tweets and on-ground activity reporting tweets represent common narratives observed across all protests. Analyzing the narrative evolution reveals that the call-to-action narrative remains consistent throughout the protests under study. Moreover, through community detection analysis conducted on the retweet network across protests, we observe the formation of narrative-centric communities.

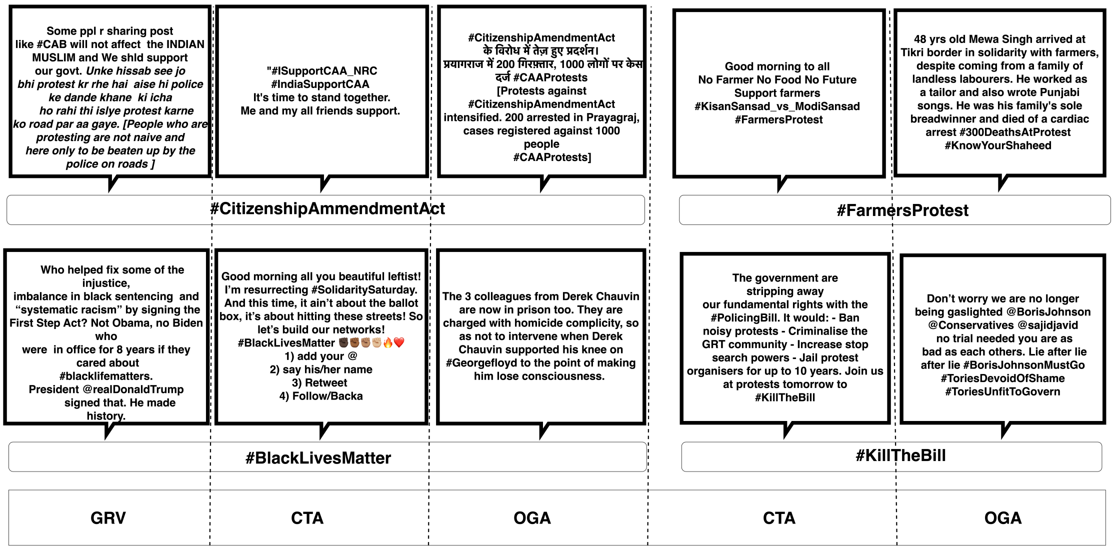
4.1 Introduction

Social media has become integral to various social movements and protests due to easy information dissemination and wider public reach [110; 56; 208; 118]. Across different socio-economic circumstances, the online protests share similar morphological features in using social media for self-organization and obtaining a more significant number of participants [76]. Narratives are verbal, graphic, or written interpretations of related events and participating actors, evolving through a given duration [158]. Using a hashtag to build a collective narrative makes Twitter one of the prime spots for conducting protest [208]. While Twitter enables a broad reach of the protest, a fine-grained analysis of various narratives present within a protest setting may help decipher the people's perception and shed light on the protest's overall focus. Understanding different narratives present in the platform is essential; likewise, it is also critical to differentiate and assess the impact produced by various participants during an online protest. Since highly influential users may be responsible for spreading information to a broader audience, the different narratives of influential users might play a crucial role in shaping the protest [208].

Previous studies on social media movements and protests have focused on different collective narratives in the campaign [142; 208]. Some narratives include information dissemination (such as personal grievances) around the topic[184], call for participation [167] or reporting of on-ground activities [203]. The grievance narrative might include personal stories of perceived injustice or other forms of hardships related to the cause. On-ground activities are narratives from people witnessing the offline protest or posts about current online activities related to the protest. The call for participation (call-to-action) narrative urges the users to participate in the cause by being part of the physical protest or using social media to tweet protest-related posts. Although the different narratives during a protest have been studied individually, a unified discussion of various narratives present within a protest is scarce [208]. To broaden the understanding of social media protests, it also becomes inevitable to understand the communication of the different participants [212]. Hence, the origin and development of various narratives during the protest require the understanding of fellow participants.

Citizenship Amendment Act, 2019 (CAA): Indian Government passed the Citizenship Amendment Act, 2019 on December 11, 2019. It allows illegal immigrants facing religious persecution in Afghanistan, Bangladesh, or Pakistan to seek citizenship in India if they have entered India on or before December 31, 2014 [39]. This led to a protest in the country with a debate on the non-secular roots of the Act. The protests were rooted in excluding other religious minority communities like Rohingya Muslims, Jews, Bahais, and Zoroastrians from seeking citizenship.

Black Lives Matter, 2020 (BLM): In a tragic turn of events, Ahmaud Arbery, a 25-year-old Black man, was shot and killed while jogging in Georgia on February 23,



Some ppl r sharing post like #CAB will not affect the INDIAN MUSLIM and We shld support our govt. *Unke hissab see jo bhi protest kr rhe hai aise hi police ke dande khane ki icha ho rahi thi islye protest karne ko road par aa gaye. [People who are protesting are not naive and here only to be beaten up by the police on roads]*

"#ISupportCAA_NRC #IndiaSupportCAA It's time to stand together. Me and my all friends support.

#CitizenshipAmendmentAct के विरोध में तेज़ हुए प्रदर्शन। प्रयागराज में 200 गिरफ़्तार, 1000 लोगों पर केस दर्ज #CAAProtests [Protests against #CitizenshipAmendmentAct intensified. 200 arrested in Prayagraj, cases registered against 1000 people #CAAProtests]

Good morning to all No Farmer No Food No Future Support farmers #KisanSansad_vs_ModiSansad #FarmersProtest

48 yrs old Mewa Singh arrived at Tikri border in solidarity with farmers, despite coming from a family of landless labourers. He worked as a tailor and also wrote Punjabi songs. He was his family's sole breadwinner and died of a cardiac arrest #300DeathsAtProtest #KnowYourShaheed

Who helped fix some of the injustice, imbalance in black sentencing and “systematic racism” by signing the First Step Act? Not Obama, no Biden who were in office for 8 years if they cared about #blacklifematters. President @realDonaldTrump signed that. He made history.

Good morning all you beautiful leftist! I’m resurrecting #SolidaritySaturday. And this time, it ain’t about the ballot box, it’s about hitting these streets! So let’s build our networks! #BlackLivesMatter ✊🏿✊🏾✊🏽✊🏻🔥❤️
1) add your @
2) say his/her name
3) Retweet
4) Follow/Backa

The 3 colleagues from Derek Chauvin are now in prison too. They are charged with homicide complicity, so as not to intervene when Derek Chauvin supported his knee on #Georgefloyd to the point of making him lose consciousness.

The government are stripping away our fundamental rights with the #PolicingBill. It would: - Ban noisy protests - Criminalise the GRT community - Increase stop search powers - Jail protest organisers for up to 10 years. Join us at protests tomorrow to #KillTheBill

Don’t worry we are no longer being gaslighted @BorisJohnson @Conservatives @sajidjavid no trial needed you are as bad as each others. Lie after lie after lie #BorisJohnsonMustGo #ToriesDevoidOfShame #ToriesUnfitToGovern

Figure 4.1: Figure showing examples of different narratives expressed by people during online protests. CTA: Call-to-action, OGA: On-ground activities, GRV: personal grievances.

2020, by three white men. [1] Another event happened in Kentucky, where the police fatally shot 26-year-old Breonna Taylor during a no-knock apartment search on March $13th$, 2020. [2] In another event, George Floyd, who was 46 years old, was arrested and killed on May $25th$, 2020, after being knelt on the neck by white police officer Derek Chauvin for over nine minutes. [3] The sequence of events led to protests in multiple cities in the US and, subsequently, in the rest of the world. #BlackLivesMatter became the hashtag for the protest that broke out in the different parts of the world [214].

Farmer’s Protest, 2020 (FP): The Indian government proposed the Farmer’s Bill on September 20, 2020, which stirred the country. The country’s farmers feared that the three laws introduced in the bill would abolish the Minimum Support Price (MSP), leaving farmers at the mercy of big corporations. Protests broke out online and offline due to the proposed bill, with people demanding it is repealed. The turn of events in the country led the Indian government to finally repeal the Act on November 09, 2021, ending the year-long protest in the country [57].

Kill The Bill Protest, 2022 (KTB): Police, Crime, Sentencing, and Courts Bill (PCSC) introduced new police powers and reviewed the present rules around crime and protests in England and Wales. The activists opposed the law due to its ability to impose conditions on any protest deemed disruptive to the local community, leading to up to

[1] https://www.nytimes.com/article/ahmaud-arbery-shooting-georgia.html
[2] https://www.nytimes.com/article/breonna-taylor-police.html
[3] https://www.nytimes.com/2020/05/31/us/george-floyd-investigation.html

10 years of jail. The punishable conditions included disrupting public properties and statues and restricting access in and out of parliament [58].

We build the framework on our previous work on narrative detection during online protests. We propose an unsupervised cluster-based approach to identify the different narratives of the protest. The primary motivation for using the cluster-based method is to leverage the semantic difference between clustered tweets and identify fine-grained separation between clusters as different narratives in a protest. The framework was used to perform a comparative study of narratives in 4 protests under study to examine dominant and converging narratives across protests. We extend the narrative detection framework to study the evolution of the converging narratives and include user-level analysis of the narratives shared during the protest. The study of the evolution of different narratives can help us understand how protests are built on the online platform. To understand users' perspectives to the narrative detection during the protest, we investigate how the top influential users engage in disseminating different narratives since influential users play a crucial role in shaping the protest. Broadly, we ask the following research questions:

RQ 1: What are the different narratives present in a protest?

RQ 2: How do different narratives evolve during the protest?

RQ 3: How do influential users contribute to different narratives in the online protest?

4.2 Data And Methodology

In this section, we discuss the dataset collected for the analysis and detail the method used for discovering the narrative clusters and user engagement in the protest.

4.2.1 Data

As a starting point for data collection, we used the trending hashtags during the protest to collect tweets around the protest. We used the Twitter API [4] to collect tweets and incrementally include the relevant hashtags through snowball sampling. Table 4.1 shows

[4] https://developer.twitter.com/en/docs/twitter-api/v1/tweets/curate-a-collection/overview

Table 4.1: Hashtags used for data collection for the 4 protests under study.

	Hashtags
CAA	#cabbill, #citizenshipamendmentact, #cab2019, #CABPolitics', #cabprotest, #caaprotest, #caa_nrc_protest, #isupportcab2019, #indiarejectscab, #indiarejectscaa, #hindusagainstcab, #hinduagainstcaa, #scstobc_against_cab, #hindussupportcab, #indiasupportscab, #isupportcaa_nrc, #muslimswithnrc, #caa_nrc_support', #caasupport, #caa_nrcprotests, #isupportcaa, #protestsagainstcaa, #protestagainstcab, #indiadoessupportcaa, #indiadoesnotsupportcaa, #indiarejectscaa, #indiarejectscaa_nrc_npr, #RejectCAA, #RejectNRC, #indiasupportscaa_nrc_npr, #indiasupportscaa, #support_caa_and_nrc, #wesupportcaa
BLM	#BlackLivesMatter, #NormalizeEquality, #blacklivesmatter, #GeorgeFloyd, #ICantBreathe, #BlackLivesMatter, #NoRacism, #BLACKLIFEMATTERS, #blacklifematters, #BlackLivesMatterUK
FP	#StandWithFarmers, #StopPrivatization_SaveGovtJob, #300DeathsAtProtest, #FarmersProtest, #MyFarmer_MyPride, #neverforget1984, #FarmersRoarInBengal, #FarmersLivesMatter, #NoFarmerNoFood, #kisanektazindabaad, #kisanmajdooriktazindabad, #kisanmorchazindabad, #BoycottBJP_4Farmers, #SaveAfghanSikhs, #KisanKaSammanPMKisan, #PMKisan, #PMKisanSammanNidhi, #KisanWithPMModi, #ModiWithFarmers, #FarmersWithModi, #RealFarmersWithModi, #isupportfarmerbills, #HonestFarmersWithPMModi, #ISupportFarmBills, #wesupportfarmerbill, #IsupportFarmReforms
KTB	#KillTheBill, #ToriesDevoidOfShame, #ToriesOutNow, #PolicingBill, #PCSCBill, #TorySleaze, #RightToProtest, #protestisnotacrime, #Wewontbesilenced, #policingbill, #ISupporttheBill, #PoliceCrackdownBill, #ProtestBill, #BristolProtest

the hashtags used for collecting tweets from the 4 protests under study.

Table 4.2: Statistics of the collected data used to analyze campaign narratives.

Protest	Start date	End date	#Tweets	#Users
CAA	Dec 07, 2019	Feb 27, 2020	11,350,276	931,175
BLM	June 03, 2020	June 29, 2020	7,183,280	3,692,495
FP	Mar 14, 2021	Aug 18, 2021	1,509,703	160,286
KTB	Jan 14, 2022	Jan 26, 2022	280,549	73,666

The collection for CAA and KTB coincides with the initial months of the protest. The collection of FP and BLM tweets was done at a later stage. The reason for the timestamps for data collection is incidental. Statistics of the collected data are present in Table 4.2. Since the initial dataset collected from Twitter may contain some noise, we perform pre-processing on the tweets to sustain tweets with rich information about the protests.

Data Pre-Processing

After we collect the initial tweets, as shown in Table 4.2, we follow the following pre-processing steps to ensure good quality of data: a) remove the mentions, URLs, and emojis. b) case-folding, where we lowercase the tweets. c) remove tweets with less than 10 words. d) split the hashtags at the capital letters.

As shown in Table 4.2, the initial data collected for CAA was $11,350,276$ from $931,175$ users. Among the total tweets, the original tweets were $1,543,805$, while Retweets / Quoted tweets were $9,806,471$. After the first pre-processing step, the tweet count was $11,302,023$. The initial data collected for FP was $1,509,703$ tweets, with $199,626$ original tweets and $1,310,077$ retweets. The first step of pre-processing reduced the FP tweet counts to $1,500,022$. The initial tweets collected for the study of KTB were $280,549$, produced by $73,666$ users. The data constituted $40,798$ original tweets and $239,751$ retweets / quoted tweets. After the pre-processing step, the total number of tweets for KTB was $278,065$. For BLM, the initially collected tweets were $7,183,280$ from $2,531,929$ users. The initial pre-processing of the tweets reduced the BLM tweets to $5,670,789$. For the selected tweets obtained after the pre-processing, we use the original tweets with mentions, URLs, and emojis for further analysis.

4.2.2 Method

Narratives in online campaigns are continuously evolving and tend to be topic driven [142; 159]. Therefore, a fixed set of labels may not always fit a given protest. Hence, we propose an unsupervised framework for identifying significant narratives of a protest, as shown in Figure 4.2. Our approach's clustering of tweets for narrative detection drives inspiration from an unsupervised user-based stance detection framework that starts with conversations on a given topic [159]. Instead of user-based detection, we use a tweet-based approach to identify the prominent narratives in a social media protest. Our complete framework is divided into four steps. In the first step, we filter active tweets for narrative detection. In the second step, we cluster the tweets from Step 1 to obtain the narratives present in a protest. In the third step, we map the tweets used for narrative detection to the original tweets in the protest to obtain narrative labels for all protest tweets. In the fourth step, we first analyze the different narrative compositions of a

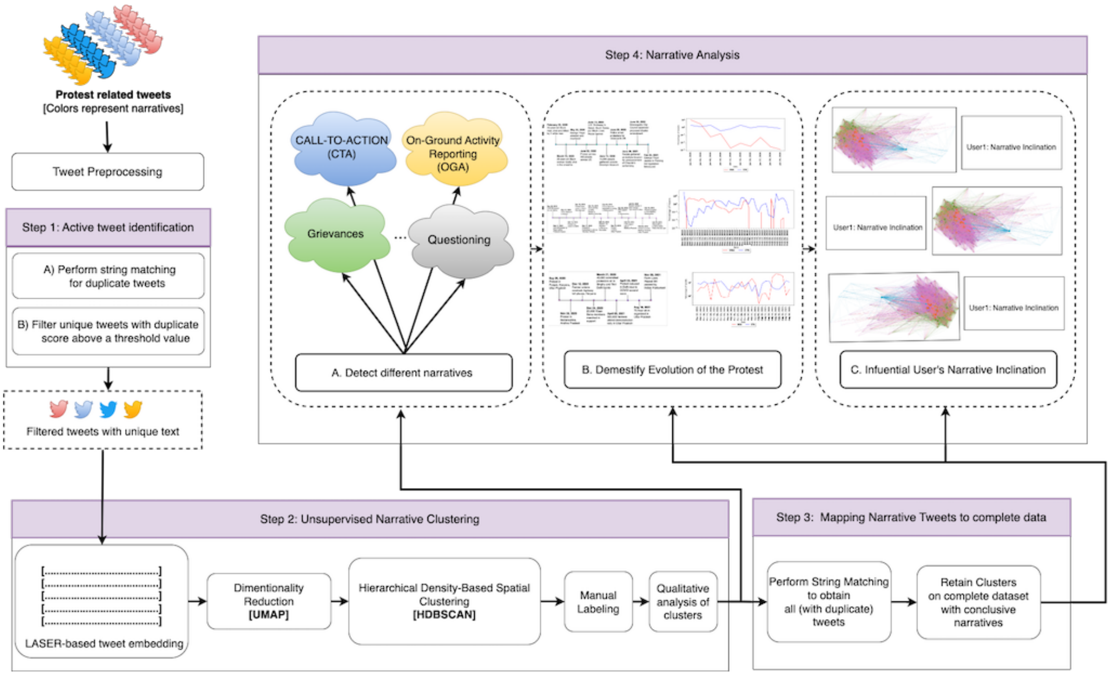

Protest related tweets
[Colors represent narratives]
Tweet Preprocessing
Step 1: Active tweet identification
A) Perform string matching for duplicate tweets
B) Filter unique tweets with duplicate score above a threshold value
Filtered tweets with unique text
Step 4: Narrative Analysis
CALL-TO-ACTION (CTA)
On-Ground Activity Reporting (OGA)
Grievances
...
Questioning
A. Detect different narratives
B. Demestify Evolution of the Protest
User1: Narrative Inclination
User1: Narrative Inclination
User1: Narrative Inclination
C. Infuential User's Narrative Inclination
Step 2: Unsupervised Narrative Clustering
[.....................................]
[.....................................]
[.....................................]
[.....................................]
[.....................................]
LASER-based tweet embedding
Dimentionality Reduction [UMAP]
Hierarchical Density-Based Spatial Clustering [HDBSCAN]
Manual Labeling
Qualitative analysis of clusters
Step 3: Mapping Narrative Tweets to complete data
Perform String Matching to obtain all (with duplicate) tweets
Retain Clusters on complete dataset with conclusive narratives

Figure 4.2: Proposed framework to demystify protest composition. The framework is divided into 4 parts. First, we identify the active tweets for the protest. Second, we utilize a clustering-based method for unsupervised narrative detection. Next, we map the narrative tweets to complete data. Finally, we perform the narrative analysis on the complete data.

protest obtained from Step 2. Here we discuss the dominant narrative in each protest and identify the converging narrative across protests. Secondly, we analyze the evolution of the converging narratives for all protests. In the third and final part, we analyze different influential users' roles in sharing the detected narratives during the protest.

Active Tweet Identification

We use a two-step process to consider a rich and unique instance of tweets around a protest. First, we use string matching to identify duplicate tweets in an online protest. We remove the hashtags and mentions to conduct string matching on the tweet text. While using most retweeted tweets is one of the approaches to identify duplicates, we also wanted to include the same text tweeted by two or more users. The practice of tweeting the same text instead of retweeting has recently gained much traction in the global south recently [106]. Secondly, we use tweets whose occurrence (duplicates) exceeds a particular threshold. The threshold adopted is based on the data size and manual intervention, where we recheck the cluster outputs with different threshold values. We choose the threshold for semantic similarity as 30 to identify the most active tweets for CAA. With 30 as the threshold, active tweets identified were $36,109$ unique tweets, which we mapped back to $7,775,268$ tweets/retweets in our dataset. For BLM, the threshold for semantic similarity was selected as 30, which led to $13,151$ unique active tweets that mapped back to $4,338,427$ tweets/retweets in the data set. For FP, we chose the threshold as 30 and obtained $7,553$ unique tweets that constituted $772,840$ total tweet/retweet in the FP dataset. The threshold for KTB was set to 5, with the total tweets considered for analysis after threshold selection being $200,946$ from $3,821$ active tweets.

Unsupervised Narrative Clustering

Once we have identified the active tweets, we first represent the tweets in the embedding space using BiLSTM encoder-based universal language agnostic sentence embedding (LASER) [20], which has proven to give the best performance for retaining linguistic information among various sentence embeddings [111]. The other motivation for using LASER is that it offers a benefit over limited resource language and code-switching texts. As the two protests under study, CAA and FP, are from India, given India's rich linguistic diversity, building models that cover as many languages as possible for a protest/campaign study becomes essential. LASER uses Moses tool [5] for pre-processing a sentence. After pre-processing, the sentence representation is 1024 dimensional.

After obtaining a 1024-dimensional representation for each tweet, we project each tweet onto a two-dimensional plane using Uniform Manifold Approximation and Projection (UMAP) algorithm [133]. UMAP attempts to project similar elements closer to each other while dissimilar elements are projected far away. The performance of UMAP has shown better projection than other techniques, including t-distributed Stochastic Neighbor Embedding (t-SNE) [199].

In the last step of unsupervised narrative detection, we cluster the two-dimensional tweet vector using hierarchical density-based clustering (HDBSCAN) [132]. HDBSCAN finds clusters of varying densities. We also tried using other clustering algorithms, including Meanshift [45] and DBSCAN [176]. However, HDBSCAN gave us the best clustering performance, determined by manual evaluation of the clusters. We manually annotate randomly selected two sets of 10 sample tweets from each cluster

[5] http://www2.statmt.org/moses/

to analyze narratives in the clusters. For all annotated samples, we calculated Cohen's Kappa [21] and found an inter-annotation agreement, and proceeded with clusters with a strong agreement between annotators (0.95).

For qualitative analysis of the labeled narrative clusters, we compute the semantic difference between different narrative clusters [159; 53]. To evaluate the semantic difference between clusters, we used a prominence score that identifies the most prominent term in each cluster to show how each narrative uniquely talks about the same issue in a different context. To suit our need, we generalize the prominence score used in the literature [159] for more than two cases. The *prominence score* uses valence score and term frequencies to distinguish cluster narratives. For each term *t*, we capture the degree of its occurrence in a set of tweets from cluster i, i.e., tf_i, as compared to all other clusters tf_j (where j ranges from $Cluster_1$ to $Cluster_n$). The *prominence score* of a term *t* is defined as a product of its valence score and its term frequency as follows:

$$V(t,i) = \log(tf_{t,i}) * (2 * \frac{\frac{tf_i}{total_i}}{\sum_{j=1}^{n} \frac{tf_j}{total_j}} - 1) \tag{4.1}$$

Using the Prominence score *Pr*, we compute each narrative cluster's top terms, emojis, hashtags, and mentions.

Mapping Narrative Cluster Tweets To Complete Data

After we have clustered and manually labeled the prime narratives clusters formed during a protest, we perform a mapping operation of narrative-identified active tweets to complete the dataset for further analysis. The mapping back of unique tweets to original tweets is a two-step process. In the first step, we perform the similarity matching of the vector of tweet embedding in Step 2 with instances of the clustered unique tweets. In the second step, we retained only the tweets for which we could obtain conclusive narratives and discarded the rest of the data from any further analysis.

Narrative Analysis

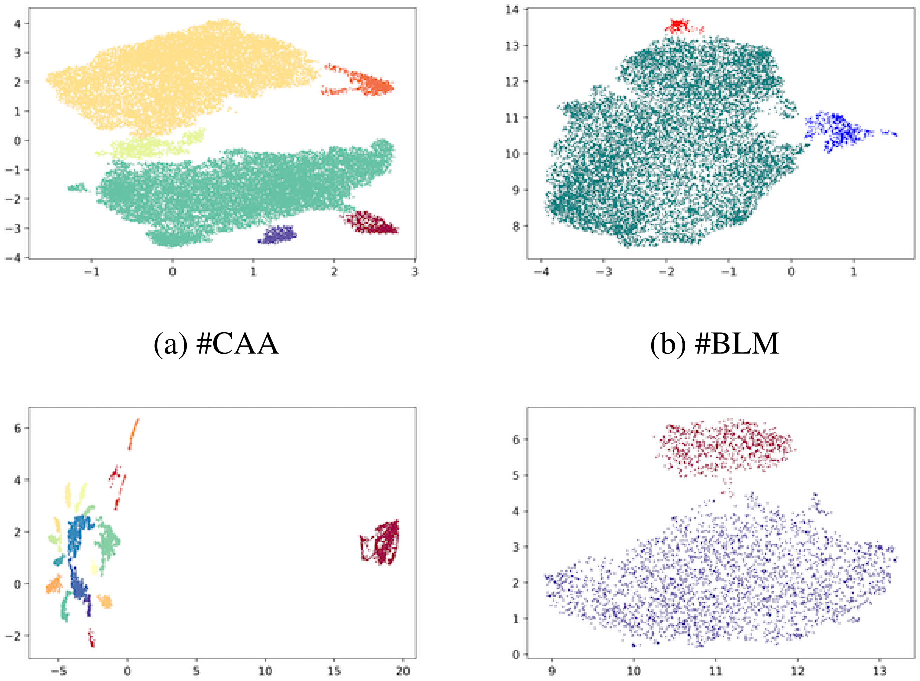
In the final step of the proposed framework, we analyze the results obtained from Step 2 and Step 3. We first analyze the narratives in each protest individually and identify the dominant narrative in each protest. Next, we analyze the converging narratives present in all protests. After identifying converging narratives, we analyze the evolution of the converging narratives in all the protests under study. In the last part of the analysis, we undertake a user-centric view of disseminating different narrative tweets during the protest. We analyze the user-retweet network structure of each protest for the Influential users defined as the top 5% users based on out-degree (represents the retweets obtained by the user) in each protest. We further analyze the role of influential users in disseminating different narratives and whether the Influential users drive any specific narrative during the protest.

Table 4.3: Main narratives present in the protests under study. Abbr; QUEST: Questioning tweets, SKEP: Skepticism tweets, CTA: Call-to-Action tweets, OGA: On-ground activity tweets, GRV: Grievances tweets.

Protest	**Narrative**	**Unique Tweets**	**#Original Tweets**	**#Retweets**
CAA	QUEST	$13,380$	$31,504$	$2,308,221$
	SKEP	$15,274$	$19,586$	$3,869,627$
	GRV	623	$1,003$	$163,185$
	CTA	865	$1,699$	$149,205$
	OGA	647	932	$94,996$
BLM	GRV	$12,071$	$8,353$	$3,547,798$
	CTA	429	298	$68,002$
	OGA	161	158	$230,818$
FP	CTA-AP	564	$1,028$	$104,552$
	OGA	530	577	$59,985$
KTB	CTA	$2,958$	2931	$173,529$
	OGA	742	730	$19,562$

4.3 Results

This section discusses the results of the three research questions under study. This constitutes Step 4 in our framework, as shown in Figure 4.2. First, we discuss the different narratives present in all protests and converging narratives across protests. Next, we discuss the evolution of the converging narratives found across different protests under study. Finally, we discuss the user-retweet network obtained in different protests and how influential users are in disseminating different protest narratives.



(a) #CAA

(b) #BLM

(c) #FP

(d) #KTB

Figure 4.3: Clusters of narratives for CAA, BLM, FP, and KTB, respectively. While CAA, FP, and BLM show oga, cta, and grievances narrative clusters among other small clusters, we only find KTB tweets divided into oga and cta narratives clusters.

4.3.1 Different Narratives In A Protest

Per RQ 1, we examine the clusters formed in each protest campaign using Step 2 in our framework. We leverage the semantic difference in the clusters to identify plausible narratives in the campaign. We only report the tweet's cluster for which we could label the narrative. To annotate protest clusters into different narratives, we leverage the previous literature on protest studies in different parts of the world [167; 184; 118].

CAA

With the duplicate threshold set as 30, the number of active tweets for clustering in CAA was $36,109$. As shown in Figure 4.3(a), 6 clusters of tweets were formed for CAA in Step 2. We manually annotate randomly selected two sets of 10 sample tweets from each cluster to analyze narratives in the clusters. Table 4.3 shows the 5 different narratives clusters in the campaign with the highest engagement. The last cluster in CAA belonged to location-specific tweets. However, the annotators could not converge on the cluster as reporting on-ground activities. Regarding engagement (i.e., tweet/retweet activity), the largest cluster showed skepticism narrative. We found skepticism in tweets towards the Act, protesters, or supporters. An example of skepticism in CAA protest is "*As #CAAProtests spread across #India, police respond with an iron fist, brutally beating unarmed protestors. They're thrashing journalists, ordering TV channels to stop airing protest footage, and shutting down the internet. Here's what they don't want you to see. #PoliceBrutality.*". The second dominant narrative for CAA was the Questioning cluster, where the tweets posed questions to the Act, politicians, and protesters for violent actions. For example, "*The police showed patience and did not shoot. Who fired at 56 policemen in Lucknow? Those who are saying that they do not have any paper, are they who are the end? Listen to the story of Pakistani Hindu.*" represents a questioning tweet from the protest. The other three important narrative clusters included grievances (GRV), call-to-action (CTA), and on-ground activities (OGA), as shown in Figure 4.1.

Table 4.4: Top terms, emojis, and hashtags identified in CAA through Prominence score in OGA and CTA narratives.

	Top terms	**Emoji**	**Top hashtags**
CTA	initiative, we, require, showcase, stronger, trending, bhaktriot, trending	👇, 🍿, 🌹, 👀, 🔃, ☠️, 🤦‍♂️, 🙏, ✊, 👈, 🐖, 🕉️, 🔱, 🏹, 🍷	#jaishriram, #jaihind, #jihadists, #solidaritypledge
OGA	Assam, Punjab, struggle, reality, curfew, Tripura, Chennai, Northeast	📣, 💐, 😥, 🤦‍♂️, 🖤, 🚩, ✌️, 👈, 👉, 👇, ❓, ❗, 😠, 👍 ⁉️	#assam, #uttarpradesh, #curfew, #tripura, #section144, #keralagovt

We compute the semantic difference between clusters using *prominence score* provided in Equation 4.1 to shed light on the clustered tweet's performance. Table 4.4 shows that the top terms for OGA include state and location information. It gives evidence of users sharing location-specific on-ground activity on social media. We exclude the prominent terms that included the various forms of CAA (e.g., Citizenship, CAB, Bill) due to their redundancy through all the narratives. The top hashtags also include states in India (i.e., Assam, Uttar Pradesh, etc.). Since the offline protest broke out in different states, the top terms and top hashtags show the prominence of states in the cluster. The top emojis used in OGA were index-pointing fingers, loudspeaker, red flag, and black heart. The top mention in OGA included news editors, reporters, and ministers. On the other hand, the top terms for CTA included words like initiative, showcase, trending, and the notion of 'we', among others. The top hashtags also had a call-to-action context, including a pledge to solidarity (#solidaritypledge). Most of the top accounts under CTA are currently suspended by Twitter. At the same time, others included political party leaders. In CAA, we found that the OGA narrative more prominently mentioned news channel personnel, while common people were mentioned mainly in CTA.

Table 4.5: Top terms, emojis, and hashtags identified in BLM through Prominence score in OGA and CTA narratives.

	Top terms	**Emoji**	**Top hashtags**
CTA	bajos, catalunya, Ireland, iamantifa, reino, paises	🌶️, ✊🏻, 🍃, ✊🏽, 👏, 🤜, 💙	#iamantifa, #breakingnews, #shekubayoh, #globalrevolution
OGA	chauvin, derek, degree, charges, murder, attorney, charged, officers, abetting, aiding, minnesota	👏, 🔴, 🚨, ➡️, ‼️, 📹, 😡, 🔵, 😧	#derekchauvin, #minneapolis, #georgefloyd, #minnesota, #mugshot, #boxing

BLM

With the duplicate threshold set to 30, 3 clusters were formed for BLM (as shown in Figure 4.3(b)) that translated from 1,3151 active tweets to 4,338,427 total tweets. Table 4.3 shows the original tweets and retweets corresponding to different narratives clusters under study. The Grievance narrative cluster (GRV) was the most prominent in BLM, with 3,556,151 in total tweets. The on-ground activity narrative cluster was the second most prominent cluster in BLM with 230,976 total tweets, followed by CTA with 68,295 total tweets. Example tweets from the three different narratives in BLM are shown in Figure 4.1. The Grievances narrative included complaints of injustice to African Americans and black women and police brutality. The CTA tweets urged participation in the protest in different parts of the world. The OGA narrative reported updates on the government's actions and updates on ground-level developments. An example of OGA with updates on awaiting government decisions includes "*UPDATE in #GeorgeFloyd case: Attorney General Keith Ellison to elevate charges against Officer*

Derek Chauvin to Second Degree Murder. Also charging other 3 officers involved with aiding and abetting second-degree murder.".

We use the *prominence score* provided in Equation 4.1 to identify the top terms, hashtags, emojis, and mentions used in each narrative. Table 4.5 shows the results obtained through the *prominence score* for top terms, emojis, and hashtags for the OGA and CTA narratives in BLM. For CTA, the top terms include iamantifa (short for I am anti-fascist). Other CTA top terms included Reino, Ireland, among other locations, indicating locations for people participation. The top emojis in CTA include raising a fist, clapping, and blue heart. The top hashtags in CTA include #iamantifa, and #globalrevolution. The top terms in OGA included Derek Chauvin, murder, abetting, and Minnesota, indicating the reporting of daily developments on the Derek Chauvin trial. The top emoji for OGA includes red and blue circles, video camera, and police-car light. The top hashtag for OGA included #derekchauvin, #minneapolis, #georgefloyd, #minnesota, #mugshot, #boxing representing the on-ground activities happening at the moment. The top mention in CTA included accounts that are now suspended or deleted. In contrast, the mentions in OGA include the handles of the US senator from Minnesota and the Attorney General of Minnesota as well as news reporters.

FP

The duplicate threshold to give the best clustering result for FP is 30. We found 20 clusters for FP from Step 2 in our framework, as shown in Figure 4.3(c). However, we consider the top 3 clusters for further analysis, constituting more than 500 unique tweets each. Table 4.3 shows the 2 major narratives preset in FP. The most dominant narrative in FP was call-to-action against politicians (CTA-AP), which mapped 564 active tweets to $105,580$ tweets. The OGA cluster is the second most dominant narrative with 530 active tweets mapped to $60,562$ tweets. The third conclusive narrative cluster under FP is CTA, which comprises 806 active tweets, mapped to $54,256$ tweets. An example tweet of CTA-against politician is "*Every farmer boycotts such a government. We have to start the "Swachh Bharat Abhiyan" by boycotting the BJP.#BoycottBJP4Farmers*". While the cluster (denoted as CTA) called for participation in support of farmers, the cluster CTA-AP (i.e., Call to action against politicians) contained tweets against the ruling government for their bill proposal. Table 4.6 shows the results obtained through the *promi-*

Table 4.6: Top terms, emojis, and hashtags identified in FP through Prominence score in OGA and CTA narratives.

	Top terms	Emoji	Top hashtags
CTA	modisansad, kisansansad, strengthen, shipping, appreciating, represent	➡️, 🚜, 🧑‍🌾, 🌾, 🐖, ✊, 🙏	#myfarmer_mypride, #boycottbjp_4farmers, #modi, #india
OGA	wife, missing, arrest, bathinda, arrested, hospital, hindutva, survived, gazipur, toolkit, rajasthan	👉, 🖤, ✊🏿, ✊, 💔, 🎥, ⚠️, 🕉️, 💵, 🤟, 👆, 👍, 💚, 🎞️	#taliban, #toolkit, #pakistan, #freeranjitsingh, #lahore, #bangladesh

nence score. The top terms for the OGA narrative for FP included terms like arrest, missing, and locations, which were in line with the on-ground activity narrative. The most prominent emojis included black heart, broken heart, video camera, and money. The OGA narrative's prominent mention included NGO handles, politicians, and news outlets. The top terms and hashtags for CTA included appreciation and farmer's pride. CTA-AP included terms like nazi and socialism. The context-specific emojis of crops, farmers, and tractors were commonly used in CTA and CTA-AP. Prominent mentions in CTA-AP were of Bollywood actors, farmer's unions, and activists. CTA mentioned the prime minister among other activist accounts and a few suspended accounts.

KTB

The duplicate threshold for KTB was set to 5, as the data collected for the protest was small. With a duplicate threshold of 5, KTB contained $200,946$ total tweets, with $3,776$ original tweets and $197,170$ retweets. Table 4.3 shows the statistics of original tweets and retweets corresponding to different narratives clusters under study. The UK protest on the policing formed 2 clusters using our framework as shown in Figure 4.3(d). Among the two clusters, more engagement was around call-to-action. Figure 4.1 shows the example of tweets from both on-ground activities and call-to-action for the protest.

Using the *prominence score* provided in Equation 4.1 to identify the top terms, hashtags, and emojis used in each narrative, we demonstrate the efficacy of the manually labeled narrative clusters. Table 4.7 shows the top hashtags, emojis, and terms identi-

Table 4.7: Top terms, emojis, and hashtags identified in KTB through Prominence score in OGA and CTA narratives.

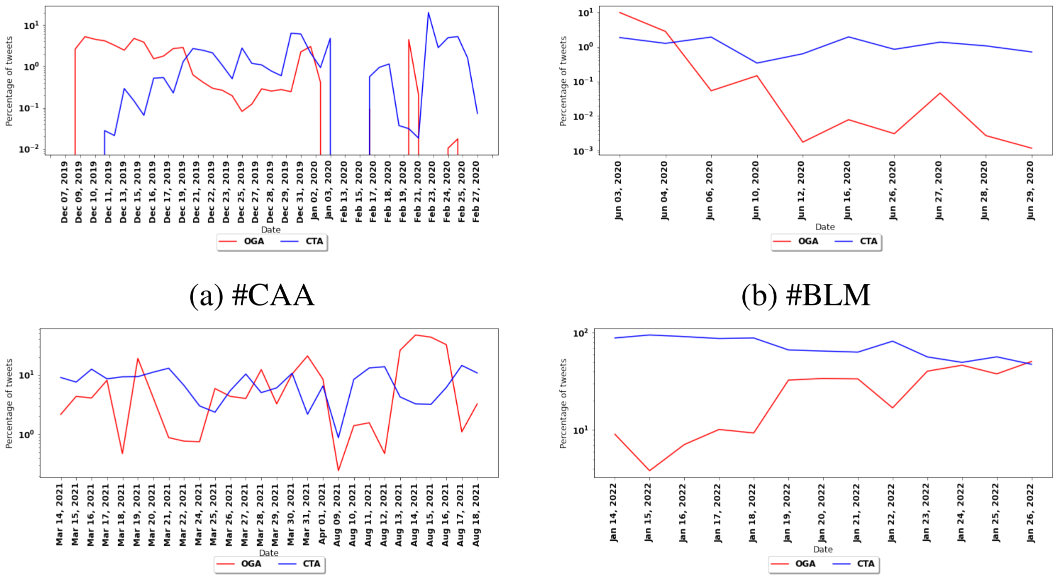
	Top terms	Emoji	Top hashtags
CTA	draconian, votes, peers, amendments, manchester, protesters, noise, saturday, activists	📢, ‼️, ⚠️, 🖤, 👀, 🎥	#killthebill, #pcscbill, #protestisnotacrime, #righttoprotest, #nationalityandbordersbill
OGA	toriesdevoidofshame, toriespartiedwhilepeopledied, toriesunfittogovern, borisjohnsonout, sue	🔥, 🤦‍♀️, 🤷‍♀️, 🐷, 🤬, 😷, 😤	#toriesdevoidofshame, #toriesoutnow, #toriesunfittogovern, #toriespartiedwhilepeopledied, #borisjohnsonout

fied in the two narrative clusters. The top prominent terms and hashtags for OGA in the KTB protest included narratives of shame against the Prime Minister and reporting of deaths. The emojis used included anger, face, facepalm, and fire. The mentions in OGA included the Prime minister's handle, members of parliament, and other politicians. While the CTA cluster top terms and hashtags included words like peers, places, and calling out activists, the top mentions included members of the green party and activists' handles.

Table 4.8: Total engagement in the narrative common across protests in our study. For FP, we report consolidated results for CTA and CTA-AP.

Protest	Narrative	Total Tweets	Tweet	Retweet
CAA	CTA	150, 904	1, 699	149, 205
	OGA	95, 928	932	94, 996
FP	CTA	159, 836	2089	157, 747
	OGA	60, 562	577	59, 985
KTB	CTA	176, 460	2, 931	173, 529
	OGA	20, 292	730	19, 562

Converging Narratives Across Protests: From the analysis of the clusters in the protests, we found that the protests might show specific clusters unique to the protest. There also exist narratives that are common across all protests. We found the presence of call-to-action (CTA) and reporting of on-ground activities (OGA) forming two persistent narrative clusters in all the protests under study. The other common narrative across protests is grievances or personal complaints [184]. While grievances play a vital role in contentious politics, they are highly subjective in nature [183]. Therefore, it becomes challenging to deduce meaningful narrative clusters for the grievances across the



(a) #CAA

(b) #BLM

(c) #FP

(d) #KTB

Figure 4.4: Evolution of the OGA and CTA narratives in the different protests, respectively.

protests. However, our proposed framework was able to form clusters with deductible characteristics for call-to-action and on-ground reporting of activities with similar features across the protests under study. The skepticism and questioning in CAA reveal the contention in the online social media about the Act. On the contrary, the FP and KTB protest was more in harmony with opposing the bill, with narratives formed majorly towards CTA and OGA. Similarly, the BLM protests contained clusters of CTA, OGA, and GRV, showing harmony in protesters towards the cause.

4.3.2 Evolution Of Narratives During Protest

Analysis of narrative shifts has been found to help us understand the story components obtained from various sources temporally [158]. In this section, we study the evolution of the converging narratives across the timeline of different protests. Since the common narratives across all the protests under study are OGA and CTA, we perform a comparative analysis of CTA and OGA to analyze how the two narratives evolve. Figure 4.4 shows the OGA and CTA narrative timeline of the 4 protests under study. To present the timeline of the OGA and CTA narratives per day, we calculate the percentage of a given narrative tweet over the total tweet produced that day. We plot the percentage of the total combined tweets per day on the y-axis in the log scale, while the corresponding dates are plotted on the x-axis. We compute the LDA on several peak days corresponding to OGA and CTA narratives to understand the discussion sequence through the timeline. We also analyze the timeline of events collected from various news articles about the protest's progress to better understand the story's evolution.

CAA

For CAA, the two most prominent narratives were skepticism (50.02% tweets) and questioning (30.09% tweets). The Grievances cluster consisted of 2.1%, while OGA and CTA constitute 1.2% and 1.9% of total tweets in the protest. The event timeline of the CAA protest is presented in Figure 4.5. Figure 4.4(a) presents the percent CTA and OGA narrative evolution for CAA daily. From the graph, we find that initially, tweets around OGA peaked on December 09, 2019, the day CAB was passed by the Lok Sabha (Lower House). We analyzed the topic on December 09, 2019, using LDA [33] from Gensim package [161]. We chose the number of topics as 4 in CAA based on the average coherence score for the dates selected for analysis. The coherence measure used was c_v, among the options present in gensim. The first topics in OGA for December 09, 2019, were around "Bill Proposal news" as it encapsulated terms *bill, citizenship, today, party*. The second topic for December 09, 2019, was "BJP role in CAA" as it covered terms like *BJP, Oppose, National, Bill*. On December 11, 2019, the Rajya Sabha passed the CAB bill. On December 11, CTA tweets started emerging, with CTA topics on the day including "Support the Act" with terms like *history, congratulation, thankful, support*. However, the percent share for OGA for December 11, 2019, was more than CTA. The OGA for December 11 includes "Bill is a Mistake" with terms like *still, mistake, together, India*. From December 21, 2019, percent CTA tweets were more than OGA tweets. We performed LDA for December 25, 2019, which reported more CTA narrative than OGA. The OGA narrative for December 25 included a topic around "People struggle" with terms like *struggle, public, resulted, look, Modi*. While CTA topic on December 25 includes "Urge to protest" with terms like *Indian, National, proclaim, we, participant, anthem*. Through multiple days in the protest, we analyzed that the topics discussed in the tweets were in line with the daily developments in the news and vice versa, with one media forming the inspiration for the other.

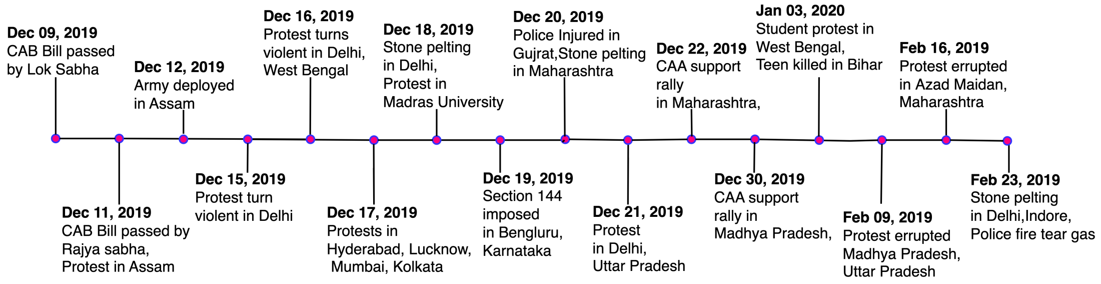


Figure 4.5: Timeline of major events during CAA protest. The timeline presents the bill's passing date by Lok Sabha, Rajya Sabha, followed by the events of the major protest that unfolded in different parts of the country.

BLM

In the BLM protest, 3 narrative clusters were formed. The most prominent cluster was formed for grievances, constituting 81% tweets during the protest timeline. The second prominent narrative during the protest was OGA which constituted 5% total tweets. The third prominent narrative was the CTA, constituting 1.5% of total tweets. The timeline of events that unfolded to form the BLM protest is shown in Figure 4.6. For the first 2 data collection dates, we found that the percentage of tweets per day for OGA was higher than CTA. On June 4, on-ground protests erupted in different parts of the US. [6] The OGA topic for the day included topic on "Charges on Derek Chauvin" with terms like *murder, change, 2nddegree, minneapolis*. The CTA narrative for the day included "Spread the protest" with terms like *people, say, spread, change, need*. On June 6, 2020, protests regarding the injustice continued in different parts of the United States. [7] On June 6; however, CTA narrative had more percent tweet than OGA narrative. The LDA topic on June 6 for OGA included "Murder charge" with terms *'murder, involved, officer, breaking, arrest*. While the CTA narrative for the day includes "March for protest" with terms *thousands, must, march, came, strong*. In #BLM protest, the percent CTA narrative remained more consistent throughout our timeline of data collection as compared to the OGA narrative.

[6] https://bushwickdaily.com/news/6540-updated-black-lives-matter-protest-schedule-for-june-4-2020/

[7] https://www.nbcnews.com/news/us-news/blog/2020-06-06-george-floyd-protests-n1226451

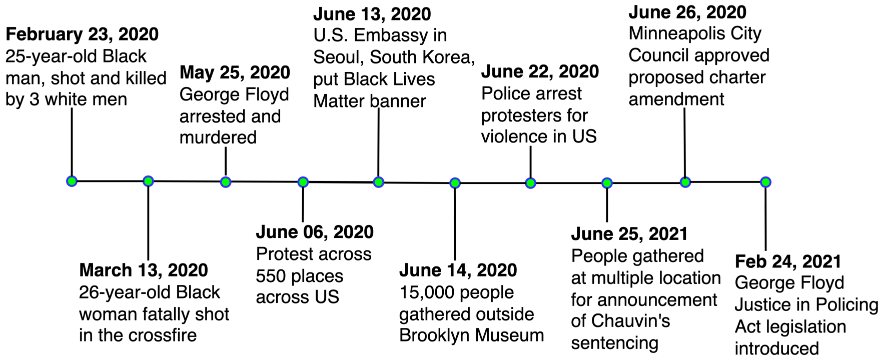

February 23, 2020
25-year-old Black man, shot and killed by 3 white men
March 13, 2020
26-year-old Black woman fatally shot in the crossfire
May 25, 2020
George Floyd arrested and murdered
June 06, 2020
Protest across 550 places across US
June 13, 2020
U.S. Embassy in Seoul, South Korea, put Black Lives Matter banner
June 14, 2020
15,000 people gathered outside Brooklyn Museum
June 22, 2020
Police arrest protesters for violence in US
June 25, 2021
People gathered at multiple location for announcement of Chauvin's sentencing
June 26, 2020
Minneapolis City Council approved proposed charter amendment
Feb 24, 2021
George Floyd Justice in Policing Act legislation introduced

Figure 4.6: Timeline of major events during BLM protest. The timeline presents the killing of a 25-year-old man by three white men, followed by the series of protest events that unfolded afterward.

FP

For Farmer's Protest 20 clusters were formed from Step 2 in our framework in Figure 4.2. In FP, The three bills under Farmer's Protest were passed by India in September 2020. The protests around the repeal of the bill started gaining momentum around September 25, 2021 [57]. The sequence of events that unfolded during the protest is presented in Figure 4.7. The bill was finally repealed by the government of India on November 29, 2021, ending a year-long protest. Figure 4.4(c) shows the timeline evolution of the Farmer's protest from March 14, 2021, till August 18, 2021, i.e., the midst of the protest. On March 14, 2021, the percent narrative for CTA was more than OGA. The LDA-based topic on the day for OGA included the topic of "Farmer's work" with terms like *farmer, think, hard, work, people, feed*. The CTA topic for the day included "Will to protest" with terms like *history, perseverance, victory, delay, depicting*. The news on March 14 included the camping of thousands of farmers, mainly from Punjab, Haryana, and Uttar Pradesh, at the Delhi border. [8] On March 19, 2021, Supreme Court-appointed panel members submitted their recommendations on the three farm laws. [9] On March 19, 2021, the OGA narrative percent surpassed CTA narrative. The LDA topic for OGA on the day included "Mislead on Bill" with terms *anything, trust, power, lie, don't, manipulate*. The CTA narrative on the day included "Take initiative" with terms *do, steps, value, 300, take*. From Figure 4.7, we observe that the CTA narrative was more consistent online. On the other hand, the OGA narrative fluctuated more, with

[8] https://www.indiatoday.in/india/story/farmers-protest-no-permanent-structures-to-be-built-at-protest-sites-farm-leaders-clarify-1779274-2021-03-14

[9] https://www.ndtv.com/india-news/supreme-court-appointed-panel-was-against-repeal-of-3-farm-laws-2834359

peaks of the narratives on specific dates, including March 19, 2021, March 25, 2021, March 31, 2021, and April 14, 2021.

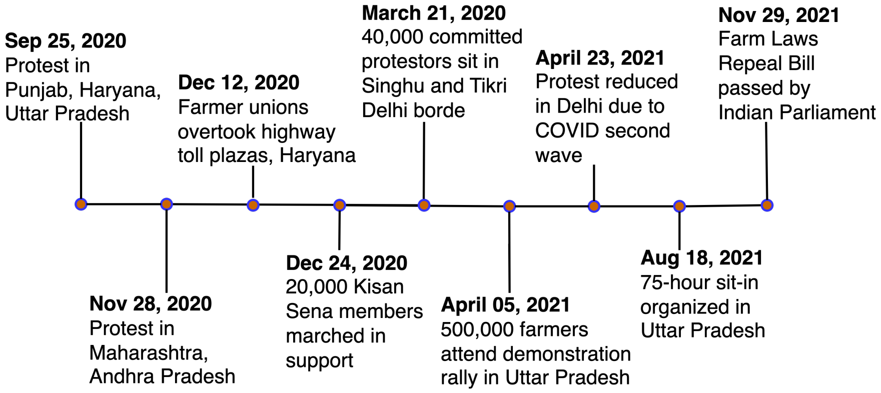

Sep 25, 2020
Protest in Punjab, Haryana, Uttar Pradesh
Dec 12, 2020
Farmer unions overtook highway toll plazas, Haryana
March 21, 2020
40,000 committed protestors sit in Singhu and Tikri Delhi borde
April 23, 2021
Protest reduced in Delhi due to COVID second wave
Nov 29, 2021
Farm Laws Repeal Bill passed by Indian Parliament
Nov 28, 2020
Protest in Maharashtra, Andhra Pradesh
Dec 24, 2020
20,000 Kisan Sena members marched in support
April 05, 2021
500,000 farmers attend demonstration rally in Uttar Pradesh
Aug 18, 2021
75-hour sit-in organized in Uttar Pradesh

Figure 4.7: Timeline of major events during FP protest. The timeline captures protests in different parts of India.

KTB

We collected data from January 14, 2022, to January 26, 2022, when the protest was gaining momentum in different parts of the UK. We found that on January 14, 2022, the percentage of tweets shared was higher for CTA than OGA. The news of the day included the opposition of the Labour house lords towards the amendments in the Police, Crime, Sentencing, and Courts bill. The LDA topics on the day for OGA included "Protest reporting" with terms *street, bill, police, Saturday, protest, right*. The LDA topic for CTA on the day included "Appeal for public support" with terms *labour, power, criminalize, amendment, you, protest, dangerous, must*. On March 22, 2022, the House of Lords rejected the proposed legislation again and demanded that the restriction on the bill be removed. The OGA topic on the day included "Brexit debate" with terms like *back, like, Brexit, brought, joined.*. The CTA topic on the day included "Protest participation" with terms including *victory, morning, community, power, criminalizing, stand, must, bill, change, over*. The CTA for the KTB protest was more persistent than OGA, however, in KTB in the week of data collection.

Narrative Evolution Analysis Across Protests: Among the 4 protests under study, the data was collected from the initial days of protest for #CAA and #KTB. However, the collection of data for #FP and #BLM is from the timeline when the protest had gained momentum. Among the 4 protests, we observed the CTA narrative as more

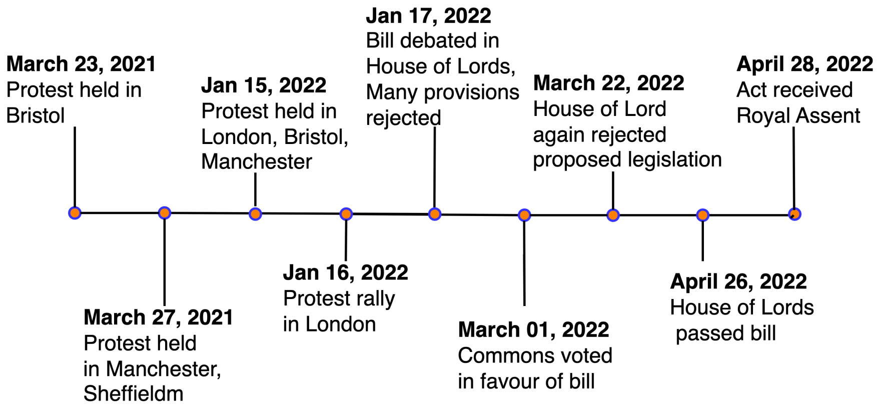

March 23, 2021
Protest held in Bristol
March 27, 2021
Protest held in Manchester, Sheffieldm
Jan 15, 2022
Protest held in London, Bristol, Manchester
Jan 16, 2022
Protest rally in London
Jan 17, 2022
Bill debated in House of Lords, Many provisions rejected
March 01, 2022
Commons voted in favour of bill
March 22, 2022
House of Lord again rejected proposed legislation
April 26, 2022
House of Lords passed bill
April 28, 2022
Act received Royal Assent

Figure 4.8: Timeline of major events during KTB protest. The timeline captures protests in different parts of the UK.

consistent throughout data collection. On the other hand, the OGA narrative fluctuated and often showed a rise whenever some significant developments were observed in protest. This analysis also supports the structure and definition of the OGA narrative cluster.

4.4 Discussion

This work investigates shared narratives and examines the evolution and communication around the shared narratives during a protest. To this end, we collect Twitter data from 4 protests from different demographic locations centered around anti-government policy or bill-related topics. We collect tweets for bill-related protests in India (CitizenshipAmendmentAct (CAA) and FarmerProtest (FP)) and the United Kingdom (KillTheBill Protest (KTB)). The inclusion of protests from different geographical and socio-political contexts enriches the study by offering a comparative lens. It helps discern patterns that may be context-specific and those that transcend geographical boundaries. For example, presence of common narrative elements such as call-to-action (CTA) and reporting of on-ground activity (OGA) were found across all four protests. However, we identify specific dominant narrative clusters for each protest. For instance, skepticism and questioning dominate the CAA protest, while CTA is prominent in KTB and FP. BLM, on the other hand, exhibits a dominant cluster focused on sharing grievances. This multi-protest approach enables a more robust understanding of how narratives evolve across diverse demographic locations and socio-political contexts. Since each protest represents a distinct anti-government policy or bill-related topic, the protests under study offers a comprehensive view of the commonalities in narrative themes. We collect data from the BLM (BlackLivesMatter) protest that led to the introduction of the George Floyd Justice in Policing Act in the US legislation. For all the 4 protests under study, we found the presence of call-to-action (CTA) and reporting of on-ground activity (OGA) narratives. The other standard narrative across protests was sharing grievances (GRV). Our analysis suggests that the narrative clusters can help reveal the underlying participant's motivation, based on which narratives are being discussed dominantly. We found skepticism and questioning were the two most dominant narratives for the CAA protest, indicating contention in public towards the bill. For KTB and FP, CTA formed the most dominant cluster indicating people's will to participate and motivate others. While in BLM, the cluster with grievances narrative was dominant, showing that people were reporting complaints and resentments for what had happened in large numbers. With the help of the prominence score, we found a pattern of emojis, hashtags, and mentions used in protest-related tweets. We found that the emojis used in the protest were mainly protest-centric. For example, the FP protests had tractor and corps as emojis, while CAA had more religious-based emojis. The mentions in the tweets provide evidence that OGA has more verified accounts tagged. In contrast, the CTA mentions more of the general public, some suspended across protests under study. In terms of narrative evolution, we saw that CTA was more consistent throughout the protest timeline, while the OGA narrative peaked around substantial developments around the protest. For capturing the communication centered around different narratives, we examine the narrative-sharing behaviors for the top 5% Influential users based on the out-degree centrality of the retweet network. Across the 4 protests under study, we found low Betweenness centrality; and high Eigenvector centrality. This indicates that across the protests, the users don't form more substantial edges between other users (Betweenness) but were connected to more Influential users (Eigenvector) and were able to have a faster flow of tweets in the network. Across the protests under study, we were able to identify narrative-centric community formation, indicating that some sub-communities centered around a single narrative.

CHAPTER 5

UNDERSTANDING OPPOSING STANCES DURING PROTEST: AUTHENTIC AND INAUTHENTIC ACTORS

In Chapter 3, we discussed the different strategies used by the counterpublics to conduct protests on social media platforms. In the network-based analysis, we established that the protest on social media combines centralized and decentralized information aggregation showing that the protest might contain organizational participation and genuine activists fighting for the cause.

This chapter focuses on protests in social media around contentious topics that may lead to divergent discourse. While we study the strategies and narratives shared during discourse, we need to account for the content incorporated by inauthentic actors. In this study, we consider user data, including user profile information [116], a network of users involved in the protest [209] as well as the content of the tweet [70], for inauthentic (bots, suspended and deleted users) and authentic actors respectively. This multifaceted approach allows us to differentiate between genuine voices and potential attempts at manipulation, providing a more accurate assessment of the protest strategies and narratives during a divergent discourse.

Towards this end, we investigate the user's perception of the #CitizenshipAmendmentAct on Twitter, as the campaign unrolled with divergent discourse in the country. We study 9,947,814 tweets produced by 275,111 users during the starting 3 months of protest. Our study analyzes user engagement, content, and network properties with online accounts divided into authentic (genuine users) and inauthentic (bots, suspended, and deleted) users. Our findings show different themes in shared tweets among protesters and counter-protesters. We find the presence of inauthentic users on both sides of the discourse, with counter-protesters having more inauthentic users than protesters. The following network of users suggests homophily among users on the same side of discourse and a connection between various inauthentic and authentic users. This work contributes to filling the gap in understanding the role of users (from both sides) in a less studied geo-location, i.e., India.

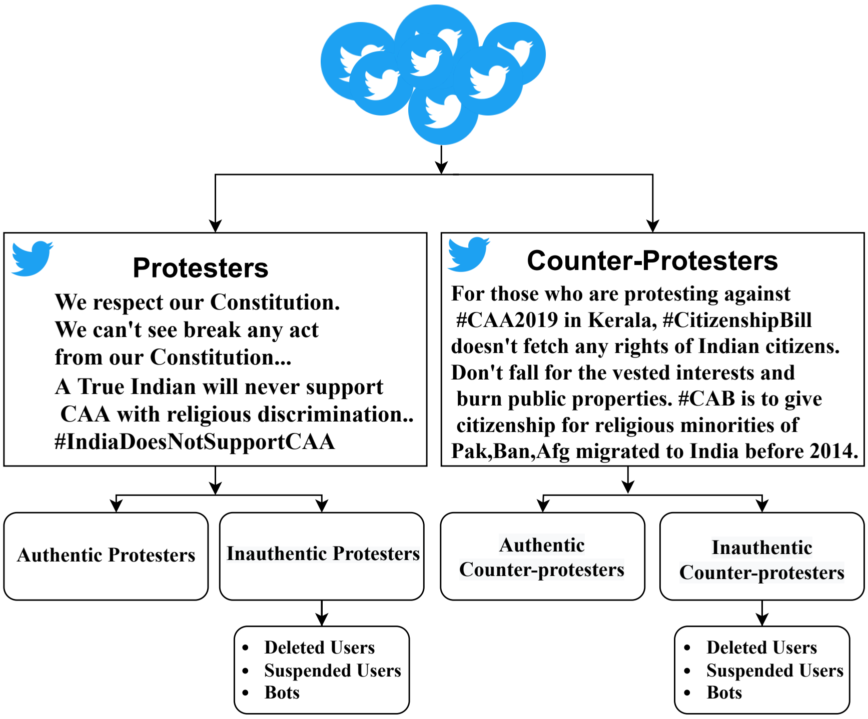

Protesters
We respect our Constitution.
We can't see break any act
from our Constitution...
A True Indian will never support
CAA with religious discrimination..
#IndiaDoesNotSupportCAA
Counter-Protesters
For those who are protesting against
#CAA2019 in Kerala, #CitizenshipBill
doesn't fetch any rights of Indian citizens.
Don't fall for the vested interests and
burn public properties. #CAB is to give
citizenship for religious minorities of
Pak,Ban,Afg migrated to India before 2014.
Authentic Protesters
Inauthentic Protesters
Authentic Counter-protesters
Inauthentic Counter-protesters
Deleted Users
Suspended Users
Bots
Deleted Users
Suspended Users
Bots

Figure 5.1: Users considered under study are divided into 4 sets. We first divide the users into protesters (users who opposed the bill) and counter-protesters (users who supported the bill). Protesters and counter-protesters are further divided into authentic and inauthentic users based on whether they were genuine users (Authentic users) or were identified as bots, suspended, or deleted by Twitter (Inauthentic users).

5.1 Introduction

In India, the first Citizenship Act was enacted in 1955, which enlisted the routes to obtain citizenship in India, which include birth, descent, registration, naturalization, and acquisition of a foreign territory. The amendment in the Act in 2019 (CAA 2019) allows the minority communities to apply for citizenship via registration or naturalization [39], with the caveat that migrants who have faced religious persecution in Afghanistan, Bangladesh, or Pakistan, can seek citizenship in India if they have entered India on or before December 31, 2014 [39]. Debate on the Act's non-secular roots was rooted in excluding other religious minority communities like Rohingya Muslims, Jews, Bahais, and Zoroastrians from seeking citizenship. Protesters deemed it unconstitutional to discriminate on religious grounds, as only certain persecuted illegal immigrants benefited from the Act. At the same time, the supporters / counter-protesters based their argument on the presumption that refugees of particular minority religious communities are more in need of asylum [39].

The enactment led to a divergent discourse on social media, with users divided in their opinion on the Act. Among the users who participated in the debate, one cohort rejected the Act, while another supported it. We define the users who reject the Act as protesters. Protesters were contested by a counter-protest campaign that questioned the protest and favored the Act. We define the users who were in favor of the Act as counter-protesters [70]. While the campaign gained traction on both Twitter and the offline world, the prevalence of manipulation of the campaign was found to be evident [87]. Given that the forms of manipulation of a discourse keep on innovating, it becomes crucial to filter the influence created by the inauthentic users in an online campaign. We define bots [179], suspended and deleted users (who tend to disseminate malicious content [1]) who participated in the discourse as *Inauthentic users*. In contrast, *Authentic users* are defined as the users who were not identified as bots, neither were suspended nor deleted. We thus study the online debate on the #CitizenshipAmendmentAct on Twitter with the participants divided into authentic and inauthentic users for both protesters and counter-protesters forming 4 set of users as shown in Figure 5.1.

Twitter has been the focus of various characterization studies involving online campaigns [70; 56; 151]. However, the characterization of a campaign concerning various sorts of authentic and inauthentic actors in discourse is limited [40]. To the best of our knowledge, we are the first to conduct a characterization study of a campaign with various users (Figure 5.1) in a less investigated setting, i.e., India. Our analysis contributes to a few recent preliminary studies on the CAA [124; 87], which provide a very coarse-grained analysis of the Act. We focus on a broader study of the Act, covering a larger dataset, multi-lingual tweets, and a richer analysis.

We analyze 275,111 users who post about topics relevant to CAA during the initial three months of the debate from December, 2019 to February 2020. We seek to understand the interplay of authentic/inauthentic users and pro- / against stance on CAA and investigate the presence and participation of inauthentic users on both sides of the discourse. For the characterization study, we first identify the stance of the participants

[1] https://help.twitter.com/en/rules-and-policies/enforcement-options

using an unsupervised stance detection approach [159]. We further study the 4 set of participants from the user, content, and network perspective to obtain a fine-grained analysis of the discourse. Broadly, we aim to answer the following research questions (RQs) through the characterization study of CAA.

RQ 1: *How are the protesters and counter-protesters involved in conducting the online campaign with respect to authentic and inauthentic users?*

Inauthentic user preferences have been studied in online campaigns, including elections [30], and more recently, the coronavirus [60]. In the CAA debate, we found the prevalence of inauthentic activity on both sides of the debate, with the online protest being highly mediated by the inauthentic users.

RQ 2: *What did the users in the discourse discuss about?*

Discourse analysis helps identify various themes in the discussion to help understand the user's perception [109]. While the themes for protesters / counter-protesters vary in CAA, we also found a difference in themes for authentic and inauthentic users in both sides, with inauthentic users posting lesser emotional content than their authentic counterparts.

RQ 3: *What was the network structure of the users?*

Network structure analysis helps examine issue alignment [209] and polarization around a controversial topic [73]. The follow network of users shows homophily, where users with similar stances follow each other more than users with opposing stances. Analysis of the follow network shows edges between authentic and inauthentic users, showing a risk of exposure of content from inauthentic users to the authentic users. Our findings reveal the interplay of inauthentic and authentic users. Prevalence of inauthentic activity was found on both sides of the debate. However, user characterization reveals that inauthentic users are more prevalent in the counter-protesters than protesters. Content analysis of the 4 set of users shows that the inauthentic users highly mediated the online protest. Emotional analysis of the content posted by the 4 set of users shows that inauthentic users use fewer emotional tweets than their authentic counterparts. Through follow the network of users, we found evidence of homophily in the network.

In this work, we contribute to the use of social media manipulation in other than western context during an online protest and study the online debate with different user's involvement in India, a country in Asia-pacific.

Table 5.1: Manually identified protest and counter-protest hashtags from trending topics during the period of data collection used for data collection.

Protest #tags	#CABProtest, #IndiaRejectsCAB, #HindusAgainstCAB, #SC-STOBC_Against_CAB, #IndiansAgainstCAB, #IndiaAgainstCAA, #CAA_NRC_Protest, #CAAprotests, #CAA_NRCProtests
Counter-protest #tags	#IsupportCAB2019, #HindusSupportCAB, #IndiaSupportsCAB, #ISupportCAA_NRC, #MuslimsWithNRC, #CAA_NRC_support, #ISupport-CAA
Ambiguous #tags	#CAB, #CABBill, #cab, #CAB2019, #CitizenshipAmendmentAct, #caa, #CABPolitics, #CitizenshipAmmendmentAct

5.2 Data

Using the official Twitter API, we collect tweets around CAA between December 07, 2019, and February 27, 2020, through daily trending hashtags around the topic. The list of hashtags used for data collection is shown in Table 5.1. Our collected data consists of 11,350,276 tweets, with 1,543,805 unique tweets and 9,806,471 retweets from 931,175 users. We first collate all the tweets from a given user to identify users actively tweeting about the topic. Hence, we consider users who have at least five tweets during the period

Table 5.2: Table with the on-ground activities coincident with peak tweet dates. Peak dates represent the dates with the highest number of tweet activity. Tweets count corresponds to a three-day rolling average of tweets calculated for each day (calculated with one day before and one day after).

Date	**Tweets**	**On-ground activities**
December 11	158,134.33	CAB passed by the upper house of parliament [50].
December 16	376,788.00	Student protests in Delhi [211].
December 17	379,699.00	Protest turns violent in Uttar Pradesh, Delhi, West Bengal and relaxed in Guwahati [18; 101].
December 20	436,616.33	Protesters turn violent with stone pelting in Gujarat, police vehicle burnt in UP, journalists detained in Kerala [4].
December 22	783,662.33	Protesters arrested, Women protest in Guwahati [178].
December 24	503,779.00	Protesters die due to bullet injury in UP [2].
December 30	276,724.33	Counter-protest rally in Madhya Pradesh, Indian-American protests in Washington [3; 100].
December 31	312,569.66	Nation wide protests [102; 5].

of data collection. The total number of users after the filtration process came down to 276,149.

Data Pre-Processing: Twitter users often use various emoticons, emojis, media links, hashtags, and other non-alphabetic characters. The informal nature of Twitter often leads to spelling and grammatical errors or incomplete sentences.

Thus, we follow the below list of pre-processing steps for the tweets before further analysis.

1. Removal of all links and mentions from the tweets
2. Removal of "RT" keyword from the beginning of retweets
3. Split of the camel case words into distinct words
4. Removal of punctuation marks
5. Removal of extra spaces
6. Replacement of digits with the word <number>
7. Case-folding where we lower-cased letters
8. Desertion of the tweet if it had lesser than three terms left after all the above steps

After the pre-processing steps, 1,038 users were disregarded for further analysis. The study conducted in the work was thus on the 275,111 users, who were most active during the campaign, and their tweets contained substantial information for further analysis. For further division of the users into authentic/inauthentic, as shown in Figure 5.1, we query the Twitter API and botometer [218] on the user IDs obtained from tweets.

Inauthentic users we consider for the study include suspended users, deleted users, and bots. Table 5.6 shows the total number of deleted and suspended users identified through querying the official Twitter API. We further collect the follower network using the official Twitter API for the users who were not deleted/ suspended/ private. We use Botometer [218], a tool used to identify a Twitter user as being automated (partially or fully) or not. Due to botometer API constraint, we collect the bot score for randomly selected 26,110 users (roughly equal to the total number of suspended / deleted accounts in our dataset). We use the *Cumulative Automation Score* (CAP) score metric provided by the API to identify a user as a bot account.

On-Ground Activity: To identify the impact of on-ground activities on opinion sharing around CAA, we manually curate the on-ground activities of the peak tweeting days, as shown in Table 5.2. The first online tweet peak was seen on December 11, 2019, which coincided with the bill passed as an Act by the Rajya Sabha (upper house) of the Indian parliament [210]. However, the highest peak was found on December 20, 2019, 9 days after the bill became an Act. On December 20, 2019, protesters around the country turned violent. A major protest was witnessed about the CAA bill in Guwahati (a north-east city of India) on December 10, 2019, which was the beginning of the chain of protests in certain parts of the country.

An Anonymized version of our data is available at `https://precog.iiit.ac.in/resources.html`

5.3 Understanding Discourse Through Unsupervised Stance Detection

To capture the fine-grained divergence among the users, we build on the previous work by [159] that uses text-feature for identification of user's stance during a political campaign. We further identify the themes in shared tweets and discuss the presence of inauthentic users in the discourse. Based on the online discourse on the Act, we identify two cohorts of users. We call the users who opposed CAA protesters. In contrast, users who share tweets in support of CAA are called counter-protesters. [159] proposed unsupervised stance detection techniques based on the text of the tweets. Another reason for the choice of algorithm is to surpass the manual annotation required in a supervised setting.

The ground truth labeling process for the seed set of users constitutes of two steps:

(1) Manual Labelling: First, we manually identify a set of hashtags indicating stance, as shown in Table 5.1. We identified 27 hashtags as counter-protest hashtags on manual inspection, which occurred in over 1.3 million tweets. The count of protest hashtags was 48, which accounted for around 1.04 million tweets. In the first step of labeling, if a user used only counter-protest hashtags and never used protest hashtags, we label the user as counter-protester. Similarly, if a user used only protest hashtags, we classify the user as a protester. In the first level of manual labeling, we identified 106,605 users as counter-protesters and 79,493 users as protesters.

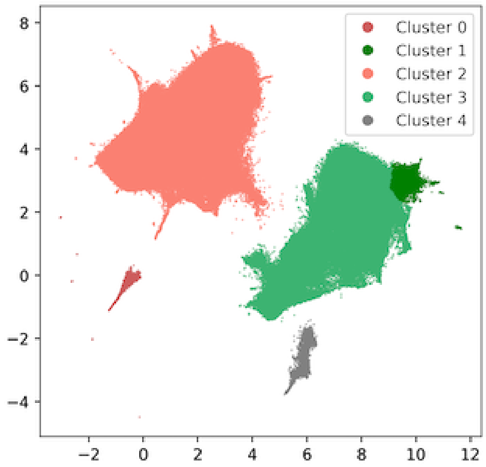


Figure 5.2: Here, Clusters 0 and 2 represent counter-protest users and Clusters 1 and 3 represent protest users. Cluster 4 had a purity below 80% and hence was not considered.

(2) Label Propagation: Around 86% of the tweets in our dataset were retweets. Based on the tweets that a user retweets, users were further labeled such that a user with at least 15 retweets from protest and none from the counter-protest side belongs to protesters. The intuition behind this approach is that the users retweet a given tweet if it aligns with their stance. We conduct this approach for two rounds. After the two rounds of label propagation, 114,977 users were identified as counter-protesters, while 79,613 were identified as protesters. The tweets of identified users were further pre-processed, and users with less than five tweets were disregarded. The final set of users after the pre-processing is 270,889.

Embedding-based Stance Detection: Word-based embedding can capture fine-grained divergence between two sets of cohorts [159]. We apply LASER (Language-Agnostic Sentence Representations)[2] to obtain 1024-dimensional embeddings of users based on their tweets. LASER is a sentence encoder trained in 93 languages, including many Indian regional languages. To obtain user-level embedding, we use the average of the vector for the filtered tweets. Users are then projected in a 2-dimensional space using Uniform Manifold Approximation and Projection (UMAP) algorithm [133]. Projection of users on lower dimensions helps overcome the curse of dimensionality [204]. UMAP projects the data elements closer if they are similar, while dissimilar data elements are

[2] https://github.com/facebookresearch/LASER

placed far apart. Projected user vectors are further clustered using hierarchical density-based clustering (HDBSCAN) [132]. Using the HDBSCAN algorithm, 5 clusters were formed with 270,889 users.

We consider clusters pure if they contain at least 30% of labeled users obtained via label propagation. We found 4 clusters have more than 80% purity of labels, as shown in Figure 5.2. Clusters 0 and 2 were identified as counter-protesters, while clusters 1 and 3 were identified as protesters' clusters according to the labeled users. Number of users identified in the 4 clusters was 263,869, with 142,839 counter-protesters and 121,030 protesters.

Topics Discussed By Users In Different Clusters:

Among the 4 clusters with high purity, the protesters are represented with shades of green, and counter-protesters are represented with shades of red, as shown in Figure 5.2. Two major clusters of opposing views (cluster 2 and cluster 3) show rich discourse on the topic. For manual inspection of assigned clusters, we randomly picked 4 sets of 10 users from each cluster and annotated all tweets for these users. We found the users in the clusters were indeed on the protester and counter-protester side, as identified through label propagation. To understand the theme of the 2 protester's clusters and 2 counter-protesters clusters, we go through all the tweets from the 4 sets. Topics discussed by the two cohorts in the 4 clusters shown in Figure 5.2 follow different themes as follows:

Cluster 0: (Counter-protesters) On a more thematic side, we found that the topics discussed by the users in Cluster 0 are mostly informative, with users sharing opinions on why CAA should be implemented.

Cluster 2: (Counter-protesters) Primary topic discussed by the users of this cluster includes questioning the protester about their actions and reasons for their disagreement with the implementation of CAA.

Cluster 1: (Protesters) Users in this cluster were tweeting about the on-ground activity of the protest, including public demonstrations, stone pelting, etc.

Cluster 3: (Protesters) Users in the cluster were posting informative tweets about CAA in the protest context.

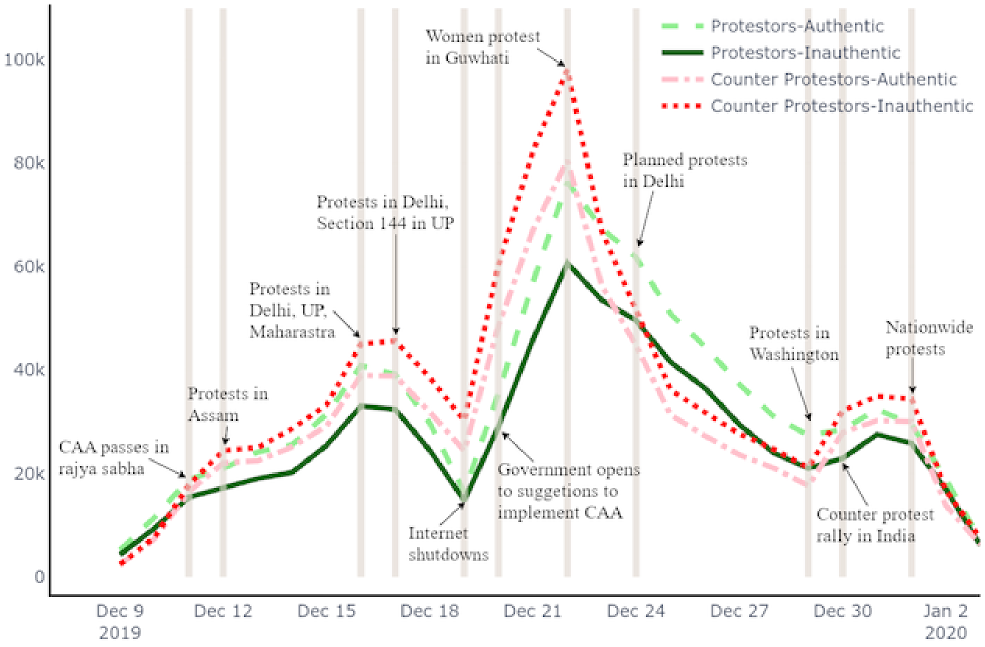


Figure 5.3: Timeline of counter-protest and protest vs. on-ground activity. Tweets produced by the Inauthentic users were more than the Authentic users during the timeline of the protest.

5.4 Content Characterization

Through content characterization, we try to understand the interplay between the online and offline activities during the period of data collection and quantify the difference in opinion among the 4 set of users.

5.4.1 Online (Twitter) Vs. Offline (On-Ground) Activity

Taking cues from previous works around planned protests [26; 138], we investigate the interplay of the online and on-ground activities during the CAA discourse, with respect to the 4 set of users in Table 5.6. Figure 5.3 shows the frequency of tweets by the 4 set of users during the 2 month of the protest period. The x-axis represents the days of protest taken as a rolling average of 3 days (one day before the date and one day after). On-ground activities corresponding to peaks in tweets are listed in the Table 5.2. First peak in the dataset was on December 11, 2019, when the CAB (Citizenship Amendment Bill) was passed by the upper house of parliament and officially became an Act [50]. Students in Assam held protest opposing the Act [1] on this day. In the initial few days, authentic protesters were more active than inauthentic protesters. While there was almost an equal proportion of authentic vs. inauthentic tweets during the initial days of passing of the bill. Another significant day was *December 16, 2019*, when students led the protest across the country, including Delhi, Maharashtra, and UP [211]. Anarchy was observed the same day in West Bengal, where people torched trains and staged sit-ins on the railway tracks [17]. Inauthentic counter-protesters made most tweets on this day, followed by authentic protesters. On *December 17, 2019*, several metro stations in Delhi [187] were closed, and Section 144[3] was imposed in UP. The previous trend of high tweets from inauthentic counter-protesters followed by high tweets from authentic protesters continued.

December 20, 2019 witnessed nationwide protest eruption, including states of Uttar Pradesh, Tamil Nadu, and Delhi [4]. The government opened to suggestions and reached out to the protesters [4]. While the inauthentic counter-protesters were more active than the inauthentic protester during the period, authentic counter-protesters made more tweets on around December 20 than authentic protesters. *December 22, 2019* had the largest peak in the dataset with on-ground counter-part of protesters being arrested and women leading the protest in Guwhati [178]. Both Inauthentic and authentic counter-protesters were more active around this day. *December 24, 2019* showed the second largest peak in the dataset, which coincided with the protester's death in Uttar Pradesh due to bullet injury [2]. Spikes on *December 30, 2019* and *December 30, 2019* found counter-protesters more actively posting than protesters. On-ground activities for the day included continued protests in different parts of the country as well as abroad in Washington [3]. Counter-protesters started rallies on *December 30, 2019* in support of CAA in different parts of the country [100]. One of the dip in tweets that we found was on *December 19, 2019* when the Internet was shut down in many parts of the country [6].

The counter-protesters had more inauthentic activities during the start of the timeline, until the largest peak. After which both authentic and inauthentic protesters showed more activity than counter-protesters. While there was a mix of authentic and inauthentic activity found in both protesters and counter-protesters, the activities of inauthentic counter-protesters were always more than than the authentic protesters. While, in case of protesters, authentic users always dominated the conversation. A common pattern in all the peaks found was that more that 90% of the authors in the timeline during any

[3] https://www.aninews.in/news/national/general-news/up-section-144-imposed-in-rampur-after-protest-against-caa20191217125542/

peak were from inauthentic users.

Table 5.3: Summary of topics for authentic and inauthentic protesters.

Authentic protesters topics	
Topic 1	india, bjp, police, muslim, student
Topic 2	police, student, hindu, assam, people
Topic 3	muslim, jamia, anti, student, delhi
Inauthentic protesters topics	
Topic 1	muslim, protest, hindu, student, protest
Topic 2	country, display, protest, acceptance, together
Topic 3	people, india, protest, police, citizenship

5.4.2 Difference In Opinion

We use LDA [33] for topic modeling and word shift graphs [69] to understand how diversified content were posted by the 4 set of users during the discourse.

Table 5.3 shows the topics discussed among the authentic and inauthentic protesters. Two of the dominant topics in authentic protesters had religious words, including hindu and muslim. The third topic included police and places of protest. While, the inauthentic protesters had one topic on religion, other 2 major topics included, citizenship, country and India as words. Table 5.4 shows the topics discussed by the authentic and inauthentic counter-protesters. While one major topic from authentic counter-protesters was support, hindu and caa, the second topic included politicians, and country. For authentic counter-protesters, best coherence score yields 2 topics. The inauthentic counter-protesters had one major topic including politicians, while another two dominant topics included citizenship, India and democracy and support as narrative. We report the most significant topics from the 4 set of users due to limited space. From the above analysis, we conclude that both protesters and counter-protesters discussed topics around religion, politician and the Act in general. However, the inauthentic users share content very similar to authentic counter-part, thus risking authentic users into believing them as authentic users. Next, we gauge the frequency of usage of various topics by the 4 set of users through word-shift graphs [69]. We use Shannon's entropy as a measure of diversity, where high Shannon entropy implies the text is less predictable [70] implying more diverse content. Figure 5.5(b) shows that protesters talked more about student, while counter-protesters talked more about hindus. We further study what do

Table 5.4: Summary of topics for authentic and inauthentic counter-protesters.

Authentic counter-protesters topics	
Topic 1	caa, support, people, anti, hindu
Topic 2	india, narendramodi, today, country
Inauthentic counter-protesters topics	
Topic 1	narendramodi, amitshah, kapilmishra_ind, delhi
Topic 2	hindu, support, indian, pakistan, citizenship
Topic 3	caa, support, democratic, india, humanitarian

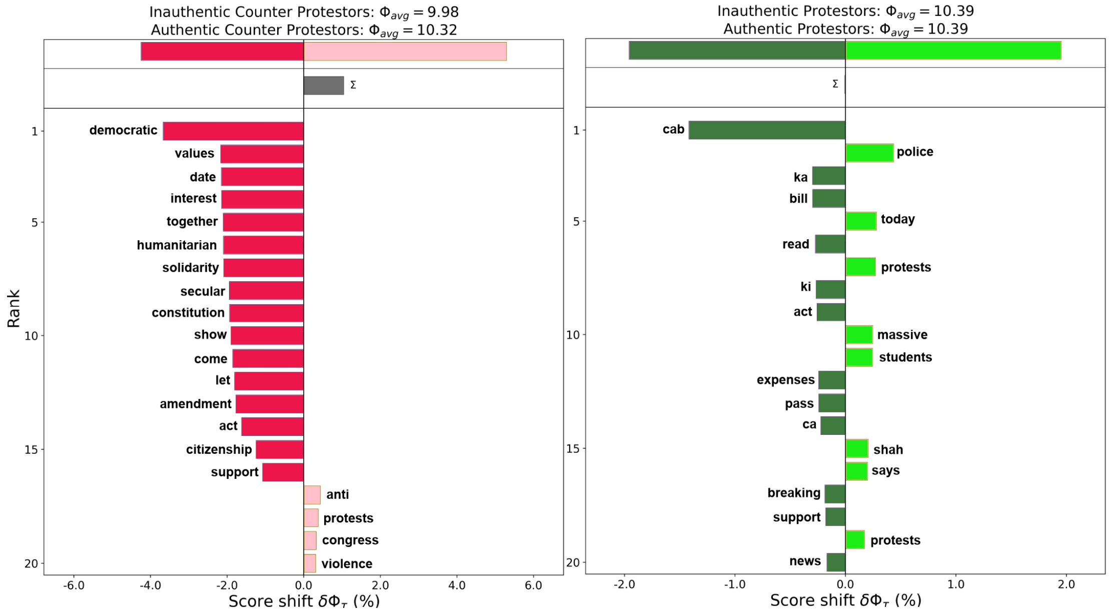
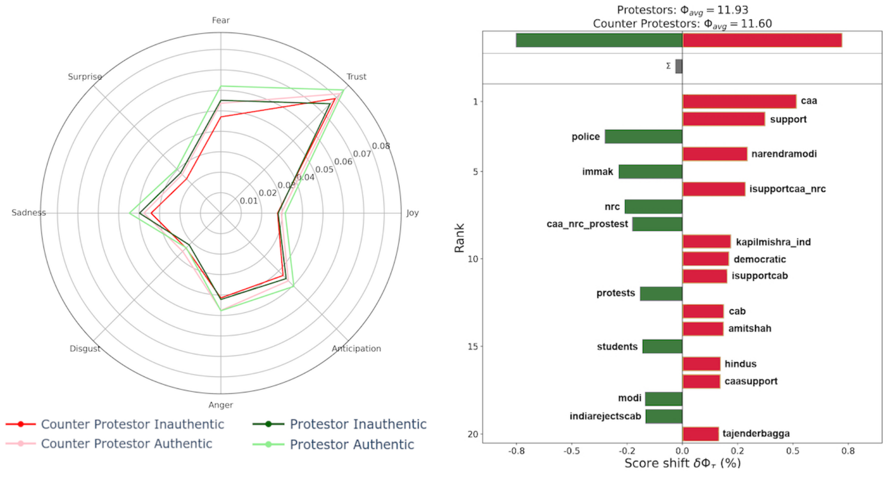


Figure 5.4: Application of word-shift graph for highligting narratives that charactrize 4 set of users.

the authentic and inauthentic protesters / counter-protesters share more frequently. Figure 5.4 shows that inauthentic counter-protesters are more appealing (e.g: humanitarian, solidarity, secular), while inauthentic protesters more frequently use words that show mistrust in government. Authentic users on both side are more frequently talking about protest and violence.

5.4.3 Emotion Analysis

We use NRC lexicon [135] that consists of 8 emotions developed from crowd-sourced manual annotation to identify the emotions of the users in the 4 set of users considered in the study. The NRC lexicon uses the plutchik's 8 wheel of emotion for English, as well as other translated Indian languages. The 8 emotions that are used in the analysis include, anger, anticipation, disgust, fear, joy, sadness, surprise, trust. Figure 5.5(a) shows that the authentic protesters had most dominant emotions for all the 8 categories.



(a)

(b)

Figure 5.5: Figure (a) presents radar plot to show the 4 set of users and their plutchik-8 emotions. Figure (b) shows the application of word shift graphs for highlighting narratives that characterize protesters and counter-protesters. Protesters are shown in green, while counter-protesters are shown in red.

The authentic counter-protesters and inauthentic protesters had almost similar emotions for fear, surprise, sadness. The inauthentic counter-protesters had least emotional content among the 4 set of users.

5.5 User Characterization

Presence Of Authentic And Inauthentic Users In Discourse: We identify users based on their authentic behavior to study the role of inauthentic users in mobilizing protests and counter-protests. As shown in Table 5.5, among the 263,869 users considered for the analysis, we found Twitter suspended 13,871 users. In comparison, 13,251 users were not found (referred to as deleted users) when queried for follower network. The number of non-authorized (private users) was 5,844. We were unable to retrieve information of 11,091 users using Twitter API. Inauthentic users obtained so far are 27,122, including suspended and deleted users. Next, we use botometer API [218] to identify bot users. Given the limitation of botometer API, we randomly pick 27,122 users from the rest of the users to query botometer for bot scores. We could retrieve bot scores for 26,110 users, out of which 14,970 were counter-protesters, and 11,140 were protesters. Table 5.6 shows the complete set of users considered for the analysis.

Findings: Through user characterization, we infer that both sides of the discourse had

Table 5.5: Distribution of suspended and deleted accounts in protesters and counter-protesters in the dataset.

	Suspended Users	Deleted User
Counter-protesters	8655 (62.39%)	7440 (56.16%)
Protesters	5216 (37.60%)	5806 (43.83%)

Table 5.6: Distribution of authentic and inauthentic users in dataset.

Total Users	53,227
Suspended Users	13,871
Deleted Users	13,246
Bots (CAP score>=0.8)	4,664
Authentic Users	21,446

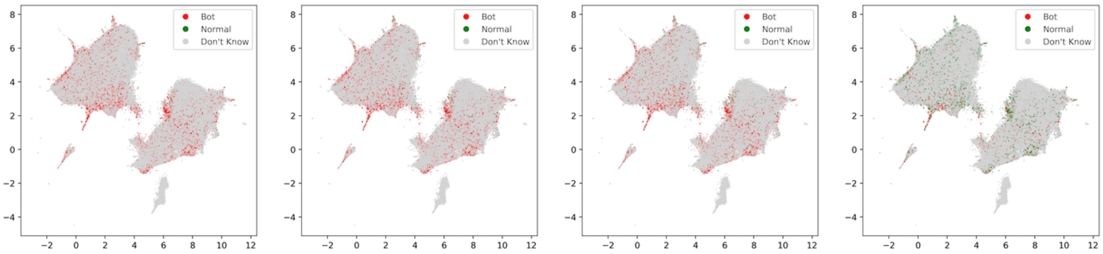
Table 5.7: Distribution of bots in the discourse with varying bot scores. P: protesters, CP: counter-protesters, T: total number of users for which bot score is known in our analysis.

Bot score (>=)	CP (% bots in CP)	P (% bots in P)	Total (% bots in T)
0.8	2,589 (17.29%)	2,075 (18.62%)	4,664 (17.86%)
0.7	11,359 (75.87%)	8,214 (73.73%)	19,573 (74.96%)
0.6	12,706 (84.87%)	9,096 (81.65%)	21,802 (83.50%)
0.5	13,500 (90.18%)	9,688 (86.96%)	23,188 (88.80%)

suspended and deleted users and bots. Counter-protesters had more than 50% suspended or deleted users on the platform, as shown in Table 5.5. Figure 5.7 shows the distribution of bots in the stance-based cluster. We notice, as shown in Figure 5.6 and Table 5.7, that as the bot score varies from 0.8 to 0.5, there is a sharp decline of bots above 0.7. This shows the presence of semi-automated accounts in the discourse.

5.6 Network Characterization

To determine if protesters and counter-protesters are in homophily and how authentic and inauthentic users are connected, we study the follow network of users in our dataset. We build a follow graph induced by the users in the dataset for network characterization. Users for whom the follow network was obtained from Twitter API exclude private accounts and accounts for which information was not obtained due to API constraints. The final follow network was obtained for 226,412 users. First, 5,000 followers were retrieved from Twitter API for each user from the sample. We consider the graph of 226,412 users as G. A directed edge from user x to user y exists if x follows y. We use this convention to ensure the network under study is campaign-specific, as partic-



(a) Bot Score>=0.5 (b) Bot Score>=0.6 (c) Bot Score>=0.7 (d)Bot Score>=0.8

Figure 5.6: Distribution of the users with varying bot scores ranging from 0.6-0.8.

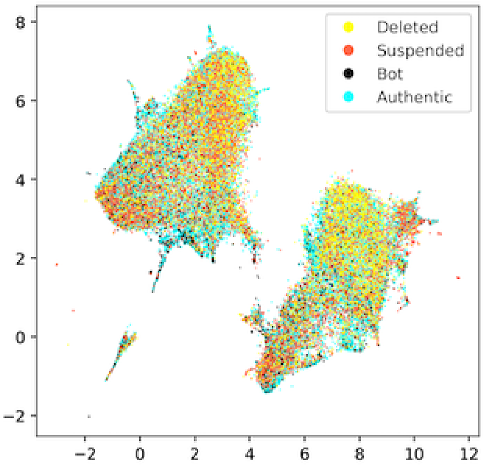


Figure 5.7: Presence of 4 set of users in the cluster.

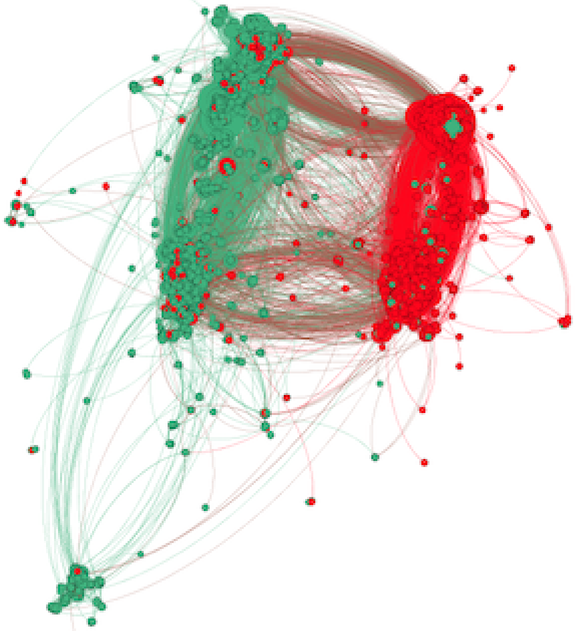


Figure 5.8: Overall follower-followee network of the protesters and counter-protesters. Protesters are represented by green color while counter-protesters by red color.

ipants in the online debate constrain the edges in the graph G. The graph G contains 21,495,449 edges, and 226,412 vertices. We found 33,278 connected components in the network. The largest strongly connected component contains 192,903 users, with 89,377 protesters and 103,526 counter-protesters. Since a strongly connected component in a directed graph is its maximal strongly connected sub-graphs, the presence of both protesters and counter-protesters in the largest strongly connected sub-graph indicates the path between the protesters and counter-protesters. The betweenness centrality of the graph G is $9.80e^{-06}$ (SD $1.388e^{-07}$), which indicates how much a node appears in the shortest path between two nodes. Since the network has very low betweenness centrality, this implies that the users in the network do not occur in many shortest paths in the follow network. The average eigenvector centrality for the network is 0.00056 (SD $4.25e^{-06}$), which shows that the users in the network are connected to influential neighbours, i.e., user-nodes which themselves have high eigenvector centrality (or high in-degree). The network density is 0.0004 indicating a sparse follow network. Figure 5.8 shows the follower-followee graph of 10,000 random users selected from 263,869 users. We experimented with different random samples of 10,000 users to check for consistency in network structure and observed a similar structure across various random sampled networks. In Figure 5.8, we observed two distinct clusters of follow network, clearly showing homophily among the users. Analysis of the graph G shows that the CAA debate on Twitter was conducted by campaigners who were connected to both sides of the debate; were not strongly connected among each other, forming a sparse network; were connected to many influential users on the platform.

Table 5.8: Network descriptive statistics for the authentic and bot accounts who participated in the discourse. $^{*}p < 0.05,^{**}p < 0.01,^{***}p < 0.001$ analyzed using unpaired Mann–Whitney U test. SD stands for Standard Deviation.

	Authentic Users		Inauthentic Users (Bots)		
Metric	Mean	SD	Mean	SD	p
Number of Followers	22.91	43.84	27.57	46.49	$***$ ($5.5e^{-32}$)
Number of Followees	22.43	61.00	29.70	72.50	$***$ ($9.07e^{-09}$)
Eigenvector Centrality	0.002	0.006	0.003	0.007	$***$ ($2.55e^{-26}$)
Betweeness Centrality	0.00011	0.0004	0.0001	0.00038	$**$ (0.01)

Follow Network For Authentic And Inauthentic Users: In order to gauge the presence of inauthentic users, we construct a graph H from a set of authentic and inauthentic users (bot scores ($>= 0.8$)).

We study the authentic and inauthentic users in the graph H and discuss the network descriptive statistics of authentic and inauthentic users. Table 5.8 shows the difference between authentic and inauthentic users with respect to various network descriptive statistics measures. We see there is a very significant difference between the followers and followees of the authentic and inauthentic users. Inauthentic users tend to have higher followers and followee than their authentic counterparts. Eigenvector centrality shows a significant difference among authentic and inauthentic users, with the bot being prominent in both measures. As a result, inauthentic users are more reachable than authentic users and have a stronger influence in the network as compared to authentic users.

5.7 Discussion

This work focuses on characterizing the discourse surrounding the Citizenship Amendment Act (CAA) on Twitter, considering the involvement of both authentic and inauthentic users. Our goal is to understand the participants' stances using unsupervised learning in a multilingual context and to identify major topics within the discourse from the perspectives of both protesters and counter-protesters. Additionally, we examine the presence and perception of various authentic and inauthentic actors in this discourse, specifically focusing on bots, suspended users, and deleted users as inauthentic actors. Users who were not categorized as inauthentic are considered authentic users.

To conduct our analysis, we collected a dataset of 9 million tweets related to the CAA using trending hashtags in India. Through our findings, we discovered the presence of inauthentic activities on both sides of the discourse. However, counter-protesters exhibited a higher level of inauthentic activity compared to the protesters. By examining the frequency of tweets over time, we observed that much of the discussion was driven by inauthentic users, who tended to post less emotional content compared to their authentic counterparts.

Regarding the content shared by authentic users, both protesters and counter-protesters predominantly focused on topics such as violence and protest. In contrast, inauthentic users strategically shared more appealing content to garner attention. Analyzing the follower network of the participants revealed the presence of homophily, where users with similar stances tended to follow each other. Furthermore, one of the largest connected components in the follower network suggested a pathway between authentic and inauthentic users, indicating the potential reachability of inauthentic users to their authentic counterparts.

This work holds significant importance as it sheds light on the dynamics of online discourse surrounding a contentious issue like the CAA. By distinguishing between authentic and inauthentic actors, we provide insights into the manipulation attempts and the presence of coordinated activities within the discourse. These findings emphasize the need for critical evaluation and awareness among social media users to discern authentic voices from inauthentic ones. Furthermore, understanding the major topics and the strategies employed by different actors in the discourse can help in developing more effective countermeasures against misinformation, polarization, and online manipulation.

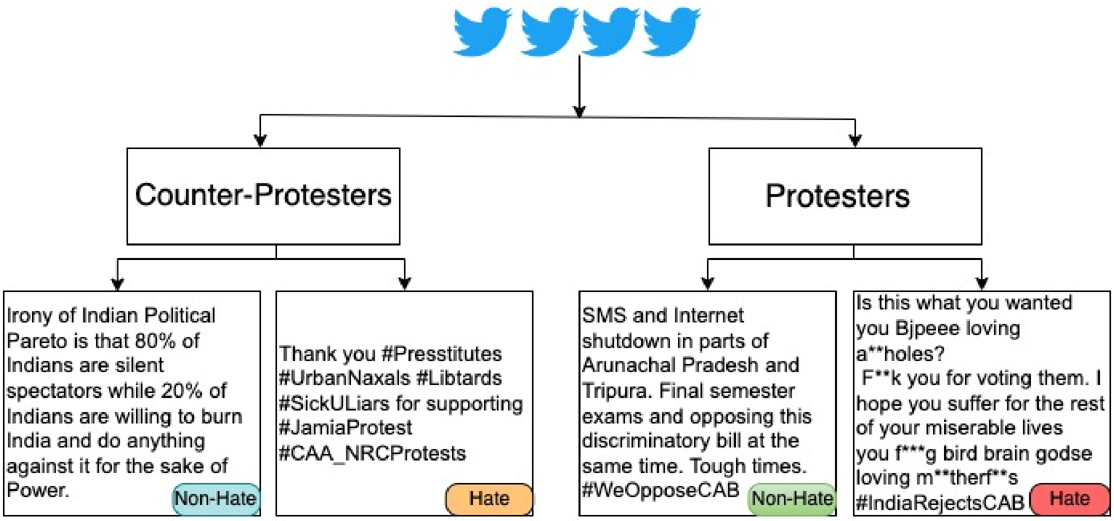
CHAPTER 6

UNDERSTANDING HARMFUL BEHAVIOR: HATE SPEECH DURING PROTEST

This chapter delves into the third primary goal of our thesis, which is to comprehend and identify detrimental users on the platform during protests. Within social media platforms, harmful behavior can encompass a wide range of activities, including the dissemination of hateful messages [169], the propagation of propaganda [95], the spread of misinformation and disinformation, and the coordinated distribution of information [149]. Understanding the various forms of harmful behavior during protests poses a significant challenge.

In this chapter, we specifically focus on hate as a potentially detrimental behavior during online protests and conduct an in-depth study on this aspect. Hateful content deliberated during the protest might shift the focus of discourse and induce a social divide. In this work, we study how hateful users exploited the elements of protest mobilization (i.e., *resources* [1] and *ability to use them*) during the divergent discourse on #CitizenshipAmendmentAct in India. Since the user's stance plays a vital role in hateful tweet detection, we build a multi-task classification model with hate speech detection as the primary task and stance detection as an auxiliary task. Our model outperforms previous models catered towards Indian tweets, with an F1-score of 0.92. We use our model to analyze the hateful users and tweets during the protest mobilization. Our key findings suggest that more hateful users produced more tweets and received faster retweets during the protest than non-hateful users. Across the opposing stances, hateful users held a more central position in the retweet network. However, hateful users who supported the bill showed more initiative through tweets/retweets to counter the protest. This work enhances understanding of a social media protest's vulnerability to hate. Our investigations provide new and nuanced insights into harmful online activities during #CitizenshipAmendmentAct protest from opposing stances and hold importance for designing offline informed interventions as well as online content moderation.

[1] We define the resources as the engagement methods on the Twitter platform (i.e., tweet, retweet, etc.)



Irony of Indian Political Pareto is that 80% of Indians are silent spectators while 20% of Indians are willing to burn India and do anything against it for the sake of Power.

Thank you #Presstitutes #UrbanNaxals #Libtards #SickULiars for supporting #JamiaProtest #CAA_NRCProtests

SMS and Internet shutdown in parts of Arunachal Pradesh and Tripura. Final semester exams and opposing this discriminatory bill at the same time. Tough times. #WeOpposeCAB

Is this what you wanted you Bjpeee loving a**holes? F**k you for voting them. I hope you suffer for the rest of your miserable lives you f***g bird brain godse loving m**therf**s #IndiaRejectsCAB

Figure 6.1: Figure showing non-hate and hate tweets in **CP** and **P** tweets, respectively. **CP** includes tweets in favor of #CAA, while **P** encompasses tweets against #CAA.

6.1 Introduction

As a protest commence, Twitter enables users to build collective narratives [208], gather supporters [166], express opinions [49], leading the way for an impactful mobilization. However, the co-existence of toxic users online can lead to targeted hate being deliberated during the protest [169]. This may cause ripples in the peaceful fabric of the society [55]. We define hate speech in the protest based on the previous literature as "any content that promotes violence against the opposing stance cohort, directly or indirectly threatening the people based on their race, ethnicity, national origin, religious affiliation, political ideology, and political affiliation" [174]. Often hate speech tends to be subjective and based on historical and temporal context. Hence, tracking the propagation of hate during protests becomes a challenging task [174]. Previous research has shown that hate speech is inevitable during an online discourse [80; 212]. Although online social media like Facebook and Twitter have significantly tackled hate speech detection, there is a need for generalized hate-speech detection for low-resource languages, which caters towards the subjectivity of the hate speech of these languages [41]. *Hate speech subsists at a convoluted intersection of freedom of expression, individual, group, and minority rights, along with concepts of liberty, dignity, and equality* [68]. However, the nature of hate-speech changes from context to context, and one definition of hate might not account for all cases, leading toward the subject of abuse [169]. Hate speech during a political discourse might not fit into one of the cases of extreme content. On the contrary, it might be an amalgamation of the discourse's composition, rooting back to the cause of the discourse in the first place. In a discourse setting, observing how the cohorts respond to a common issue is critical [71]. Some of the early methods to detect hate speech include dictionary-based approach [82], bag-of-word approach [37], and feature-based approach [170]. The more recent approaches include using deep learning architectures [25; 224]. Recently, multi-lingual models have enabled multi-lingual hate-speech detection [13; 169]. While much work on hate speech has been done on the text level, user-level detection is still nascent [155; 129; 128; 54]. While hate speech on social media is rising in general [129], the study of political discourse reveals that party affiliation, gender, and ethnicity as reasons for individuals resorting to posting hate speech for political leaders [186]. The policies induced by the government have also been found to show discourse in public [169; 192; 212]. Understanding the user's viewpoint in discourse involves *stance detection*, which aims to infer the author's viewpoint depending upon linguistic cues, the author's identity, and social interactions [11]. Stance detection in social media has been performed on statement level [136; 52] or user level [10; 122; 123; 53]. Previous research has proposed embedding-based user-level stance detection on specific targets [159]. The psycholinguistic analysis of target-specific hate towards a group shows more religious context than directed hate towards a person [61]. Target-oriented user-level hate detection has also been explored for COVID-19 pandemic [88; 16].

We focus on the target-oriented hate speech spread during a divergent discourse in India. The detection of hate speech in India in multilingual settings has shown the best results with LASER embedding and Logistic Regression [13]. Recently, multi-task learning has improved performance on various NLP tasks [185; 215], where performance is sensitive to the task at hand [134]. In context with Indian languages, multi-task learning has shown promising results in joint modeling on sentiment with cyberbullying detection [125], bail prediction of Hindi legal corpus [107], and stance detection [171].

To study the spread of hate during #CAA, we use the dataset and unsupervised user's stance detection as described in Chapter 5. Taking cues from previous research on stance-aware hate speech detection during a discourse [80], we further performed hate speech detection on all the tweets by the user. We use the identified stance of users (users who support #CAA: Counter-Protesters (**CP**) and those who oppose #CAA: Protesters (**P**)) to randomly sample 2,000 tweets proportionately from both sides and

Table 6.1: Table showing the statistics of the users and their respective tweets on **CP** and **P** sides, respectively.

Stance	**User Count**	**Tweet count**
CP	74,829	1,717,091
P	53,853	1,050,177
Total	128,682	2,767,268

manually label them for the presence of hate speech. Examining the array of political stances alongside hate speech provides valuable insights into attitudes and hostility towards a predetermined target [54]. Figure 6.1 shows the example of hate and non-hate tweets during the #CAA protest for opposing stances. Through the users' identified stances and hateful tweets, we built a multi-task classification model for hate speech detection that outperformed the previous baselines. We use the model further to classify all the stance-aware tweets during the mobilization. Using a clustering-based approach on the propensity to produce hate, we further divide the users from low to high hate intensity. Next, we perform a fine-grained analysis of the users and their tweets to analyze the hateful content shared by the two sides during the protest. More precisely, to perform the fine-grained analysis, we address the following Research Questions concerning opposing stances:

RQ1: How can we characterize the spread of hate speech during the protest?

RQ2: What is the role of users with varying hate intensities in spreading hate?

RQ3: How did the community perceive the hate during protest mobilization?

Our work takes account of the temporal spread of hate speech during the protest in a less explored country, India. This study is the first to use a multi-task framework for stance and hate-speech detection during protest mobilization. We focus on identifying hate speech in the context of the protest and not any ethnic/religious groups alone.

6.2 Data & Methodology

6.2.1 Data

To understand the mobilization of hate speech during the #CAA protest, we collect 11,350,276 tweets from 931,175 users during the starting 3 months (December 07, 2019 to February 27, 2020) of the protest. Chapter 5 describes the process of data collected and relevant pre-processing steps followed.

Table 6.2: Table showing the statistics of the users and their respective tweets on **CP** and **P** sides, respectively.

Stance	User Count	Tweet count
CP	74,829	1,717,091
P	53,853	1,050,177
Total	128,682	2,767,268

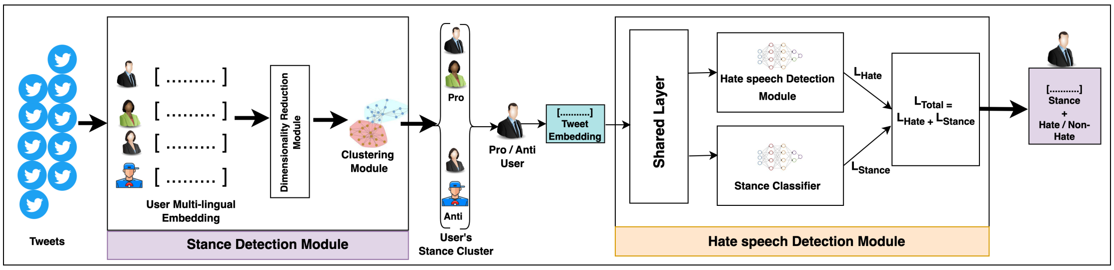

Tweets
[.........]
[.........]
[.........]
[.........]
User Multi-lingual Embedding
Dimensionality Reduction Module
Clustering Module
Stance Detection Module
Pro
Anti
User's Stance Cluster
Pro / Anti User
[..........] Tweet Embedding
Shared Layer
Hate speech Detection Module
Stance Classifier
L_{Hate}
L_{Stance}
$L_{Total} = L_{Hate} + L_{Stance}$
[..........] Stance + Hate / Non-Hate
Hate speech Detection Module

Figure 6.2: Figure showing our proposed framework to detect user-level stance followed by tweet-level hate speech detection for stance-aware users.

6.2.2 Methodology

We present our approach in Figure 6.2. Our approach consists of two building blocks: (i) a user-level stance detection module and (ii) a tweet-level hateful content detection module.

User-Level Stance Detection

We build upon the unsupervised stance detection algorithm proposed in recent literature [159; 141] to identify the user's stance. First, we identify hashtags in our dataset as counter-protest and protest hashtags. Our dataset had 27 counter-protest hashtags and 48 protest hashtags. The detection of the user's stance is carried out in 6 iterative steps: (i) identify the seed users whose tweets only contain hashtags from either **CP** or **P** side, i.e., hashtag-based labeling (ii) include users who retweet the users identified in *step (i)* at least k-times (k = 15), i.e., label propagation, (iii) create 1024-dimensional user embedding through LASER (Language-Agnostic Sentence Representations)[2] by taking an average of the vector of the filtered tweets, (iv) project users on a 2-dimensional space

[2] https://github.com/facebookresearch/LASER

using Uniform Manifold Approximation and Projection (UMAP) algorithm [133], (v) cluster the 2-dimensional embedding using hierarchical density-based clustering (HDBSCAN) [132], (vi) check the purity of cluster based on users identified through seed set and label propagation, to detect stance of the clusters. Step (i) yielded 106,605 **CP** users and 79,493 **P** users. The HDBSCAN algorithm yielded 5 clusters with 270,889 users. We consider clusters pure if they contain at least 30% of labeled users obtained via label propagation and show at least 80% purity of labels. We found 4 such clusters for #CAA, where two clusters belonged to **CP** (142,839 users), and the other two belonged to **P** (121,030 users).

Final Dataset

After identifying a substantial number of users for **CP** and **P** in the discourse, to ensure the richness of the conversations, we considered tweets that showed 1,000 or more occurrences (either original tweets with the exact text or retweets, combined) in our analysis. With the tweets dropping less than 1,000 occurrences, we performed a second iteration of user filtration for users with less than 5 tweets in the dataset. The final data statistics we worked with included 128,682 users who accounted for 2,767,268 tweets in our dataset. Table 6.2 shows the distribution of the final dataset that we use in our analysis.

The hateful content shared during a protest can be onerous [129], with a mix of language and cultural diversity [169]. We manually annotated 2,000 tweets (1,000 from each **CP** and **P**). The annotation of the hate speech in tweets was done by 2 groups consisting of 4 annotators. The annotators constituted a research scholar and 3 undergraduate students studying at the university. Due to the hate and derogatory content in the tweet, annotators were advised to take breaks between annotations to account for sound mental health. We used Cohen's Kappa [21] for calculating the inter-annotation agreement among the two groups of annotators. We calculated Cohen's Kappa and found an inter-annotation agreement of 0.91, showing a strong agreement between the annotators. We resolved the differences in the remaining tweet annotations. Next, we develop the multi-task classifier using the annotated data to identify whether the rest of the tweets are hateful. The final annotated data consisted of 131 hate and 869 non-hate tweets from **CP** and 114 hate and 886 non-hate tweets from **P**. Thus the annotated sample of 2,000 tweets contained 245 hate tweets and 1,755 non-hate tweets.

Table 6.3: Table showing the performance of various baselines and the proposed Multi-task learning framework for hate speech detection in Tweets.

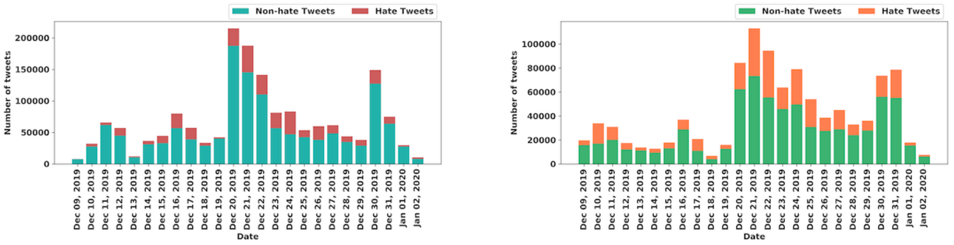
Model	Accuracy	F1-score (weighted)	Precision	Recall
LASER+LogR	0.64	0.64	0.65	0.64
LASER+LSTM	0.73	0.73	0.73	0.73
mBERT + LogR	0.80	0.80	0.81	0.80
mBERT+MTL	0.93	0.93	0.92	0.93
LASER+MTL	**0.95**	**0.93**	**0.93**	**0.95**

Hateful User Detection

We use the annotated data to build the classifier for hate speech detection. We also explore various strategies from previous literature and compare them with our proposed framework.

Baselines: Being aware of the challenges of social media corpus, such as multilingual, unstructured, and noisy, we experiment with previously proposed hate prediction pipelines as our baseline. For the baselines, we use the hyperparameters as presented in the respective papers. For the first baseline, we implemented LASER with Logistic Regression [14] and obtained an F1-score of 0.64. We use the model proposed by [169] for the second baseline that uses LASER with LSTM and obtained an F1-score of 0.73. For our third baseline, we use mBERT, a version of BERT trained on a multilingual corpus, to create sentence embeddings of length 128. The representation is then passed through Logistic Regression Layer. The F1-score for mBERT+LR is 0.80. Previous literature has found that multilingual hate speech detection for low-resource languages performs best with LASER embeddings [14]. Hence we use LASER to form tweet embedding and F1-score as the metric to gauge the model performance, catering to the imbalanced data set.

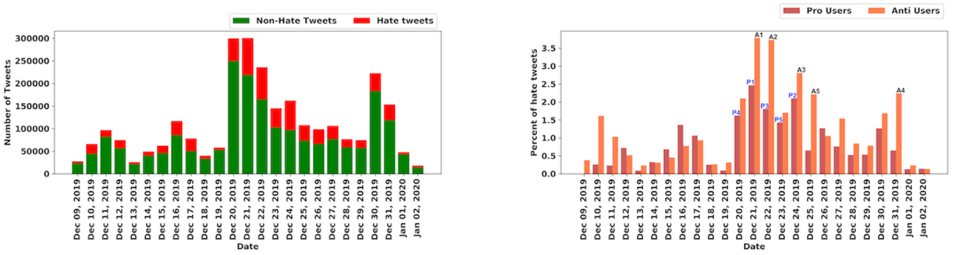
Multi-task Hate speech classifier Model: Multi-task models have performed well in various classification tasks in the recent past [125; 215; 171]. Hence, leveraging the initial annotated dataset, we built a Multi-Task Learning (MTL) framework for hate speech detection. In our MTL model, hate prediction is the main task, whereas we keep stance prediction as an auxiliary task. The intuition behind using stance as an auxiliary prediction task is that it will help the model capture the tweet's context better



(a)

(b)

Figure 6.3: Distribution of hate tweets during the discourse. Figure (a) shows the tweet timeline for **CP** tweets. Figure (b) shows the tweet timeline for **P** tweets.



(a)

(b)

Figure 6.4: Figure (a) shows the total tweet timeline divided into hate and non-hate tweets. Figure (b) shows the distribution of percentage hate tweets during the CAA discourse for **CP** and **P**, respectively.

and guide it toward stance-aware hate speech detection. Another advantage of using the MTL model is that it reduces the risk of overfitting when the data is imbalanced. Hence, we employ hard parameter sharing in our MTL model, which reduces the risk of overfitting [27; 168]. Input to the model is a single tweet: t_i. Contextualized representation is generated for each tweet using a multilingual sentence encoder [163]: h_i. The sentence embedding is fed into a standard single-layer transformer architecture (shared transformer). For both hate and stance predictions, classification heads (fully connected MLP layers) are placed on top of the transformer embedding to get their respective classifications. We use standard cross entropy loss (L_{hate} and L_{stance} respectively) for training. Both the losses are equally weighted; hence, the total loss comes out as shown in Equation 7.1.

$$L_{total} = L_{hate} + L_{stance} \tag{6.1}$$

6.3 Spread of Hate Content

We use our proposed model to classify all tweets under study as hateful or non-hateful to conduct an in-depth analysis of the protest, concerning the opposing stances. Our first set of analyses focused on the spread of hateful content (i.e., *ability to use the resources in Protest mobilization*). With the help of a fine-grained analysis of tweets shared during the protest, we inspect and compare the spread of hateful content on each side of the divergent discourse.

Figure 6.3 shows the stacked plot of the hate and non-hate tweets shared during the study period. Table 6.4 shows the 4 most hateful days in terms of frequency of hateful tweets being shared for **CP** and **P**, respectively. We found that the day of the highest frequency of tweets in our dataset, i.e., December 21, 2019 is also the day with the highest number of the hateful tweet for both **CP** and **P**. The **CP** had 42,279 hateful tweets on the day, while **P** witnessed 39,764 hateful tweets. The second most hateful tweet dissemination for **CP** was on December 24, 2019, with 36,129 tweets, while the second most number of hateful tweets produced by **P** was on December 22, 2019, with 39,148 tweets. Figure 6.4 compares the percentage of hate tweets generated per day under **CP** and **P** tweets, respectively. On comparing **CP** percent hate tweets to that of **P** hate tweets, **P** posted more hate tweets during the discourse. These results suggest that more tweets were generated from **CP** users. However, the **P** tweets produced more percent hate daily during the protest.

Table 6.4: Table showing the date and corresponding hate produced by the **CP** and **P** content, respectively, in descending order from December 2019.

Date	CP	Date	P
Dec 21	42,279 (2.46%)	Dec 21	39,764 (3.78%)
Dec 24	36,129 (2.10%)	Dec 22	39,148 (3.72%)
Dec 22	31,022 (1.80%)	Dec 24	29,491 (2.80%)
Dec 20	27,876 (1.62%)	Dec 31	23,519 (2.23%)

Table 6.5 shows the topics the **CP** and **P** hate tweets contained during the highest hateful tweets frequency days. The LDA topics common in hate tweets of both sides included topics on muslims, imprisonment, and traitors, as shown in Table 6.5. December 21, 2019 recorded the most hateful tweets from both **CP** and **P** tweets. The **CP** hateful tweets contained topics of the destruction of public property and violence for

Table 6.5: Topics discussed during the days with the most hateful tweets spread in the dataset.

Date	CP	P
Dec 20	Location, Terror, Muslim, Revolution, Religion, Violence, Propaganda	Killing, Location, Religion, Student, Muslim, Propaganda, Violence
Dec 21	Location, Killed, Warning, Name-calling, Hindu, Demonstration	Muslim, Violence, Location, Name-calling, Demonstration
Dec 22	Demonstration, Attack, Police, Religion, Women, Politicians, Attack	Demonstration, Location, Name-calling, Muslims, Slogans
Dec 24	Jihad, Violence, Demonstration, Religion, Name-calling, Student, Brutality, Muslim	Death, Arrest, Minister, Demonstration, Student, Brutality, Location
Dec 31	Sadhguru, Secularism, Demonstration, Slogan, Minister, Name-calling	Sadhguru, Demonstration, Trolls, Respect, Fraud, Propaganda, Name-calling, Slogans

Table 6.6: Users divided into Hate intensity for Counter-Potesters and Protesters along with their respective tweets.

Cluster centers	Score range	CP		P	
		Users	Tweets	Users	Tweets
2.79	0-8	63,434	843,935	43,355	460,951
14.29	9-27	9,967	618,420	8,812	375,071
40.44	28-123	1,428	254,736	1,686	214,155

December 21, 2019, while **P** hateful tweet topics included genocide, propaganda, and the wrongdoings of the media. December 24 was the second most hateful day for **CP** and the third most hateful for **P**. On December 24, the **CP** hate topics included Jihad, Violence, Religion, and brutality. The **P** topics contained major topics of death, Arrest, and Ministers/politicians. Location topics on both sides included the places of protests, rallies, and violence. The name-calling included terms like tukde-tukde gang [3], cheap, bhakts (devotee), gaddar (traitor), pseudobhakts, Nazi, Ma*bhakt (slang with devotee), Modia (Media in favor of **CP**). The slogans included Kagaz nhi dikhayenge (will not show papers), zindabad (a cheer), illallah (god is one).

Discussion: We investigate the real-world incidents through the lens of news that erupted during the protest, coinciding with the most hateful days for **CP** and **P** tweets. In Figure 6.4, P4 shows the 4^{th}-most hateful day in **CP**. On this day, news of injuries to police erupted in different parts of India, including Gujrat, Maharashtra, Delhi, and

[3] https://www.bingedaily.in/article/who-is-the-tukde-tukde-gang-and-where-did-they-come-from

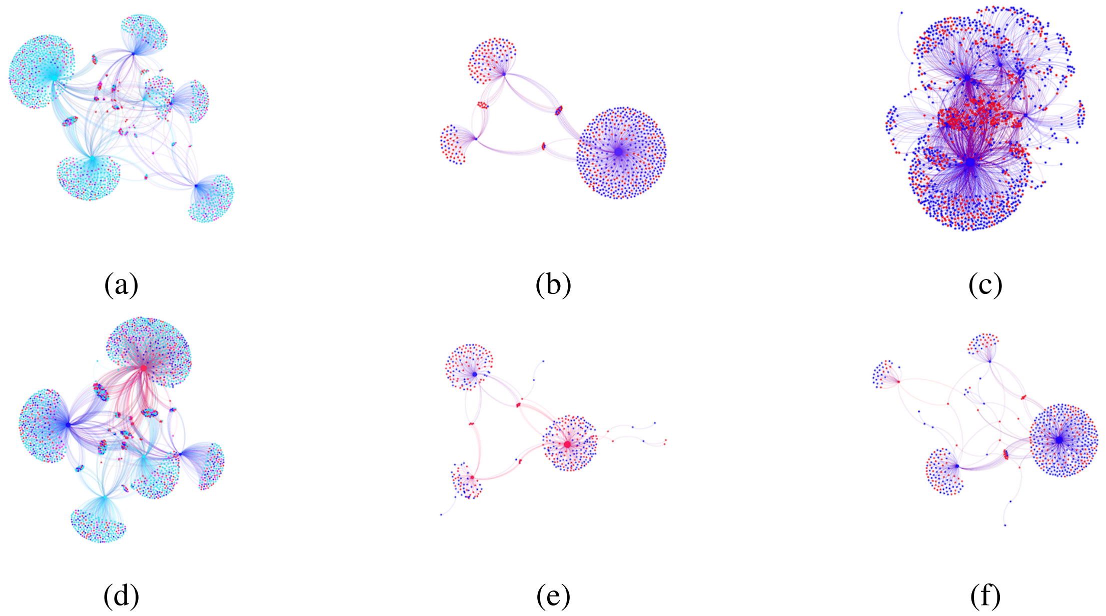


Figure 6.5: Figure showing Retweet network of users' interaction with different hate intensities for **CP** and **P** users. The size of the nodes corresponds to retweet frequency. Figure a-c presents the **CP** retweet network. Figure d-f presents the **P** retweet network. Color scheme: Dark blue: old low-intensity users, Light-blue: new low-intensity users, Red: old mid and high hate-intensity users, and Pink: new mid and high-intensity users. We find that the number of hateful users increased towards the most hateful day reported in tweets.

Uttar Pradesh.[4] The next significant spike is P1 and A1, which is the most hateful day for both **CP** and **P**. P1 and A1 coincide with the news of peaceful protests against CAA in Delhi and Bhopal.[5] Reports of violent protests led to Section 144 being introduced in different parts of the country.[6] On the peaks P3 and A2 (December 22, 2019), rally for **CP** started in Mumbai.[7] Violent protests with arrests were reported in different parts of the country.[8] P5 (December 23, 2019) predominantly witnessed **CP** rallies in the country.[9] With the P2 and A3 peaks, news of death due to bullet injury was new development.[10]

[4] https://www.freepressjournal.in/mumbai/**P**-protest-3-st-buses-stoned-in-hingoli

[5] https://www.freepressjournal.in/india/caa-protests-at-jama-masjid-peaceful

[6] https://www.freepressjournal.in/india/caa-protests-after-violent-protests-section-144-imposed-in-gorakhpur

[7] https://www.freepressjournal.in/mumbai/mumbai-hundreds-come-together-in-support-of-caa

[8] https://www.freepressjournal.in/india/65-arrested-350-booked-for-violence-during-caa-protests-in-ghaziabad

[9] https://www.freepressjournal.in/india/bjp-launches-social-media-campaign-to-reach-out-to-minorities-regarding-caa-nrc

[10] https://www.freepressjournal.in/india/caa-protest-14-of-16-upprotesters-died-of-bullet-injury

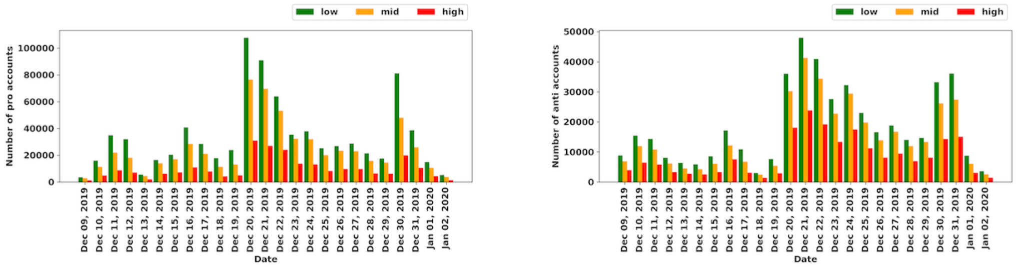
Table 6.7: Table showing the FRT for low, mid, and high hate-intensity users in **CP** and **P**, respectively. The time is reported in (Hour:Minute:Second) format. $^{*}p < 0.05,^{**}p < 0.01,^{***}p < 0.001$, – not significant, analyzed using unpaired Mann–Whitney U test.

FRT (H:M:S)	**CP**	**P**	*p*
Low	$01:36:27$	$01:28:49$	*(0.03)
Mid	$0:59:07$	$01:33:10$	–(0.33)
High	$03:03:00$	$02:02:58$	–(0.07)

6.4 Hate Content Spreaders

To propose that a given user is hateful is a much more difficult task, as one user can post multiple tweets during the time, which may or may not be hateful. Hence, to perform a fine-grained analysis of the hateful users, it is essential to identify users based on their propensity to be hateful [165]. We divide the users in our dataset based on how much hateful content they have posted. While the lowest number of hateful tweets can be 0, the highest number of hateful tweets a user posted in our dataset was 123. We divided the users from least to most hateful based on the threshold values using k-means [105] as a clustering algorithm on the count of hate tweets by users. The k-means algorithm initially selects k points in space as an initial guess for centroid hate score, followed by assigning all the remaining points to the nearest centroid. The procedure is reiterated until no points switch clusters or all iterations are completed. We experimented with different values for k and found the best fit for k as 3, based on the elbow method. The statistics of users divided into the 3 clusters, based on our best k-value, are shown in Table 6.6.

According to the number of hateful tweets the users posted, we categorized them as low, mid, and high hateful users, based on the ranges of hate tweets they posted, i.e., $0-8$, $9-27$, and $28-123$, respectively adopted from k-means clustering. The three tiers allow us to control better users' distribution based on their hate intensity. The median values for **CP** are 2.0, 12.0, and 35.0 for low, mid, and high intensities, respectively. The median values for hate scores in **P** are 3.0, 12.0, and 36.0 for low, mid, and high intensities, respectively. Table 6.6 shows that the distribution of users with low hate intensity is maximum in both **CP** and **P** users. In comparison, the high-intensity hateful users were the least in distribution. Once we have divided the users according to their hate intensity, we perform further analysis to explore how the different sets of users affect the discourse on the platform.



(a)

(b)

Figure 6.6: Figure shows users' activity with different hate intensities on the platform. Figure (a) shows the activity for **CP** users. Figure (b) shows the activity for **P** users.

Figure 6.6 (a) and (b) show the number of hateful user's account participation for **CP** and **P** users, respectively. On any given day, low-hate-intensity users produced the highest frequency of tweets on both sides. The number of most-hateful users peaked on December 20 for **CP**, while for **P**, the most-hateful users peaked on December 21. The distribution of high-hate intensity users in both **CP** and **P** users suggests that hateful users slowly cluster around the mean, after which they show declining trends. This result contrasts with the analysis of users with different hate intensities in other more free speech-advocating social media platforms such as Gab, where the number of hateful users shows a steadily increasing trend in general [129] suggesting that although hate on social media platforms is generally increasing, hate users relevant to a targeted protest is synchronized with the protest, as it shows a mid-peak with lower start and end tendency [49]. For **CP** users, we found that the average tweets produced in descending order are by mid-hate users (0.069), high-hate users (0.060), and low-hate users (0.046), respectively. At the same time, retweets in **CP** users produced with the average in descending order are high-hate (178.32), mid-hate (61.97), and low-hate (13.25), respectively. On the **P** side, the average number of tweets made by mid-hate users (0.046) is more than high-hate users (0.040), followed by low-hate users (0.029). The average number of retweets in **P** users in descending order are high-hate(126.97), followed by mid-hate (42.51) and low-hate (10.60) users, respectively. This result suggests that the more hateful users (mid-hate and low-hate combined) produced more tweets during the protest.

6.5 Hate Content in Protest Mobilization

To study the mobilization of hateful content during the #CAA protest, we explore the tweet-retweet interactions on the platform by the users with different hate intensities. We first observe each user's average tweets and retweets with different hate intensities during the protest. For **CP** users, we found that the average tweets made by mid-hate users (0.069) and high-hate users (0.060) were more than low-hate users (0.046). Similarly, on the **P** side, mid-hate users (0.046) and high-hate users (0.040) made more tweets on average than low-hate users (0.029). This result suggests that the more hateful users also produced more tweets during the protest. Among the hateful users, the mid-hate users were more actively tweeting than the high-hate users during the protest. Regarding retweets, the average retweets done by low-hate **CP** (13.25) and **P** (10.60) users were the lowest, followed by that of mid-hate **CP** (61.97) and **P** (42.51) users. The high-hate users for **CP** (178.32) and **P** (126.97) produced the maximum retweets in the dataset. Although the high-hate users produced maximum retweets during the protest, the retweets of hateful content reveal that mid-hate users made maximum retweets (38.74%) to the hateful tweets in the overall protest, followed by high-hate users (18.18%).

Next, to understand how the tweets done by different hate-intensity users were perceived (i.e., user engagement), we compute the First Retweet Time (FRT) for each hate-intensity group [129]. FRT is defined for a set of users U $\in U_{low}, U_{mid}, U_{high}$ as given in the equation 6.2.

$$FRT_U = \frac{1}{|U|} \sum_{u \in U} RT_u \tag{6.2}$$

The First Retweet Time (FRT) essentially calculates the average time taken to get the first retweet for a post made by user u. We calculate FRT for the 3 set of hate-intensity users to gauge whether users of certain hate-intensity received faster retweets. On the overall dataset, we observe that the mid-hate users receive a retweet in the least time, i.e., within 1 hour and 12 minutes. The high-hate users received retweets most passively (2 hours 22 minutes) among the three (statistically significant) sets of users. The low-hate users received a retweet in 1 hour and 25 minutes on average. We further analyze the FRT for the 3 set of users for both **CP** and **P** groups.

Table 6.7 compares FRT for **CP** and **P** users with different hate intensities. We find that the low-intensity users for **P** receive retweets faster than **CP** users ($p < 0.05$). However, the comparison of FRT for mid and high-hate **CP** and **P** users is not statistically significant.

Table 6.8: Table showing the Network descriptive statistics for **CP** and **P** users. $^{*}p < 0.05,^{**}p < 0.01,^{***}p < 0.001$, – not significant, analyzed using unpaired Mann–Whitney U test.

Measures	**Intensity**	**CP**	**P**	***p***
Closeness centrality	Low	0.0007	0.0006	–
	Mid	0.0009	0.0008	–
	High	0.001	0.0009	***
Indegree Centrality	Low	$8.08e^{-05}$	6.78^{-05}	***
	Mid	0.0003	0.0002	–
	High	0.0006	0.0005	***
Outdegree Centrality	Low	$8.34e^{-05}$	$8.04e^{-05}$	***
	Mid	$7.16e^{-05}$	$7.20e^{-05}$	***
	High	$4.65e^{-06}$	$3.98e^{-06}$	***

Next, we explore the descriptive network statistics for the **CP** and **P** users to understand information flow patterns during the protest. We compute three descriptive network statistics: closeness centrality, indegree centrality, and outdegree centrality. The outdegree centrality shows the communication power of the user (measured using the number of retweets the user receives), while indegree centrality shows the initiative user takes during the protest (measured through the tweets and retweets a user does). The closeness centrality measures how a given user is close to all the other users in the network. Table 6.8 shows that indegree-centrality is significantly more for **CP** low and high-hate users than **P** users. The out-degree centrality is significantly more for mid-hate, **P** users. The low and high-hate, **CP** had significantly more out-degree centrality.

Above results suggest that low and high-hate **CP** users were more active in taking the initiative, measured through in-degree centrality. The low and high-hate **CP** users also held significantly more vital communication during the protest than **P** low and high-hate users. On average, a larger closeness centrality indicates more central nodes in the network. The closeness centrality measure reveals that **CP** users have more central high-hate-intensity users in the network.

To understand the evolution of the **CP** and **P** users, we further analyze the retweet network of both sides. Due to space limitation, we show the retweet network for 3 days based on spikes in the users (Figure 6.4). We plot the Retweet network on a random 10% sample for the days and visualize the largest connected component for selected days. Figure 6.5 shows the retweet network for the **CP** and **P** users on the respective

Table 6.9: Table showing the Inauthentic behavior (Bots and Suspended Accounts) for a stratified sample of users from CP and P stances. NH-Total: Total Non-Hate users, H-Total: Total hateful users, NH-Bot: Non-Hate bots. NH-Sus: Non-Hate suspended users, H-Bot: Hateful Bots, H-Sus: Hateful suspended users.

	Date	Total Users	NH-Total	H-Total	Sample	NH	NH-Bot	NH-Sus	Hate	H-Bot	H-Sus
CP	Dec 09	6,026	3,148	754 (12.5%)	1198	595	143 (2.4%)	220 (3.6%)	603	442 (7.3%)	75 (1.2%)
	Dec 16	33,858	24,233	1,383 (4.0%)	5,297	4,185	603 (1.7%)	1,553 (4.5%)	1,112	807 (2.3%)	138 (0.4%)
	Dec 20	48,274	37,706	1,389 (2.87%)	7,240	6,124	808 (1.6%)	2,047 (4.2%)	1,116	812 (1.6%)	135 (0.2%)
P	Dec 09	10,070	5,631	1,148 (11.4%)	1,979	1,030	245 (2.4%)	215 (2.13%)	949	656 (6.5%)	74 (0.73%)
	Dec 16	17,726	10,941	1,471 (8.2%)	3,075	1,847	354 (1.9%)	440 (2.4%)	1,228	845 (4.7%)	93 (0.5%)
	Dec 20	27,716	18,830	1,627 (5.8%)	4,055	2,700	337 (1.2%)	665 (2.3%)	1,355	930 (3.3%)	99 (0.3%)

dates. For each day, we color code new low-hate users as dark blue if they were present previously in our dataset (old users) and light blue if the users are newcomers on the day (new users). We combine the low-hate and high-hate users as hate users for better visualization. We color-code old low-hate and high-hate users red, and new hate users are color coded as pink. We build the first network graph for December 09, which is the second day of our data collection. Both **CP** (Figure 6.5(a)) and **P** (Figure 6.5(d)) show a hub-like structure where one user is connected to many nodes, indicating retweet relation between one-to-many users. A key observation for December 09 is that a hateful user holds the center position in the **P** network. The next peak we observe is on December 16. The retweet network for both **CP** (Figure 6.5(b)) and **P** (Figure 6.5(e)) shows that the largest connected component was formed by old users (dark blue color). At the same time, the **P** retweet network had two hateful central nodes. On December 20, the hate users were more prominent for both **CP** and **P**. From the evolution of the network from Dec 09 to Dec 20, we find hateful content increasingly seeping into the retweet network as the hateful content peaked on December 21, 2019.

6.6 Manipulation of Hate Speech

Previous research has shown that bots dominate the discussion during online discourse, indicating discourse manipulation [196]. Hence, we conclude our analysis by comparing the hateful behavior of the protesters and counter-protesters with clear signs of inauthenticity to gauge potential manipulation in the protest. For inauthentic behavior, we focus on Twitter suspension [11] and Bot behavior as proxies. We use Botometer API [219], using the universal scores ≥0.7 to identify bot behavior. The purpose of

[11] https://help.twitter.com/en/managing-your-account/suspended-twitter-accounts

identifying the bots in discourse is to analyze the deliberate sharing of hate speech during the protest. Catering to the rate limitation for Botometer API, we sample 26,632 users in a stratified manner, considering 20% users from both sides (15,677 users in **CP** and 10,955 users in **P**) for checking bot scores and Twitter suspension. We monitored the presence of the inauthentic actors on December 09, December 16, and December 20, 2019, respectively, for both **CP** and **P** sampled users. Table 6.9 shows the percentage of bots and suspended users for the three days under consideration. We found that our sampled dataset's bots and suspended users constituted $<$10% inauthentic actors daily. We also found that inauthentic activity was not necessarily limited to hateful users but was also present in non-hate users on both sides of the protest.

6.7 Discussion

In this work, we study how the two significant elements of protest mobilization (resources and the ability to use them) were exploited to spread hate during the #CAA protest in India. We first divide the users into counter-protesters (CP) and protesters (p) based on an unsupervised stance detection framework in a multi-lingual setting. We further annotate tweets for hateful content and train a multi-task classifier for hate speech detection, with hate detection as the main task and stance detection as an auxiliary task. After we have trained the model, we classify the rest of the tweets for all stance-aware users as hateful or not. We use the above results to understand the spread of hate during the protest from content, user, and network perspectives. Our key findings reveal that most hateful day coincides with the day with highest percentage of tweets present in collected data. On dividing the users based on hate intensity, we find that more hateful users also produced more tweets and retweets during the protest. The mid-hate users, i.e., those who produced 9-27 hate tweets during the protest, made maximum retweets to hateful content and received fast retweets to their content. Among counter-protesters and protesters, low and high-hate counter-protesters exerted more initiative to participate and more communication power. The counter-protesters high-hate users exercised more central positions than their counterparts.

Limitations: The study of social media protests involves several caveats. Hashtags used in the data collection might not cover the complete picture of the discussion.

The dependence on a single media platform and public APIs is another limitation of the work [194; 35]. Although Twitter makes data publicly available, it is intrinsically sensitive.

CHAPTER 7

UNDERSTANDING HARMFUL BEHAVIOR: COORDINATED BEHAVIOR DURING PROTEST

So far, we have focused on the strategies used by protesters to conduct a protest, the narratives shared by the participants, and the presence of inauthentic behavior in the protest. We also focused on the harmful behavior during the protest in the form of hate speech. However, harmful behavior on social media has most recently been seen as a multi-faceted problem, where multiple forms of harmful behavior might be in interplay. For example, Hristakieva *et al.* [93] focused on the spread of propaganda by coordinated communities. By conducting a combined analysis of propaganda and coordination, valuable evidence regarding the destructive nature of coordinated communities was revealed, which would not be accessible through separate investigations.

Different narratives shared during protests, often coordinated, play a strategic role in shaping collective opinions, making it essential to decipher them. As users engage in online protests by sharing collective narratives, they may also become susceptible to various harmful and malicious influence operations under the hood of protest participation, including coordinated harmful behavior. This problem becomes more aggravated when sharing harmful content is done in a coordinated mechanism involving genuine and inauthentic accounts (bots, suspended users) to enable larger outreach.

In this chapter, we examine the narratives shared by coordinated communities over the enactment of the Citizenship Amendment Act (#CAA) by the Indian Government. We first examine the narratives shared by coordinated communities of opposing stances, i.e., protesters (who opposed the Act) and counter-protesters (who supported it) on Twitter. Next, we analyze different narrative-sharing coordinated communities from opposing stances concerning various inauthentic behavior ranging from user suspension, the presence of bots, and the presence of hateful content. Our analysis indicates that the most hateful, strongly coordinated community of counter-protesters (showing skepticism and grievances in tweets) and protesters (showing questioning and grievances in tweets) also showed a high degree of inauthentic behavior (i.e., bots and suspended users). We also identified strongly connected communities spreading hate during the protest, with a low degree of inauthentic behavior. To summarize, our study offers unique insights into the harmful nature of coordinated communities that were previously not investigated.

7.1 Introduction

Collective narratives shared during the protest include evolving verbal, graphic, or written interpretations of related events in a given duration [158]. Chapter 4 explains the presence of various narratives in a social media-mediated protest such as personal grievance, call for action, and reporting of on-ground activities [203], while contentious topics have shown additional questioning and skepticism narratives. Apart from posing individual risks such as hate content, malicious users engage in coordinated actions to amplify the dissemination and reach of manipulation [146].

This chapter focuses on the coordinated social media efforts by the opposing stances (**CP** and **P**) during the discourse around Citizenship Amendment Act (#CAA) in India, considering authentic, inauthentic, and hateful actors. We define coordination as an exceptional similarity between a group of users in terms of retweets, hashtag usage, and mentions [146]. The protest participants may coordinate their way into sharing different collective narratives during the protest [208]. We first identify the stance of different users in the #CAA discourse using the unsupervised user-based stance detection technique described in Chapter 5. Next, we build a user-similarity network using some of the most significant coordination mechanisms for online protest (i.e., hashtags, mentions, and retweets). Since the coordination of different narratives created by users may help gain an elevated online reach, our first research question is:

RQ1. What protest-centric narratives were shared by the coordinated communities during the discourse? We use the unsupervised narrative detection technique proposed by [140] to identify the narratives shared by the communities obtained from the user similarity network. Different communities obtained by opposing stances incorporate different agendas and exert different levels of coordination that we aim to decipher. Our major finding suggests skepticism and grievances were the two most dominant narratives shared by the largest communities on both sides. While the next highest community in **CP** and **P** shared tweets about questioning and reported on-ground activities. However, the extent is not clear. Although we can gauge the presence of coordination among different narratives in coordinated communities, the extent or pattern of coordination for building these narratives is unclear. Hence, we attack this as our second research question:

RQ2. What is the extent of coordination exerted by different protest-centric narratives-based communities? We derive the network's multi-scale backbone by retaining statistically relevant links and nodes. Through an iterative process, we detect communities among increasingly coordinated user subsets. Our approach avoids fixed thresholds and allows for studying coordination across the spectrum, from weak to strong. Through different network measures, we analyze the characteristics property of the strongly coordinated communities and their narrative focus. It is to be noted that the division of users into opposing stances(**CP** and **P**) also paves the way for derogatory and hateful comments [169] from authentic and inauthentic users, where users on either side may produce hate towards the opposing side [169]. However, identifying inauthentic coordinated communities during an online protest becomes challenging since online protests are inherently coordinated [190]. The multi-faceted vulnerability of users on the platform during protests brings us to our third research question:

RQ3. What harmful behavior was exerted by the most coordinated communities during the discourse? Researchers have raised concerns about the risk of mass manipulation of public opinion through disinformation campaign [34; 24; 153] and social bots [197; 40; 119]. Here, we focus on spreading hate as potentially harmful behavior during the protest. Since hate can be spread by authentic and inauthentic users (bots), we perform a fine-grained analysis on the production of hate by different levels of coordinated communities and decipher how hateful communities are different from inauthentic activities. Our analysis finds a mix of hateful and bot-based harmful communities, indicating the complexity of identifying different forms of vulnerability in online protests.

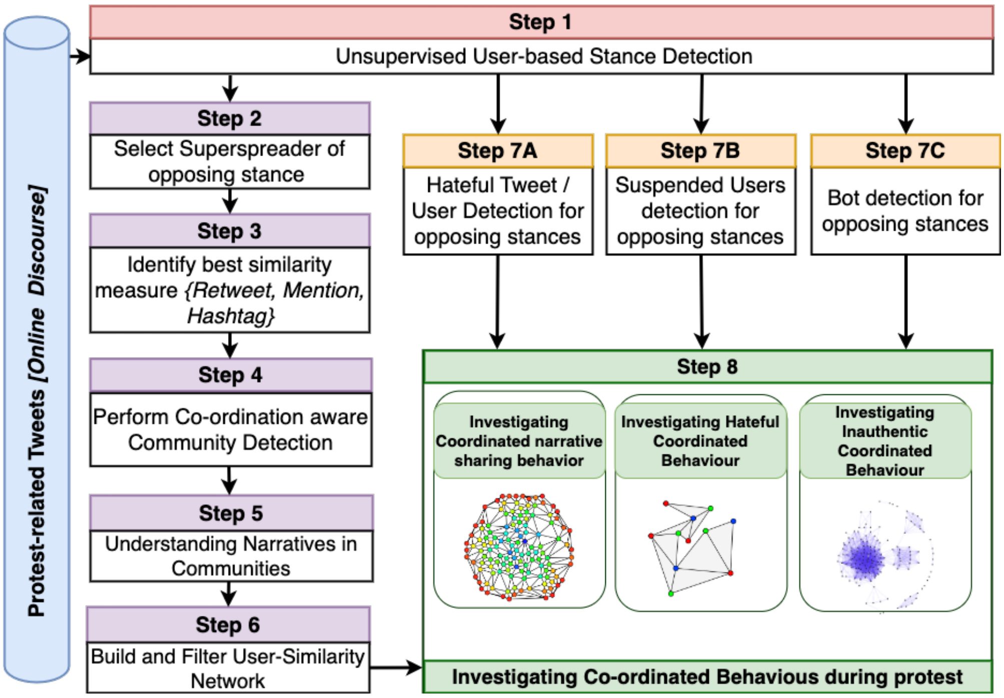

Protest-related Tweets [Online Discourse]
Step 1
Unsupervised User-based Stance Detection
Step 2
Select Superspreader of opposing stance
Step 3
Identify best similarity measure {Retweet, Mention, Hashtag}
Step 4
Perform Co-ordination aware Community Detection
Step 5
Understanding Narratives in Communities
Step 6
Build and Filter User-Similarity Network
Step 7A
Hateful Tweet / User Detection for opposing stances
Step 7B
Suspended Users detection for opposing stances
Step 7C
Bot detection for opposing stances
Step 8
Investigating Coordinated narrative sharing behavior
Investigating Hateful Coordinated Behaviour
Investigating Inauthentic Coordinated Behaviour
Investigating Co-ordinated Behavious during protest

Figure 7.1: Overview of our approach based on combining the analysis of hateful user detection and various inauthentic activities with coordination for examining coordinated harmful behavior.

7.2 Data and Methodology

7.2.1 Data

We use the official Twitter API to collect 11,350,276 tweets from 931,175 users about the CAA protest between December 07, 2019, and February 27, 2020, through daily trending hashtags around the protest. The detail of data collection and data pre-processing in described in detail in Chapter 5.

7.2.2 Method

This section first describes our approach to identifying the narrative-based coordinated communities formed during the online discourse of #CAA protest. Next, we examine the strength of various narrative communities formed in opposing stances and gauge the presence of inauthentic and harmful behavior in them. We define harmful behaviors as the tendency to share tweets with hateful content that may fuel disharmony. The inauthentic users may tend to manipulate users or create disharmony during the protest. We consider the Twitter bot-like behavior and Twitter suspension as ground truth for inauthenticity and harmful behavior, respectively. Figure 7.1 shows the 8 steps followed in our framework for the fine-grained analysis of the protest under study.

Unsupervised Stance Detection

Based on the online discourse towards the CAA, we identify the two cohorts of users as protesters (**P**) and counter-protesters (**CP**), based on whether they were against the Act or were in favor of it, respectively. To identify users from opposing cohorts, we build upon the unsupervised method for user-based stance detection proposed by [159]. To perform stance detection, we first identify the protest and counter-protest hashtags in our dataset through a manual investigation as shown in Table 5.1. We identified 27 counter-protest hashtags and 48 protest hashtags that we use for stance detection carried out in 6 steps: (i) *Hashtag-based labeling*: Identification of users who only tweet with hashtags from either protesters or counter-protesters side, (ii) *Label propagation*: include users who retweet the users identified in step (i) at least k-times (k = 15) [159], (iii) *Embedding creation*: create 1024-dimensional user embedding obtained from taking the average of the vector of the filtered tweets for each user using LASER (Language-Agnostic Sentence Representations)[1], (iv) *Dimensionality reduction*: use Uniform Manifold Approximation and Projection (UMAP) algorithm [133] to project users in 2-dimensional space, (v) *Clustering*: cluster the 2-dimensional embedding using density-based approach, such as Hierarchical Density-Based Clustering(HDBSCAN) [132], (vi) *Cluster purity*: use the identified stance produced from label propagation of step (i) to label the stance of the cluster, if the cluster is pure (i.e., contains at least 30% labeled users obtained via label propagation and has at least 80% purity of labels). Step (i) for Unsupervised stance detection yielded $106,605$ **CP** and $79,493$ **P** users. Step (ii) increased users set to $114,977$ **CP** and $79,613$ **P** through label-propagation. We perform pre-processing on users obtained through Step (ii) to remove users with less than 5 tweets, bringing the dataset to $270,889$ users. Step (iii) creates 1024-dimensional embedding, and step (iv), projects users onto 2-dimensional space. In step (v), we obtain 5 clusters for $270,889$ users. On performing purity analysis on the obtained clusters in step (vi), we found 4 clusters have more than 80% purity of labels, 2 from the **P** and 2 from the **CP** side. The final dataset used for further analysis

[1] https://github.com/facebookresearch/LASER

included $263,869$ users, divided into $142,839$ **CP** and $121,030$ **P**, as obtained from the clusters.

	CP	**P**	**Total**
Users	7,480	5,383	12,863
Total Tweets	732,550	434,784	1,167,334
Retweet	732,035	434,611	1,166,646
Tweets	515	173	688

Table 7.1: Statistics of the total engagement produced by **CP** and **P** *superspreaders* during the online protest.

Superspreaders Identification

To ensure the richness of the conversation during the discourse, we filtered tweets with 1,000 or more occurrences from the dataset. The occurrence may be from either tweets or retweets, identified through simple string matching. Next, we perform another iteration of user filtration to remove users with less than 5 tweets that may have been removed from the occurrence-based tweet filtration process. This reduced dataset to 2,767,268 tweets from 128,682 users, divided into $1,717,091$ tweets by $74,829$ **CP** and $1,050,177$ tweets by $53,853$ **P**. Finally, to analyze coordinated communities, we consider *superspreaders* [146] from the opposing stances [152] defined as the top 10% users from both sides with the highest retweets. Table 7.1 shows the statistics of engagement produced by *superspreaders* for the opposing stances. Our final dataset for further analysis consists of $12,863$ users who produced $1,67,334$ tweets/retweets.

Unsupervised Narrative Detection

To identify narratives shared by opposing stances, we build upon the unsupervised collective narrative detection technique proposed in previous literature [140]. The process can be divided into the following steps: (i) Identification of active tweets in the protest, where we perform string matching on the complete dataset (**CP** and **P** combined) and consider tweets with semantics duplicates above 30. We obtained $36,109$ active tweets corresponding to $7,878,996$ tweets/retweets in the dataset in this step, (ii) Projection of the tweets onto a two-dimensional plane using UMAP, (iii) Clustering of the projected tweet vectors using HDBSCAN, (iv) Manually annotating randomly selected 2 sets of 10 tweets per cluster. The calculation of Cohen's Kappa [21] produced a strong

	Counter-Protesters (CP)	Protesters (P)
SKEP	This is not a student protest; this is not even a protest against #CitizenshipAmendmentAct. I smell an international conspiracy, proxy battle to weaken India, using some students as fronts. Not 1 provision in bill is anti minority. People of India will reject violence.	Why not BJP doing something which benifit directly to hindus like for jobs , education , police reforms and reduce ground corruption for Hindus? Because BJP is failed want to divert important issue and fooling hindus in the name of CAB that's why I m against #HindusAgainstCAB
QUEST	Today, some people of Assam and some states (Tripura, Kerala) are protesting on the road against #CAB Are all the people of these states against #CAB? If they are against, then are Muslim refugees more valuable than Hindu refugees? If valuable then why and for what??	Where is it justified by the BJP government to enter the library hostel of Jamia University and release tear gas, beat up youths? Can students not protest against the #CAB2019 on the soul of the constitution?
OGA	What does #CABBill have to do with 'Indian' Muslims? Yes, the basis of citizenship is 'religious', that means citizenship to the minorities of Pakistan, Afghanistan, Bangladesh who have faced religious persecution but that too foreigners. So how is this #CitizenshipAmendmentBill against the Muslims of India?	Delhi Police burnt the buses, is the government trying to set the country on fire. This is what it seems by looking at these pictures. #CABProtest Delhi Protest: Did police burn buses in Delhi?
GRV	People throw stones in India and expect that they don't die. #isupportka_nurse'	Indian women showing the world how its done.. You can't You can kill refugees from Pakistan and Afghanistan? #CAA_NRC_Protests #CAAProtest #IndiaAgainstCAA #NRC_CA
CTA	if u don't stand today against illegal refugee then this will be your future.. #ISupportCAA_NRC #IndiaSupportsCAB #IStandWithRajnikanth	I don't support CAA NRC. I have voted for the nation, its time you also voice your opinion right here. #IndiaDoesNotSupport-CAA_NRC

Table 7.2: Table showing the example tweets of different narratives present in the *counter-protesters* CP and *protesters* P side.

agreement between annotators (0.95), (v) Map the clustered and cross-verified labeled narratives to our dataset presented in Table 7.1 for further analysis.

Extent of Coordination Computation

To measure the coordination in the discourse, we build upon the network-based coordination detection from previous literature [146]. Rather than analyzing the overall protest for coordination, we suggest detecting coordination within specific user stance clusters. This approach can help to unravel the intricate dynamics and shared narratives within a particular stance, which may not be apparent when examining coordination across the entire dataset. To this end, we focus on the *superspreaders*, i.e., the top 10% users in the discourse, as shown in Table 7.1, who were collectively responsible for 42% of total engagement (tweets/retweets) in the #CAA protest. Once the *superspread-*

ers have been identified, we first select the coordination measure that best represents coordination from the pool of coordination mechanisms identified in the literature, i.e., co-retweet, co-hashtag, and co-mention [145; 94]. We start with computing the TF-IDF-weighted vector for the three mechanisms used for coordination detection, so that we discount popular and emphasize relevant tweets/mentions/hashtags. For co-retweet, we compute the TF-IDF vector of the Tweet IDs the user has tweeted. For the co-hashtag, we compute the TF-IDF vector on the author's hashtags throughout the protest. For co-mention, we compute the TF-IDF vector on the set of all user mentions done by the author in their tweets. Next, we compute the similarity between the corresponding vectors of the *superspreaders* using cosine similarity. The pair of *superspreaders* and their cosine similarity result in an undirected weighted user-similarity network. Next, we use a multiscale backbone to retain statistically significant network structure [177]. Finally, we use the Louvain community detection algorithm to identify communities within stance clusters. To decipher the strength of coordination, we perform network dismantling and remove nodes and edges iteratively based on the moving edge weight threshold. More formally, we remove the weak edges at each iteration till we have exhausted all the nodes in the network. Since the edge weight is used as a proxy for coordination, each subsequent network represents a different extent of coordination, measured by the corresponding value of the moving threshold. For every user, the coordination score corresponds to the threshold value at which the node gets disconnected from the rest of the network. Among the 3 mechanisms under consideration, we found mention showed the strongest coordination behavior in #CAA, as it retained users over the higher threshold values (0.8 to 1). Hence, we perform further analysis on the similarity mechanism as *mentions* with the edge weight corresponding to the users who the two connected nodes have mentioned in their tweets during the protest.

Hate Speech Detection

From the active set, which consists of 128,682 users and 2,767,268 tweets, we manually annotate 2,000 tweets (1,000 from each **CP** and **P**) to identify hate speech. Two groups of four annotators performed the annotation, and Cohen's Kappa showed a strong agreement of 0.91 [21]. We resolved differences and ended up with 245 hate tweets and 1,755 non-hate tweets in our annotated sample, with 131 hate and 869 non-hate tweets from

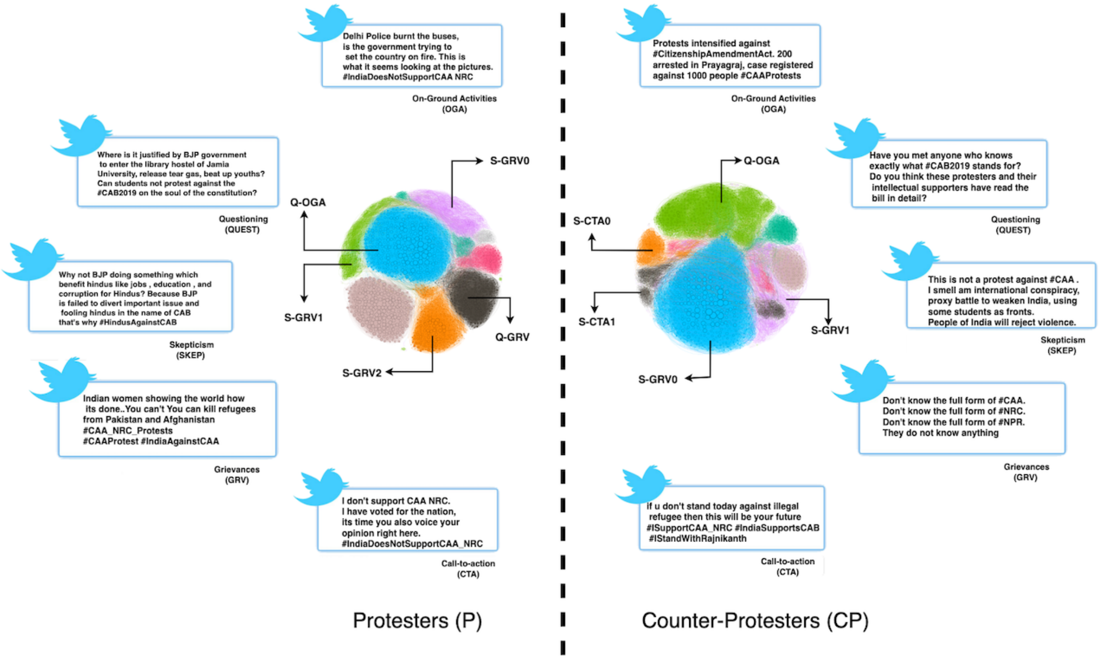
CP and 114 hate and 886 non-hate tweets from **P**. We utilized our annotated data to develop a hate speech detection classifier and experimented with previously proposed pipelines for hate detection in protest-related tweets. We used LASER with Logistic Regression as our baseline [14] and achieved an F1-score of 0.64. To improve the F1 score, we developed a multi-task learning (MTL) framework for hate speech detection, with stance prediction as an auxiliary task. The MTL model helped capture the tweet's context and reduced the risk of overfitting from imbalanced data [27; 168]. We used a multilingual sentence encoder to generate contextualized representations for each tweet [14], fed into a shared transformer architecture. Classification heads were placed at the transformer embedding to get the hate and stance predictions. We used standard cross-entropy loss for training the model. The data was split into the train, validation, and test sets with a ratio of 70:10:20, and we found that random under-sampling of the majority class worked best for our framework (LASER-MTL). The final dataset for training and testing comprised 490 samples, with 245 being hate tweets. The model was trained for 15 epochs with a learning rate of $5*10^{-5}$ and a batch size of 8 and obtained an F1 score of 0.93. We selected the saved model with the best performance to predict hate on the complete dataset. We manually annotated 50 stratified samples from each opposing stance to verify the expected class and found the model's efficacy reasonable. Hence we used the detected hate classes for further analysis of each community's hate score. We compute the hate score of the i-th community c_i as follows:

$$H_c(C_i) = \Phi(H_u(u_j) \forall u_j \in c_i, \tag{7.1}$$

where Φ is the community-level aggregation function.

7.3 Analysis

In this section, we discuss the results obtained from the analysis of the coordinated communities formed during the protest. Using our best-identified mechanism for coordination (i.e., *mention* as described under the Method Section), we first perform an unsupervised narrative detection on the communities obtained from the user-similarity network for **CP** and **P**, to analyze the broad narrative themes in the communities. Further, we investigate the coordination pattern in the obtained communities through net-



Delhi Police burnt the buses, is the government trying to set the country on fire. This is what it seems looking at the pictures. #IndiaDoesNotSupportCAA NRC

Where is it justified by BJP government to enter the library hostel of Jamia University, release tear gas, beat up youths? Can students not protest against the #CAB2019 on the soul of the constitution?

Why not BJP doing something which benefit hindus like jobs , education , and corruption for Hindus? Because BJP is failed to divert important issue and fooling hindus in the name of CAB that's why #HindusAgainstCAB

Indian women showing the world how its done..You can't You can kill refugees from Pakistan and Afghanistan #CAA_NRC_Protests #CAAProtest #IndiaAgainstCAA

I don't support CAA NRC. I have voted for the nation, its time you also voice your opinion right here. #IndiaDoesNotSupportCAA_NRC

Protests intensified against #CitizenshipAmendmentAct. 200 arrested in Prayagraj, case registered against 1000 people #CAAProtests

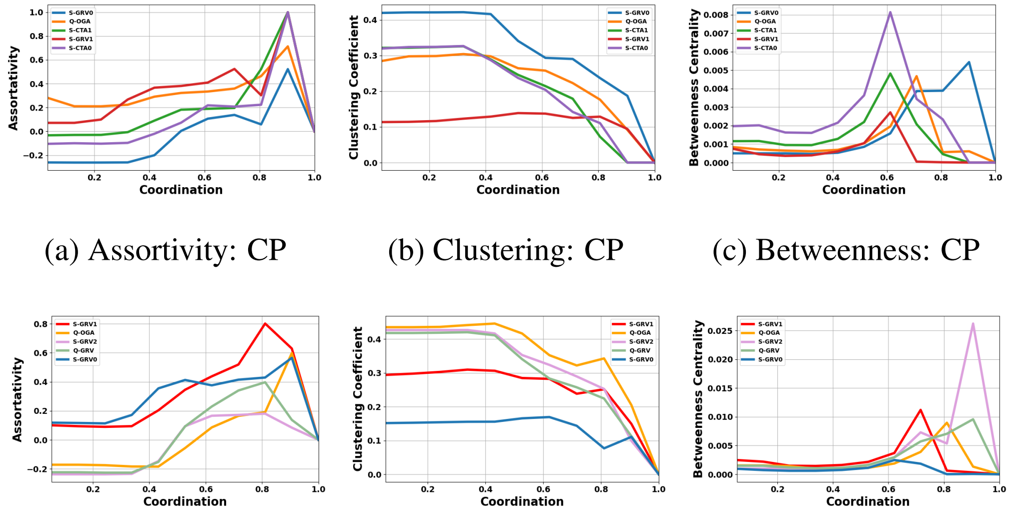
Have you met anyone who knows exactly what #CAB2019 stands for? Do you think these protesters and their intellectual supporters have read the bill in detail?

This is not a protest against #CAA . I smell am international conspiracy, proxy battle to weaken India, using some students as fronts. People of India will reject violence.

Don't know the full form of #CAA. Don't know the full form of #NRC. Don't know the full form of #NPR. They do not know anything

if u don't stand today against illegal refugee then this will be your future #ISupportCAA_NRC #IndiaSupportsCAB #IStandWithRajnikanth

Figure 7.2: Communities obtained on the user-similarity network from mention metric for *superspreaders*. A total of 9 communities were formed in P, and 8 communities were formed in CP. Narrative labels are written for the top 5 communities in opposing stances, with P narratives on the left side and CP narratives written on the right side.



(a) Assortivity: CP

(b) Clustering: CP

(c) Betweenness: CP

(d) Assortivity: P

(e) Clustering: P

(f) Betweenness: P

Figure 7.3: Figure showing the relationship between computed network measures for each coordinated community as a function of the extent of coordination.

work measures. The network measures used in the analysis help to quantify the extent of coordination in the obtained communities. Finally, we combine the community narrative labels with coordination and hate scores to study hate as a function of coordination in each community. Finally, we compare our analysis for hateful coordinated communities with the presence of Twitter suspended users and bots to check for the role of different authentic and inauthentic users in the dissemination of hate in the communities.

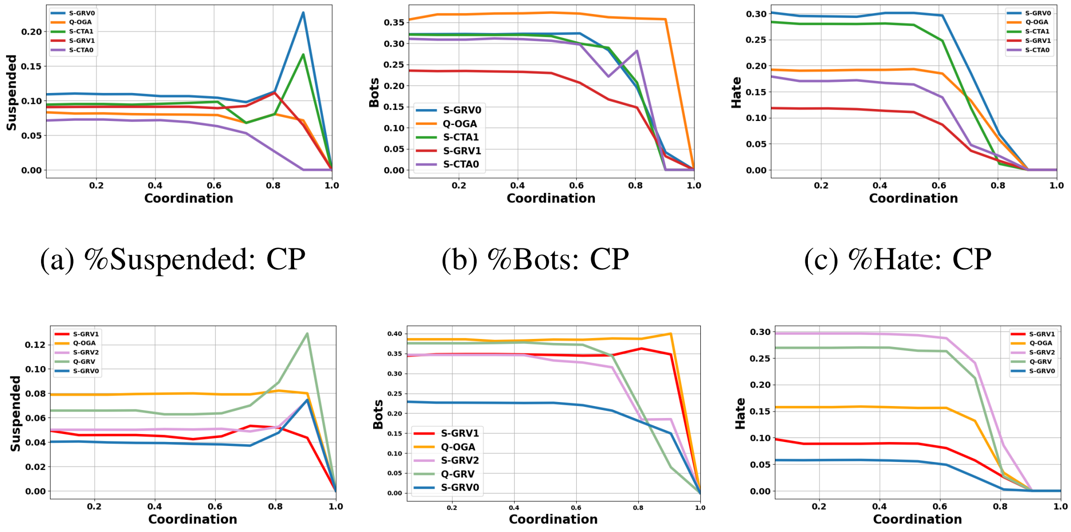
7.3.1 Forming Coordinated Communities

In this section, we analyze the narratives in the CAA protest and their dissemination by the different communities in opposing stances (**CP** and **P**). Table 7.2 shows the example of the narratives identified in the cluster on both sides. *Questioning* clusters are characterized by the presence of questions posed by the protest participant, which may vary from questioning CAA to questioning the protesters. *Skepticim* cluster's characteristics include doubt of the users either towards the legitimacy of the protest or any doubts toward the protest conduct, etc. *Call-to-Action* are the tweets that request the users to participate in the protest. While *on-ground activity* clusters are characterized by users reporting the real-world events as they unfold. *Skepticism* constituted 56% tweets in *CP* and 65% tweets in *P* forming the most dominant narrative on opposing stances. *Questioning* contained 38% tweets in *CP* and 28% tweets in *P*, forming the second most dominant collective narrative in the protest. Using the 'mention' similarity mechanism and performing community detection, we found 8 communities in **CP** and 9 communities in **P** as shown in Figure 7.2. Depicted by Step 5 in our methodology (Figure 7.1), we use an unsupervised narrative detection technique, where we cluster the unique active tweets and further use string matching of clustered tweets with all the tweets produced by *superspreaders* to identify their tweet narratives. We identified clusters of *questioning* (*quest*), *skepticism* (*skep*), *grievances* (*grv*), call-to-action (*cta*), and on-ground activities (*oga*) on both sides. Table 7.3 provides detailed statistics of the 5 largest communities formed in the **CP** and **P** sides.

We use the identified narratives present in the communities to label them for further analysis. This step also helps us to understand the collective narrative shared by the users as a function of coordination for each community. The communities' naming convention is based on the gradient of narratives present in the communities. As shown in Table 7.3, since **CP** and **P** both contained majority narratives as *skepticism* or *questioning*, we start the community name with S (for *skepticism*) or Q (for *questioning*), based on the which narrative between *skepticism* or *questioning* had more tweets. Next, we compare the number of tweets from non-dominant narratives (*grv*, *oga*, *cta*),

Name	Users	SKEP	QUE	GRV	CTA	OGA
			CP			
S-GRV1	36.08%	1,963	684	1,209	895	318
Q-OGA	19.67%	392	1,062	450	442	513
S-GRV0	12.47%	623	292	362	354	153
S-CTA1	11%	776	39	327	368	42
S-CTA0	6.28%	266	195	164	206	164
			P			
S-GRV0	37.03%	1826	159	1252	203	154
S-GRV1	13.09%	430	318	317	72	260
Q-OGA	10.41%	266	290	202	44	239
Q-GRV	9.06%	460	24	322	33	23
S-GRV2	8.91%	437	40	282	43	48

Table 7.3: Distribution of the different narratives present in the **CP** and **P** coordinated communities.SKEP: Skepticism, QUE: Questioning, GRV: Grievances, CTA: Call-to-Action, OGA: On-ground Activities. We found that skepticism and grievances community (S-GRV) contained maximum number of users in both **CP** and **P**.



(a) %Suspended: CP

(b) %Bots: CP

(c) %Hate: CP

(d) %Suspended: P

(e) %Bots: P

(f) %Hate: P

Figure 7.4: Figure showing the relationship between the percent suspended users, percent bots, and mean hate for each coordinated community as a function of the extent of coordination for the opposing stances.

and the majority of the non-dominant narrative is chosen to complete the name. The largest community formed for **CP** contained 36.08% users and showed *skepticism* and shared *grievances* (S-GRV1). The second largest community in **CP** contained dominantly *questioning* and *oga* using the same convention (Q-OGA). The third largest community with 12.47% users shared *skepticism* and shared *grievances* dominantly (S-GRV0). *Call-to-action* formed the dominant narrative for the fourth and fifth largest communities for *CP*, both accompanied by *skepticism* (S-CTA1, S-CTA0 respectively). In **P**, however, the two largest communities (37.03% and 13.09% users respectively) show *skepticism* and *grievances* as the majority narrative (S-GRV0 and S-GRV1 respectively). The third largest community shared *questioning* and *oga* (Q-OGA). In contrast, the fourth largest community shared *questioning* and *grievances* as the majority narrative (Q-GRV). The fifth-largest community shared *skepticism* and *grievances* as the majority narrative (S-GRV2). In summary sharing of tweets showing *skepticism* with *grievances*, and *questioning* with *oga* narratives were dominantly present across the communities of *superspreaders* in the protests. Seemingly, **CP** shared skeptical narratives more dominantly, while *grievances* was the focus for more **P** communities. We also found that *cta* featured more dominantly in **CP** communities.

7.3.2 Strength of Narrative Based Coordinated Communities

As per RQ2, we analyze network measures for communities to understand the pattern of coordination in opposing stances. To decode the coordination pattern in the communities, we perform community detection iteratively on sub-networks obtained through repetitively removing edges with weight lesser than the moving threshold and the nodes disconnected in the process, depicted by Step 6 in the Figure 7.1.

First, we explore the structural properties of the coordinated communities with the help of different network measures. Figure 7.3(a) and Figure 7.3(d) shows the assortativity of the coordinated network for **CP** and **P** respectively. Assortativity measures the tendency of nodes to be connected to similar nodes in the network. The communities appear strongly assortative for both **CP** and **P**, especially as we move towards higher coordination extent. This shows that the users were connected with similar users, forming a clique [2] of coordinated users. In **CP**, the communities with the highest increasing trend of assortativity towards a higher coordination extent were S-CTA1, S-CTA0, and S-GRV1, while in **P**, it was S-GRV1, S-GRV0, and Q-OGA communities. Next, Figure 7.3(b) and Figure 7.3(e) show the clustering coefficient vs the extent of coordination of the **CP** and **P** communities, respectively. The clustering coefficient measures how the nodes are clustered together in the network, i.e., whether all the nodes know each other. Both **CP** and **P** communities show well-organization and clustered in the network, given by the decreasing trend of the communities. The highest clustering is exerted by S-GRV0, Q-OGA, and S-CTA1, respectively, for **CP**, while the three highest

[2] clique of coordinated users refers to a group or network of users working together in a coordinated fashion

clustering coefficients towards stronger coordination for **P** are Q-OGA, S-GRV2, and S-GRV1, respectively. Next, we see the betweenness centrality for the communities, which measures how much influence a particular node has on the flow of information in the graph. Figure 7.3(c) and Figure 7.3(f) show the betweenness centrality vs. the extent of coordination of the **CP** and **P** communities, respectively. We witness that towards the greater extent of coordination, the betweenness centrality shows falling trend for both **CP** and **P** communities. For **CP**, the community with higher betweenness centrality towards stronger coordination was S-GRV0 (coordination $\simeq$ 0.9), while one exceptional betweenness centrality community for **P** was S-GRV2 (coordination $\simeq$ 0.9)

7.3.3 Harmful Coordinated Communities

In this section, we combine the narrative labels with the coordination and hate score to study the trend of hate as a function of coordination for communities in opposing stances. Given that the ground truth for harmful behavior is not present in the real-world event, to validate our finding about harmful coordinated communities, we compare our findings of hateful and coordinated communities with established harmful and inauthentic behaviors, i.e., users suspended [3] by Twitter and automation (bot score), respectively. Due to rate-limit, we sample 50% *superspreaders* from each opposing stance (4,262 users from **CP** and 3,463 users from **P**) and used Botometer API [219] to identify the bot, based on universal scores having a value $\geq$0.7.

Hate Speech

Figure 7.4 (c) & (f) shows the mean hate produced by communities of varying threshold values in **CP** and **P** communities, respectively. For **CP**, we find that S-GRV0, S-CTA1, and Q-OGA communities show plateaux structure up to threshold 0.6, after which communities show a marked declining trend. The communities S-CTA0 and S-GRV1 were the least hateful and also showed a declining trend below threshold 0.6, implying that there was not much coordination in the hateful users of the 2 communities. In **P**, the highest hateful content ration-bearing communities were S-GRV2, Q-GRV, and Q-OGA, respectively. S-GRV2, Q-GRV, and Q-OGA showed plateaux structure

[3] https://help.twitter.com/en/managing-your-account/suspended-twitter-accounts

till threshold 0.7, after which the communities showed a declining trend. S-GRV1 and S-GRV0 showed the lowest degree of hate and were not strongly connected. For **CP**, S-GRV0, S-CTA1, and Q-OGA were the three most hateful communities marked by a coordination score of $\simeq 0.6$, based on the elbow method. Among the coordinated communities in **P**, Q-GRV, Q-OGA, and S-GRV2 contain the most hate score and shown by plateaux structure up to threshold 0.7 (hence coordination score of $\simeq 0.7$).

Harmful & Inauthentic Activity

It is crucial to identify the deliberate sharing of hateful content during discourse as it can shift the focus and create social divisions. Hence, we conclude our work by comparing our hate and coordination score to the established signs of inauthenticity (bots) and harmfulness (suspended users). Figure 7.4(a),(d) shows the percentage of suspended users present in the coordinated communities in **CP** and **P**, respectively. In **CP**, the suspended users show strongly coordinated behavior for S-GRV0 and S-CTA1, showing the clique formation of the harmful communities (coordination value $\simeq 0.9$). Another community in **CP**, where the percentage of suspended users showed a rising trend towards a strongly connected community, was S-CTA1. In **P**, the suspended users show strongly coordinated behavior for Q-GRV. S-GRV2 and S-GRV0 showed indifference for the percent suspended users until the highest threshold value (coordination score $\simeq$ 0.9). Q-OGA seemed indifferent to the coordination threshold variation, while S-GRV1 shows a decreasing trend. In summary, S-GRV0 from **CP** and Q-GRV from **P** showed the strongest coordination among the identified suspended users. The Inauthentic behavior is investigated through the presence of bots in the coordinated communities for **CP** and **P**. In **CP**, out of 4,262 users, 2,630 users were identified as bot accounts, while 2,153 out of 3,463 **P** users were identified as bots. In **CP**, we found that Q-OGA exerted the strongest coordination among the bots, presented by the plateaux structure, as shown in Figure 7.4(b); however, the bot scores showed indifference to the changing threshold of coordination. S-CTA0 was found to be the only community that showed stronger coordination of bots, characterized by a rising trend. However, in the case of **P**, the Q-OGA and S-GRV1 narrative showed a plateaux structure up to coordination score $\simeq 0.8$, after that, showed a rising trend. Q-GRV, S-GRV2, and S-GRV0 showed decreasing coordination, indicating that the user at the highest coordination were not bots.

Overall, among the communities formed in **CP**, S-GRV0 appears to be the most harmful community, shown by high hate validated by having the largest share of suspended and the second largest share of bots, showing the highest disassortativity (indicates the formation of hubs), and highest clustering coefficient (indicates nodes had acquaintance with each other). Another harmful community identified in **CP** was S-CTA1, which had high hate, bot, and suspended user share; however, it showed an assortative network with a very low clustering coefficient, indicating a more distributed harmful behavior. The two communities with the least problematic for **CP** were found to be S-CTA0 and S-GRV1. For **P**, we found Q-GRV was the most harmful community, characterized by the highest suspended users and second highest hateful content, with a disassortative network structure and low clustering coefficient. Another problematic community was Q-OGA, which had second highest suspended, highest bot activity, and third highest bot accounts. S-GRV0 community in **P** was the least harmful with the least hateful content, hence being considered an authentic community. Another community, S-GRV1, showed more bot activity and less hateful content, while S-GRV2 contained fewer bots but more hateful behavior indicating the complexity of understanding the different forms of coordinated community and its malicious behavior. Due to the pursuit of spreading hate, we mark S-GRV2 as harmful, despite the community's low bot and suspended scores.

7.4 Discussion And Future Work

We conducted a novel analysis of hate and coordination during online discourse with opposing stances. Using the 2019 Citizenship Amendment Act discourse in India as a case study, we identified the distinct narratives shared by coordinated communities with opposing stances and gauged the level of coordination among them. Additionally, we utilized hate as a metric for harmful activity and assessed the presence of inauthenticity in different communities based on the bot and suspended account behavior. Among the counter-protester's (**CP**) coordinated communities, S-GRV0 (skepticism and grievances) was the most harmful, with high hate, high suspended users, and high bot presence. However, another coordinated community with a similar narrative i.e., S-GRV1 (skepticism and grievances), was authentic. In the protester's (**P**) coordinated communities, Q-GRV (questioning and grievances) was the most harmful, with the highest number of suspended users and high hate, followed, while S-GRV0 (skepticism and grievances) was found to be authentic. We also discovered communities that exhibited fewer bots but more hate S-GRV2 (skepticism and grievances) in **P**, indicating the multi-faceted harmful behavior during the online protest conduct. Differentiating between similar narrative communities from authentic and inauthentic sources is left as a topic for future research.

CHAPTER 8

CONCLUSIONS, LIMITATIONS AND FUTURE WORK

This chapter provides a comprehensive overview of our research on understanding various aspects of online social media-mediated protests, including strategies employed, shared narratives, and harmful behaviors. Our investigation encompasses protests facilitated by social media platforms across different regions worldwide, with Twitter as the primary analysis platform.

To begin, we examined activists' strategies to conduct online protests through the lens of protest over the cause of death of Indian actor #ShushantSinghRajput. Next, we delved into understanding the objectives of these projects, mainly through the lens of collective narratives shared during the protests. By analyzing the narratives prevalent in these online movements, we gained insights into the motivations and intentions of the participants. We analyzed the collective narratives present in the protest with the help of 4 protests under study, i.e., #CitizenshipAmendmentAct, and #FarmerProtest in India, the #KillTheBill protest in the United Kingdom, and the #BlackLivesMatter movement in the United States of America. Next, we acknowledged the presence of genuine and inauthentic users within the protest landscape and thoroughly analyzed the contributions made by authentic and inauthentic actors involved in protests in a discourse setting of #CitizenshipAmendmentAct. Finally, we addressed harmful behavior by focusing on coordinated-inauthentic activities, hateful activities, Twitter suspension, and Twitter Bots during the protests.

In this chapter, we briefly discuss the summary of each of the thesis's contributions in Section 8.1 and present the limitations and future work in Section 8.2.

8.1 Summary

This thesis's focus of objectives is twofold. Firstly, we aim to understand how social media facilitates the achievement of protest goals and uncover collective narratives shared during the protest. Secondly, we seek to identify harmful behaviors during the divergent discourse of a protest. We divide the thesis into four parts: (i) Understanding the strategy and objective used for online protest sustenance, (ii) Detecting and analyzing collective narratives shared during protests, (iii) Detecting and analyzing the opposing stances during the protest inclusive of authentic and inauthentic actors, (iv) Detecting and analyzing harmful behavior during protest.

8.1.1 Understanding Strategies

Our first contribution examined how Twitter activists build a diverse global support network and challenge the dominant narrative during an online protest. As a case study for studying the growth of online protest, we analyzed the strategy and objective of the protest surrounding the cause of the death of Indian actor Sushant Singh Rajput (#SSR) on Twitter. Despite the cause of death being reported as a suicide by the officials, a counterpublic movement emerged on Twitter, discussing alternative theories such as nepotism and murder, leading to an online protest on various social media, including Twitter. This study sheds light on how hashtag activism can evolve into connective action by examining the mechanisms of generative role-taking, hashtag-based storytelling, and issue alignment among diverse activist groups. The application of the connective action framework to analyze the counterpublic campaign surrounding the untimely death of Sushant Singh Rajput (SSR) on online social media provided valuable insights into protest strategies. Understanding generative role-taking through the construction of a user retweet network revealed the importance of influential information generators, which have a shorter path to reach fellow activists. Additionally, it highlighted the active connections maintained by top drivers. This knowledge helped comprehend the dynamics of information dissemination within the campaign and the roles played by key participants. Identifying the most consistent hashtag, such as #justiceforssr, and the peak usage of #candle4ssr provideed valuable insights into the effective mobilization of the counterpublic campaign. Hashtags served as rallying points, allowing activists to coordinate their efforts, express solidarity, and amplify their message. Analyzing these hashtag patterns informed our understanding of effective communication strategies in online protests. Moreover, the community detection analysis conducted on the retweet network reveals the presence of clique formation. This indicated a combination of centralized and decentralized information aggregation, with densely connected top generators and some individuals having sparse connections. Recognizing these patterns helped in comprehending the structure and organization of the counterpublic campaign, as well as the interplay between different activist groups. Overall, this research contributes to our understanding of protest strategies by demonstrating how the connective action framework can be applied to study online social media campaigns. It provided insights into the dynamics of information diffusion, the role of hashtags in storytelling and mobilization, and the formation of communities within the activist network. These findings can inform future protests and movements and aid in the development of more effective strategies for online activism.

8.1.2 Detecting And Analyzing Different Narratives

Next, we studed and examined collective narratives shared during protests and their role in shaping and advancing collective opinions. To this end, we collected Twitter data from 4 protests from different demographic locations centered around anti-government policy or bill-related topics. We collected tweets for bill-related protests in India (CitizenshipAmendmentAct (CAA) and FarmerProtest (FP)) and the United Kingdom (KillTheBill Protest (KTB)). We also collect data from the BLM (BlackLivesMatter) protest that led to the introduction of the George Floyd Justice in Policing Act in the US legislation. For all the 4 protests under study, we found the presence of call-to-action (CTA) and reporting of on-ground activity (OGA) narratives. Another standard narrative across protests was sharing grievances (GRV). Our analysis suggests that the narrative clusters can help reveal the underlying participant's intention, based on which narratives are being discussed dominantly. We found skepticism and questioning were the two most dominant narratives for the CAA protest, indicating contention in public towards the bill. For KTB and FP, CTA formed the most dominant cluster indicating people's will to participate and motivate others. While in BLM, the cluster with grievances narrative was dominant, showing that people were reporting complaints and resentments for what had happened in large numbers. With the help of the prominence score, we found a pattern of emojis, hashtags, and mentions used in protest-related tweets. We found that the emojis used in the protest were mainly protest-centric. For example, the FP protests had tractor and corps as emojis, while CAA had more religious-based emojis. The mentions in the tweets provide evidence that OGA has more verified accounts tagged. In contrast, the CTA mentions more of the general public, some suspended across protests under study. In terms of narrative evolution, we saw that CTA was more consistent throughout the protest timeline, while the OGA narrative peaked around substantial developments around the protest. For capturing the communication centered around different narratives, we examine the narrative-sharing behaviors for the top 5% Influential users based on the out-degree centrality of the retweet network. Across the 4 protests under study, we found low Betweenness centrality; and high Eigenvector centrality. This indicated that across the protests, the users didn't form more substantial edges between other users (Betweenness) but were connected to more Influential users (Eigenvector) and could have a faster flow of tweets in the network. Across the protests under study, we identified narrative-centric community formation, indicating that some sub-communities centered around a single narrative.

8.1.3 Detecting And Characterizing Opposing Stances, For Authentic And Inauthentic Actors

Since contentious topics are prone to divergent discourse, we delved into the opposing stances formed during an online protest in the next part of the thesis. We use India's #CitizenshipAmendmentAct protest as a case study to investigate the opposing stances and the content they shared during the discourse. We also analyzed the follower network of the opposing stances. Our investigation of the opposing stances accounted for different authentic and inauthentic actors on the platform and compares their shared content and network structure. We contribute to being the first study to perform a fine-grain analysis of the contention around the #CitizenshipAmendmentAct on Twitter regarding opposing stances and authenticity vs. inauthenticity combined. Our goal is to understand the participants' stances using unsupervised learning in a multilingual context and to identify major topics within the discourse from the perspectives of both protesters and counter-protesters. Additionally, we examined the presence and perception of various authentic and inauthentic actors in this discourse, specifically focusing on bots, suspended users, and deleted users as inauthentic actors. Users who were not categorized as inauthentic are considered authentic users.

To conduct our analysis, we collected a dataset of 9 million tweets related to the

CAA using trending hashtags in India. Our findings revealed the presence of inauthentic activities on both sides of the discourse. However, counter-protesters exhibited a higher level of inauthentic activity than protesters. By examining the frequency of tweets over time, we observed that much of the discussion was driven by inauthentic users, who tended to post less emotional content than their authentic counterparts.

Regarding the content shared by authentic users, both protesters and counter-protesters predominantly focused on topics such as violence and protest. In contrast, inauthentic users strategically shared more appealing content to garner attention. Analyzing the follower network of the participants revealed the presence of homophily, where users with similar stances tended to follow each other. Furthermore, one of the largest connected components in the follower network suggested a pathway between authentic and inauthentic users, indicating the potential reachability of inauthentic users to their authentic counterparts.

This work holds significant importance as it sheds light on the dynamics of online discourse surrounding a contentious issue like the CAA. By distinguishing between authentic and inauthentic actors, we provide insights into the manipulation attempts and the presence of coordinated activities within the discourse. These findings emphasize the need for critical evaluation and awareness among social media users to discern authentic and inauthentic voices. Furthermore, understanding the major topics and the strategies employed by different actors in the discourse can help develop more effective countermeasures against misinformation, polarization, and online manipulation.

8.1.4 Detecting And Analyzing Harmful Behavior

Among the harmful behaviors, we first focus on disseminating hateful content during online protests. To this end, we study how hateful users exploited the elements of protest mobilization (i.e., *resources* defined as the engagement methods on Twitter such as tweeting, retweeting, etc. and *ability to use them*) during the divergent discourse on #CitizenshipAmendmentAct in India. Since the user's stance plays a vital role in hateful tweet detection, we build a multi-task classification model with hate speech detection as the primary task and stance detection as an auxiliary task. Our model outperforms previous models catered towards Indian tweets, with an F1-score of 0.92. After we have trained the model, we classify the rest of the tweets for all stance-aware users as hateful or not. We use the above results to understand the spread of hate during the protest from content, user, and network perspectives. Our key findings revealed that the most hateful day coincides with the highest peak during the protest. On dividing the users based on hate intensity through k-means clustering algorithm over the frequency of hate tweets produced by all users, we find that more hateful users also produced more tweets and retweets during the protest. The mid-hate users, i.e., those who produced 9-27 hate tweets during the protest, made maximum retweets to hateful content and received fast retweets to their content. Among counter-protesters and protesters, low and high-hate counter-protesters exerted more initiative to participate and more communication power. The counter-protesters high-hate users exercised more central positions than their counterparts.

To delve deeper into the harmful activities in play during a protest, we combine different forms of harmful behavior with inauthenticity in our final part of the thesis. We use #CitizenshipAmendmentAct as a case study and decipher the various forms of inauthentic activities (bots, suspended users) and harmful behavior (hate speech and coordinated inauthentic behavior) exerted by the opposing stances during the online discourse. To this end, we identified the coordinated communities in the opposing stances, marked by the exceptional similarity between two users through different mechanisms such as hashtags, retweets, and mentions. Using the 2019 Citizenship Amendment Act discourse in India as a case study, we identified the distinct narratives shared by coordinated communities with opposing stances and gauged the level of coordination among them. Additionally, we utilized hate as a metric for harmful activity and assessed the presence of inauthenticity in different communities based on the bot and suspended account behavior. Among the counter-protester's (**CP**) coordinated communities, S-GRV0 (skepticism and grievances) was the most harmful, with high hate, high suspended users, and high bot presence. However, another coordinated community with a similar narrative, i.e., S-GRV1 (skepticism and grievances), was authentic. In the protester's (**P**) coordinated communities, Q-GRV (questioning and grievances) was the most harmful, with the highest number of suspended users and high hate, followed, while S-GRV0 (skepticism and grievances) was found to be authentic. We also discovered communities that exhibited fewer bots but more hate S-GRV2 (skepticism and grievances) in **P**, indicating the multi-faceted harmful behavior during the online protest conduct. Differentiating between similar narrative communities from authentic and inauthentic sources is left as a topic for future research.

8.2 Limitations and Future Work

In this section, we discuss the limitations and future work related to our thesis.

8.2.1 Limitations

In this thesis, a data-driven approach is employed to address the research questions and explore the topic thoroughly. Data-driven methodologies rely on the systematic analysis of empirical data to derive insights and draw conclusions. While this approach has several advantages, it is vital to recognize and discuss the inherent limitations to ensure the integrity and credibility of the research findings. We first acknowledge the limitations of our approaches.

- **Identifying Real Participants in the Protest**: Since we started the protest study through hashtag-based data collection, it imposes the problem of identifying whether the users participating in the protest are the concerned set of people. Although the analysis of tweets for a protest provides a glimpse of an overall discussion on the topic, the actual victims might remain unheard or unreached in the population. One of the ways to go about this problem is to identify the actual participants, map their social media profile, and perform analysis. However, this approach may lead to ethical concerns about user privacy invasion. Another major bottleneck for this approach is that it might be very expensive.

- **Data Sampling Limitations**: It is important to acknowledge the limitations of our dataset, particularly regarding the data sampling process. Our dataset may suffer from inherent biases because we only have access to a 1% sample of Twitter data from the REST API. For example, this sampling method may underestimate the presence of protest strategies or harmful behavior during the observed protests on Twitter. Despite this limitation, our dataset includes ample protest tweets, user profiles, and network data to gain valuable insights per the research questions. However, it is challenging to obtain an ideal and completely unbiased dataset.

- **Limited Coverage of Protests**: Our data collection process focused primarily on Twitter, which means we may have missed some popular developments of the protests that may be prevalent on other social networks. However, Twitter still provides a substantial sample of public content, allowing us to capture a significant portion of the campaigns under investigation.

- **Dataset Completeness**: Our data collection heavily relied on APIs provided by Twitter. While this approach ensured compliance with platform policies and terms of service, it may result in missing some posts related to the protests.
- **Real-World Information Completeness**: Although the data we are working on gives an overview of the online world, such as Twitter, it is hard to comprehend the real-time protest progress offline. We use the news articles as an alibi to map the online and offline protest conduct. However, the accuracy of mapping all the offline developments in the online world in real-time is still an open problem.
- **Limited Protests Under Study**: We conduct experiments with protests in India, the UK, and the US. Since the protests are subjective in nature and rooted in many social and political factors, increasing the protest-relevant data from other parts of the world may increase the inclusivity of our findings.

We believe there are several future directions that researchers can partake in for each of the contributions made in respective chapters. We discuss some of the future directions in the section below.

8.2.2 Future Work

In this section, we briefly discuss some future directions that the researchers can partake in to take the computational analysis of the protests further.

Understanding Strategies and Collective Narratives

Strategies and narratives employed during protests are subjective and constantly evolving. As a result, the study of protest participation remains relevant and ongoing. Continually examining protests can unveil new strategies and narratives that resonate with the public at the time. Additionally, further exploration is needed to gain a deeper understanding of the shared grievances that are the foundation of every protest [162]. Future researchers can conduct in-depth analyses of shared grievances across different protests and may propose methods to identify common formats of grievances across various movements. Future research can also use a multi-modal approach to protest study, where protest issue classification from placards can be combined with network and post to perform an effective and in-depth analysis [121].

Understanding Harmful Activities

In this thesis, we have successfully identified and examined harmful activities, including hateful activity, coordinated inauthentic behavior, bots, and suspended users, within protest-related discourse. However, it is important to recognize that the harmful activities present in protests can be diverse and encompass other aspects such as propaganda [95], fake news [34], disinformation [81], and more. Future researchers can explore the intricate interplay of various harmful behaviors within protests. Recent work by Valecha *et al.* [198], has found that threat and coping-related issues positively affect fake news sharing in health-related issues. In the future, how fake news sharing is affected by protest-related issues can be a promising direction. Another potential avenue for future research involves conducting a comparative analysis of different harmful activities during protests, aiming to discern the underlying intentions of these harmful users. Furthermore, investigating these factors on platforms other than Twitter remains an under-explored research direction. It is worth noting that one of the significant challenges in conducting such studies may arise in computational analysis for low-resource languages.

Unsupervised Protest Event Detection

So far, we have delved into understanding the strategies adopted, the narratives shared, and the harmful behavior in protests. The two major concerns that we aimed to address while conducting the studies were (i) to focus on the online protests in non-western countries, covering the research gap of catering to low-resource languages, (ii) to conduct the protest study in an unsupervised fashion, catering to the subjectivity and nature of the protest. In this section, we discuss another dimension of computational analysis of protest, i.e., protest event detection in non-western countries. Protest event detection aims to identify and extract pertinent data from a text about specific categories of events related to what, where, and when a protest might occur. There has been a lot of work done towards protest event prediction in Western countries and in a supervised manner [137]. However, the problem of protest event detection with respect to unsupervised settings and low-resource languages has been understudied [222].

The early detection of protests is very important for taking early precautionary measures. However, the main shortcoming of protest event detection is the scarcity of sufficient training data for specific language categories, making it difficult to train data-hungry deep learning models effectively. Therefore, cross-lingual and zero-shot learning models are needed to detect events in various low-resource languages. As a first step towards this future direction, we uses a multi-lingual cross-document level event detection approach using pre-trained transformer models developed for the dataset obtained from Shared Task 1 at CASE 2023 [98]. The dataset was spread over multiple languages (English, Spanish, Portuguese, Turkish, Urdu, and Mandarin). With this work, we emphasize towards detection of the protest event for low-resource language in a zero-shot fashion. Our system achieves an average F_1 score of 0.73 for the document-level event detection task. Our approach secured 2^{nd} position for the Hindi language in subtask 1 with F_1 score of 0.80.

Related Work: Early detection of ongoing and past events exploited feature-based approaches to detect events [115]. The early data-driven approaches [90] and knowledge-driven and rule-based approaches missed the semantic relationship in the data [51]. Other early approaches for event detection include machine learning models such as SVM and decision trees [175]. Recent deep learning approaches proposed in the literature [7] improve event detection. However, they are not generalized for low-resource languages. To address the data scarcity problem for low-resource languages, researchers have recently used the pre-trained language model GPT-2 to generate training samples [205]. Targeting the issues with scarce availability of low-resource languages, the CASE 2021 subtask introduced the multi-lingual crisis event detection dataset, which focuses on the zero-shot and few-shot detection of protest and crisis event [97].

Data: The dataset we use for protest event detection was obtained from CASE 22 shared task created in the process presented in [96]. The data is such created that some news documents contain protest event information, while some news document does not contain any protest events. The data provided for training are highly imbalanced and provided for only 3 languages. The testing data contains 7 languages, with documents from additional 4 languages apart from training data. Table 8.2 provides the details of the training data provided in the shared task. Table 8.3 presents the test data for the Task. Given that no training data is present for Hindi, Mandarin, Turkish and Urdu, the task of document event detection becomes a zero-shot classification problem.

Language	**Model**	**macro-F1**
English	mBert+Softmax	0.76
	XLM-Roberta+LSTM	0.74
	XLM-Roberta+Sigmoid	**0.77**
	XLM-Roberta+Sigmoid (U)	0.72
Spanish	**mBert+Softmax**	**0.69**
	XLM-Roberta+LSTM	0.63
	XLM-Roberta+Sigmoid	0.64
	XLM-Roberta+Sigmoid (U)	0.63
Portuguese	mBert+Softmax	0.68
	XLM-Roberta+LSTM	0.71
	XLM-Roberta+Sigmoid	**0.76**
	XLM-Roberta+Sigmoid (U)	0.72

Table 8.1: Test results for English, Spanish, and Portuguese documents, as reported in the shared task, for which training data was available. U: Under-sampled data.

Language	**Label 1**	**Label 0**	**Total**
English (En)	1,912	7,412	9,324
Spanish (Es)	131	869	1,000
Portuguese (pt)	197	1,290	1,487

Table 8.2: Statistics for the training Data available for Shared Task 1, subtask 1: Document-level crisis event prediction.

Language	**Documents**
English	3,871
Hindi	268
Mandarin	300
Spanish	400
Portuguese	671
Turkish	300
Urdu	299

Table 8.3: Statistics for the test Data for testing for Shared Task 1, subtask 1: Document-level crisis event prediction.

Since we experiment with mBERT (cased) and other sentence-based embeddings, we do not lowercase our document corpus before training. We also do not conduct any language-specific pre-processing to keep the preprocessing step language agnostic. However, we removed any URLs, or a single occurrence replaced repeated symbols. We also removed any extra spaces present in the data.

Methodology: Transformer-based models have recently gained success in various multilingual NLP tasks such as offensive content detection [19] and various zero-shot cross-lingual tasks [225; 113]. We experiment with different multi-lingual models and analyze how the different models perform on the downstream task of document classification in subtask 1. We design the document classification problem as a sequence classification problem [89; 83]. In our approach, we use different transformer models including XLM-Roberta [46], mBERT [59], and encoder-decoder-based LASER [20] to generate embedding from the documents. We experiment with different layers on top of the multi-lingual sentence embedding. Our preliminary analysis found that transformer-based XLM-Roberta with a sigmoid layer outperformed other models in the F1 score. Therefore, our approach proposes the XLM-Roberta model with a sigmoid classification layer for event prediction. XLM-Roberta is pre-trained on unlabeled Wikipedia text and CommonCrawl Corpus of 100 languages. XLM-Roberta has a vocabulary size of 25,000 and uses SentencePiece tokenizer [112]. We fine-tuned the model for our task with the training data provided. The training data was highly imbalanced. However, oversampling and under-sampling methods didn't provide any marginal improvement in the model's output as per our experiments. XLM-R belongs to an unsupervised representation learning framework as it doesn't use any cross-lingual resources [46]. XLM-R has L = 12 transformers, with H = 768 attention heads with A = 12, and 270M parameters. The maximum token size for input for XLM-R is 512 tokens. The token size of 512 is less for creating document-level creation, as a lot of information might not be captured. However, breaking the sentences into 512-length tokens might lead to an incorrect labeling process for different sentence splits [83]. Due to the limitation of our system, our final approach uses a 256-length token for document embedding creation. The learning rate was $2.75e^{-05}$, the batch size for training was 32, and the training was done for 20 epochs. The total training time taken for the XLM-R-based model was approximately 2 hours. Since we use the Sigmoid layer on the top of XLM-R, the final decision boundary for 0/1 was taken based on the probability of 0.6 for all cases.

Language	Model	macro-F1
Hindi	mBert+Softmax	0.71
	XLM-Roberta+LSTM	0.75
	XLM-Roberta+Sigmoid	**0.80**
	XLM-Roberta+Sigmoid (U)	0.77
Turkish	mBert+Softmax	0.69
	XLM-Roberta+LSTM	0.70
	XLM-Roberta+Sigmoid	**0.74**
	XLM-Roberta+Sigmoid (U)	0.69
Urdu	mBert+Softmax	0.67
	XLM-Roberta+LSTM	**0.72**
	XLM-Roberta+Sigmoid	0.71
	XLM-Roberta+Sigmoid (U)	0.73
Mandarin	**mBert+Softmax**	**0.75**
	XLM-Roberta+LSTM	0.71
	XLM-Roberta+Sigmoid	0.75
	XLM-Roberta+Sigmoid (U)	0.73

Table 8.4: Test results for Hindi, Mandarin, Turkish and Urdu documents, as reported in the shared task. Training data was not provided for the above language. Hence classification is done in a zero-shot setting.

For training of all models, we use the Nvidia RTX 3090 GPU system with an installed Cuda version of 11.3. For training, we combined the training data from the 3 languages, English, Spanish, and Portuguese, as shown in Table 8.2. We performed at a 90:10 split for training and testing, respectively. The split was done randomly but stayed the same for all the experiments with models to obtain the result on the same set of datasets. The score we demonstrated for document-level classification was the F1-macro metric, which was selected as an evaluation metric for our models. We performed experiments with different epoch numbers and batch sizes with the same experimental setup.

Results: In this section, we demonstrate our results from various models. We elaborate on the results from different models for each language. Table 8.1 shows the result for English, Spanish, and Portuguese language, for which we had training data. We found that XLM-Roberta with the Sigmoid layer outperformed for English and Portuguese tasks; however, the best model for Spanish was multilingual BERT with the softmax layer. Table 8.4 presents the results for the zero-shot classification for the respective languages. Our best model, the XLMRoberta+Sigmoid model, obtained a macro-F1 score of 0.80 for Hindi and secured 2^{nd} in the shared task. For Turkish, the best model also came out as XLMRoberta+Sigmoid, with macro-F1 as 0.74. For the

Urdu language, XLMRoberta+LSTM slightly outperformed the proposed model. For Mandarin, however, the best F1-score was obtained from the mBERT+Softmax model and XLMRoberta+Sigmoid.

Conclusion: We focus on the future directions for computational protest analysis, toward protest event detection in non-western countries and low-resource languages. The main motivation of this work is to emphasize the need for unsupervised event detection methods that can cater to low-resource languages. We focus on protest event detection at the docuemt level for low-resource language. We explored various multilingual and zero-shot approaches and showed results across the languages in subtask 1. We propose XLM-Roberta with a Sigmoid layer for classifying crisis events in zero-shot and low-resource language settings. Our system achieved an average F1 score of 0.73. Among the given languages, our proposed approach secured 2^{nd} place in the Hindi document event classification task. While comparing with our approach, the multilingual BERT with softmax layer obtained better results for Spanish and Mandarin, with the result for Spanish securing the 4^{th} spot in the shared task.

8.3 Ethical Concerns

This thesis has carefully addressed several ethical concerns associated with data collection from Twitter for protest-based analysis in India, the UK, and the US. Firstly, strict measures were implemented to protect data privacy and confidentiality. All personal identifying information of Twitter users involved in protests was anonymized and encrypted to ensure their identities remained secure throughout the analysis. Additionally, for future research, we only release the text of the tweets and refrain from any personally identifiable information being made publicly available. Our aim in this thesis is to focus on the collective wisdom of the crowd rather than individual perspectives on protest activities. Throughout the study, a responsible and balanced approach was adopted when analyzing and interpreting user-generated content. Any potentially harmful or incendiary content was handled with sensitivity, and efforts were made to disseminate the findings responsibly, focusing on the broader context and societal implications rather than singling out specific individuals or groups. Special care was taken to avoid misinterpretation or misrepresentation of tweets, ensuring that the findings accurately reflected the sentiments and perspectives expressed by the participants. Steps were also taken to verify and validate the data to enhance the credibility and reliability of the research outcomes. To reduce bias and improve diversity in the annotation task, we recruited annotators from different parts of the country to conduct annotations.

REFERENCES

[1] (2019). CAB protest: Students clash with police near Assam secretariat. *Economic Times*.

[2] (2019). Anti-CAA Protests Highlights December 24: Rahul and Priyanka Gandhi stopped outside Meerut by police, returns Delhi. *jagran News Desk*.

[3] (2019). Indian-Americans protest CAA third time in nine days, author=IANS. *freepressjournal*.

[4] (2019). Breaking news on December 20, author=India TV News Desk. *newindianexpress*.

[5] (2019). Flash protest in Hyderabad again CAA, NRC on New Year's eve: Six detained. *The News Minute.*.

[6] (2019). Internet shutdowns. URL `https://internetshutdowns.in/static-page/caa-protest/`.

[7] **Ahmad, F.**, **A. Abbasi**, **B. Kitchens**, **D. A. Adjeroh**, and **D. Zeng** (2020). Deep learning for adverse event detection from web search. *IEEE Transactions on Knowledge and Data Engineering*.

[8] **Akbar, S. Z.**, **A. Sharma**, **H. Negi**, **A. Panda**, and **J. Pal** (2020). Anatomy of a rumour: Social media and the suicide of sushant singh rajput. *arXiv preprint arXiv:2009.11744*.

[9] **Akhtar, S.**, **V. Basile**, and **V. Patti** (2021). Whose Opinions Matter? Perspective-aware Models to Identify Opinions of Hate Speech Victims in Abusive Language Detection. URL `http://arxiv.org/abs/2106.15896`.

[10] **Aldayel, A.** and **W. Magdy** (2019). Your stance is exposed! analysing possible factors for stance detection on social media. *Proceedings of the ACM on Human-Computer Interaction*, **3**(CSCW), 1–20.

[11] **ALDayel, A.** and **W. Magdy** (2021). Stance detection on social media: State of the art and trends. *Information Processing & Management*, **58**(4), 102597.

[12] **Alonso, O.**, **V. Kandylas**, and **S. E. Tremblay** (2018). How it Happened: Discovering and Archiving the Evolution of a Story Using Social Signals. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, (1), 193–202.

[13] **Aluru, S. S.**, **B. Mathew**, **P. Saha**, and **A. Mukherjee** (2020). Deep Learning Models for Multilingual Hate Speech Detection.

[14] **Aluru, S. S.**, **B. Mathew**, **P. Saha**, and **A. Mukherjee**, A deep dive into multilingual hate speech classification. *In* **Y. Dong**, **G. Ifrim**, **D. Mladenić**, **C. Saunders**, and **S. Van Hoecke** (eds.), *Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track*. Springer International Publishing, Cham, 2021.

[15] **Amenta, E.** and **M. P. Young** (1999). Democratic states and social movements: Theoretical arguments and hypotheses. *Social Problems*, **46**(2), 153–168. URL `http://www.jstor.org/stable/3097250`.

[16] **An, J.**, **H. Kwak**, **C. S. Lee**, **B. Jun**, and **Y.-Y. Ahn**, Predicting anti-Asian hateful users on Twitter during COVID-19. *In Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021.

[17] **ANI** (2019). Amid protests, Mamata Banerjee announces mega rally in Kolkata to protest against CAA, NRC today. *freepressjournal*.

[18] **ANI** (2019). CAA protest row: Section 144 imposed in UP Mau's Hajipura Chowk area. *freepressjournal*.

[19] **Arango, A.**, **J. Pérez**, **B. Poblete**, **V. Proust**, and **M. Saldaña** (2022). Multilingual resources for offensive language detection. *WOAH 2022*, 122.

[20] **Artetxe, M.** and **H. Schwenk** (2019). Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, **7**, 597–610.

[21] **Artstein, R.** and **M. Poesio** (2008). Inter-coder agreement for computational linguistics. *Computational linguistics*, **34**(4), 555–596.

[22] **Badawy, A.**, **A. Addawood**, **K. Lerman**, and **E. Ferrara** (2019). Characterizing the 2016 Russian IRA influence campaign. *Social Network Analysis and Mining*, **9**(1).

[23] **Badawy, A.** and **E. Ferrara** (2018). The rise of jihadist propaganda on social networks. *Journal of Computational Social Science*, **1**(2), 453–470.

[24] **Badawy, A.**, **K. Lerman**, and **E. Ferrara** (). Who Falls for Online Political Manipulation?

[25] **Badjatiya, P.**, **S. Gupta**, **M. Gupta**, and **V. Varma**, Deep learning for hate speech detection in tweets. *In Proceedings of the 26th international conference on World Wide Web companion*. 2017.

[26] **Bahrami, M.**, **Y. Findik**, **B. Bozkaya**, and **S. Balcisoy** (2018). Twitter reveals: Using twitter analytics to predict public protests. *CoRR*, **abs/1805.00358**. URL `http://arxiv.org/abs/1805.00358`.

[27] **Baxter, J.** (1997). A bayesian/information theoretic model of learning to learn via multiple task sampling. *Machine learning*, **28**, 7–39.

[28] **Bennett, W. L.** and **A. Segerberg** (2012). The logic of connective action: Digital media and the personalization of contentious politics. *Information, communication & society*, **15**(5), 739–768.

[29] **Bennett, W. L.** and **A. Segerberg**, *The logic of connective action: Digital media and the personalization of contentious politics*. Cambridge University Press, 2012.

[30] **Bessi, A.** and **E. Ferrara** (2016). Social bots distort the 2016 us presidential election online discussion. *First monday*, **21**(11-7).

[31] **Bimber, B.**, **A. Flanagin**, and **C. Stohl**, *Collective action in organizations: Interaction and engagement in an era of technological change*. Cambridge University Press, 2012.

[32] **Bittner, M.**, **D. Dettmar**, **D. Morejon Jaramillo**, and **M. J. Valta** (2020). Virtual tribes: Analyzing attitudes toward the LGBT movement by applying machine learning on Twitter data. *Springer Proceedings in Complexity*, 157–175.

[33] **Blei, D. M.**, **A. Y. Ng**, and **M. I. Jordan**, Latent dirichlet allocation. *In Advances in neural information processing systems*. 2002.

[34] **Bovet, A.** and **H. A. Makse** (2019). Influence of fake news in Twitter during the 2016 US presidential election. *Nature Communications*, **10**(1), 1–14. URL `http://dx.doi.org/10.1038/s41467-018-07761-2`.

[35] **Boyd, D.** and **K. Crawford** (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, **15**(5), 662–679.

[36] **Bozarth, L.** and **C. Budak**, Is slacktivism underrated? measuring the value of slacktivists for online social movements. *In Proceedings of the International AAAI Conference on Web and Social Media*, volume 11. 2017.

[37] **Burnap, P.** and **M. L. Williams** (2016). Us and them: identifying cyber hate on twitter across multiple protected characteristics. *EPJ Data science*, **5**, 1–15.

[38] **Burns, A.** and **B. Eltham**, Twitter free iran: an evaluation of twitter's role in public diplomacy and information operations in iran's 2009 election crisis. *In Communications Policy & Research Forum 2009*. 2009.

[39] **Chandrachud, A.** (2020). Secularism and the citizenship amendment act. *Indian Law Review*, **4**(2), 138–162. URL `https://doi.org/10.1080/24730580.2020.1757927`.

[40] **Chang, H.-C. H.**, **E. Chen**, **M. Zhang**, **G. Muric**, and **E. Ferrara** (2021). Social Bots and Social Media Manipulation in 2020: The Year in Review. URL `http://arxiv.org/abs/2102.08436`.

[41] **Chopra, S.**, **R. Sawhney**, **P. Mathur**, and **R. R. Shah**, Hindi-english hate speech detection: Author profiling, debiasing, and practical perspectives. *In Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34. 2020.

[42] **Cinelli, M.**, **E. Brugnoli**, **A. L. Schmidt**, **F. Zollo**, **W. Quattrociocchi**, and **A. Scala** (2020). Selective exposure shapes the facebook news diet. *PloS one*, **15**(3), e0229129.

[43] **Clauset, A.**, **M. E. Newman**, and **C. Moore** (2004). Finding community structure in very large networks. *Physical review E*, **70**(6), 066111.

[44] **Cohen, S.**, **W. Nutt**, and **Y. Sagic** (2020). Sushant Singh Rajput's father files FIR against actor's friend for abetting suicide.

[45] **Comaniciu, D.** and **P. Meer** (2002). Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on pattern analysis and machine intelligence*, **24**(5), 603–619.

[46] **Conneau, A.**, **K. Khandelwal**, **N. Goyal**, **V. Chaudhary**, **G. Wenzek**, **F. Guzmán**, **E. Grave**, **M. Ott**, **L. Zettlemoyer**, and **V. Stoyanov**, Unsupervised cross-lingual representation learning at scale. *In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 2020.

[47] **Conti, M.**, **S. K. Das**, **C. Bisdikian**, **M. Kumar**, **L. M. Ni**, **A. Passarella**, **G. Roussos**, **G. Tröster**, **G. Tsudik**, and **F. Zambonelli** (2012). Looking ahead in pervasive computing: Challenges and opportunities in the era of cyber–physical convergence. *Pervasive and mobile computing*, **8**(1), 2–21.

[48] **Contributors, I. T.** (2020). SSR death case: Study decodes how BJP politicians pushed 'murder' narrative. [Online; accessed 06-October-2020].

[49] **Costa, J. M.**, **R. Rotabi**, **E. L. Murnane**, and **T. Choudhury** (2015). It is not only about grievances: Emotional dynamics in social media during the brazilian protests.

[50] **Damini Nath, V. S.** (2019). After a heated debate, Rajya Sabha clears Citizenship (Amendment) Bill. *The Hindu*.

[51] **Danilova, V.** and **S. Popova**, Socio-political event extraction using a rule-based approach. *In OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*. Springer, 2014.

[52] **Darwish, K.**, **W. Magdy**, and **T. Zanouda**, Trump vs. hillary: What went viral during the 2016 us presidential election. *In International conference on social informatics*. Springer, 2017.

[53] **Darwish, K.**, **P. Stefanov**, **M. Aupetit**, and **P. Nakov**, Unsupervised user stance detection on twitter. *In Proceedings of the International AAAI Conference on Web and Social Media*, volume 14. 2020.

[54] **Das, M.**, **P. Saha**, **R. Dutt**, **P. Goyal**, **A. Mukherjee**, and **B. Mathew**, You too brutus! trapping hateful users in social media: Challenges, solutions & insights. *In Proceedings of the 32nd ACM Conference on Hypertext and Social Media*. 2021.

[55] **Dash, S.**, **D. Mishra**, **G. Shekhawat**, and **J. Pal** (2022). Divided we rule: Influencer polarization on twitter during political crises in india. *Proceedings of the International AAAI Conference on Web and Social Media*, **16**(1), 135–146.

[56] **De Choudhury, M.**, **S. Jhaver**, **B. Sugar**, and **I. Weber** (2016). Social Media Participation in an Activist Movement for Racial Equality. Technical report. URL `http://www.wired.com/2015/10/how-black-lives-matter-uses-`.

[57] **desk, E. W.** (2021). Farmers end year-long protest: A timeline of how it unfolded. URL `https://indianexpress.com/article/india/one-year-of-farm-laws-timeline-7511961/`.

[58] **desk, T. B. I. W.** (2022). What are the Kill the Bill protests? *The Big Issue*.

[59] **Devlin, J.**, **M.-W. Chang**, **K. Lee**, and **K. Toutanova** (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[60] **Dunn, A. G.**, **D. Surian**, **J. Dalmazzo**, **D. Rezazadegan**, **M. Steffens**, **A. Dyda**, **J. Leask**, **E. Coiera**, **A. Dey**, and **K. D. Mandl** (2020). Limited role of bots in spreading vaccine-critical information among active twitter users in the united states: 2017–2019. *American Journal of Public Health*, **110**(S3), S319–S325.

[61] **ElSherief, M.**, **V. Kulkarni**, **D. Nguyen**, **W. Y. Wang**, and **E. Belding**, Hate lingo: A target-based linguistic analysis of hate speech in social media. *In Proceedings of the International AAAI Conference on Web and Social Media*, volume 12. 2018.

[62] **Fast, E.** and **E. Horvitz**, Identifying dogmatism in social media: Signals and models. *In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, 2016.

[63] **Ferrara, E.** (). WHAT TYPES OF COVID-19 CONSPIRACIES ARE POPULATED BY TWITTER BOTS? Technical report.

[64] **Ferrara, E.**, **O. Varol**, **C. Davis**, **F. Menczer**, and **A. Flammini** (2016). The rise of social bots. *Communications of the ACM*, **59**(7), 96–104.

[65] **Field, A.**, **G. Bhat**, and **Y. Tsvetkov** (2019). Contextual Affective Analysis : A Case Study of People Portrayals in Online # MeToo Stories. (Icwsm).

[66] **Freelon, D.**, **C. D. McIlwain**, and **M. D. Clark** (2017). Beyond the Hashtags: #Ferguson, #Blacklivesmatter, and the Online Struggle for Offline Justice. *SSRN Electronic Journal*.

[67] **Freeman, L.** (). Centrality in social networks conceptual clarification. *Social Network*, **1**, 215.

[68] **Gagliardone, I.**, **D. Gal**, **T. Alves**, and **G. Martinez**, *Countering online hate speech*. Unesco Publishing, 2015.

[69] **Gallagher, R. J.**, **M. R. Frank**, **L. Mitchell**, **A. J. Schwartz**, **A. J. Reagan**, **C. M. Danforth**, and **P. S. Dodds** (2020). Generalized word shift graphs: A method for visualizing and explaining pairwise comparisons between texts. *CoRR*, **abs/2008.02250**. URL `https://arxiv.org/abs/2008.02250`.

[70] **Gallagher, R. J.**, **A. J. Reagan**, **C. M. Danforth**, and **P. S. Dodds** (2018). Divergent discourse between protests and counter-protests: #blacklivesmatter and #alllivesmatter. *PLOS ONE*, **13**(4), 1–23.

[71] **Gallagher, R. J.**, **A. J. Reagan**, **C. M. Danforth**, and **P. S. Dodds** (2018). Divergent discourse between protests and counter-protests: #BlackLivesMatter and #AllLivesMatter. *PLoS ONE*, **13**(4).

[72] **Garimella, K.**, **G. D. F. Morales**, **A. Gionis**, and **M. Mathioudakis** (2018). Quantifying Controversy on Social Media. *ACM Transactions on Social Computing*, **1**(1), 1–27.

[73] **Garimella, K.**, **G. D. F. Morales**, **A. Gionis**, and **M. Mathioudakis** (2018). Quantifying controversy on social media. *Trans. Soc. Comput.*, **1**(1).

[74] **Gerbaudo, P.** and **E. Treré** (2015). In search of the 'we' of social media activism: introduction to the special issue on social media and protest identities. *Information Communication and Society*, **18**(8), 865–871.

[75] **Germani, F.** and **N. Biller-Andorno** (2021). The anti-vaccination infodemic on social media: A behavioral analysis. *PLoS ONE*, **16**(3 March), 1–14.

[76] **González-Bailón, S.**, **J. Borge-Holthoefer**, **A. Rivero**, and **Y. Moreno** (2011). The dynamics of protest recruitment through an online network. *Scientific Reports*, **1**, 1–7.

[77] **Goode, B. J.**, **S. Krishnan**, **M. Roan**, and **N. Ramakrishnan** (2015). Pricing a protest: Forecasting the dynamics of civil unrest activity in social media. *PLoS ONE*, **10**(10), 1–25.

[78] **Gorrell, G.**, **M. E. Bakir**, **I. Roberts**, **M. A. Greenwood**, **B. Iavarone**, and **K. Bontcheva** (2019). Partisanship, propaganda and post-truth politics: Quantifying impact in online debate. *The Journal of Web Science*, **7**.

[79] **Grčar, M.**, **D. Cherepnalkoski**, **I. Mozetič**, and **P. Kralj Novak** (2017). Stance and influence of twitter users regarding the brexit referendum. *Computational social networks*, **4**(1), 1–25.

[80] **Grimminger, L.** and **R. Klinger**, Hate towards the political opponent: A twitter corpus study of the 2020 us elections on the basis of offensive speech and stance detection. *In Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. 2021.

[81] **Grulcan Elmas, T.**, **R. Overdorf**, and **K. A. Epfl** (2021). Tactical Reframing of Online Disinformation Campaigns Against The Istanbul Convention. Technical report.

[82] **Guermazi, R.**, **M. Hammami**, and **A. B. Hamadou**, Using a semi-automatic keyword dictionary for improving violent web site filtering. *In 2007 Third International IEEE Conference on Signal-Image Technologies and Internet-Based System*. IEEE, 2007.

[83] **Gürel, A.** and **E. Emin**, Alem at case 2021 task 1: Multilingual text classification on news articles. *In Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*. 2021.

[84] **Gurukar, S.**, **D. Ajwani**, **S. Dutta**, **J. Lauri**, **S. Parthasarathy**, and **A. Sala**, Towards quantifying the distance between opinions. *In ICWSM*. 2020.

[85] **Haider, S.**, **L. Luceri**, **A. Deb**, **A. Badawy**, **N. Peng**, and **E. Ferrara** (2020). Detecting Social Media Manipulation in Low-Resource Languages. URL `http://arxiv.org/abs/2011.05367`.

[86] **Haidt, J.**, Moral Psychology and the Misunderstanding of Religion1. *In The Believing Primate: Scientific, Philosophical, and Theological Reflections on the Origin of Religion*. 2011.

[87] **Hari, K. S.**, **D. Aravind**, **A. Singh**, and **B. Das**, Detecting propaganda in trending twitter topics in india—a metric driven approach. *In Emerging Technologies in Data Mining and Information Security*. Springer, 2021, 657–671.

[88] **He, B.**, **C. Ziems**, **S. Soni**, **N. Ramakrishnan**, **D. Yang**, and **S. Kumar**, Racism is a virus: Anti-asian hate and counterspeech in social media during the covid-19 crisis. ASONAM '21. Association for Computing Machinery, New York, NY, USA, 2022.

[89] **Hettiarachchi, H.**, **M. Adedoyin-Olowe**, **J. Bhogal**, and **M. M. Gaber**, Daai at case 2021 task 1: Transformer-based multilingual socio-political and crisis event detection. *In Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*. 2021.

[90] **Hogenboom, F.**, **F. Frasincar**, **U. Kaymak**, and **F. De Jong** (2011). An overview of event extraction from text. *DeRiVE@ ISWC*, 48–57.

[91] **Horawalavithana, S.**, **K. W. Ng**, and **A. Iamnitchi**, Drivers of Polarized Discussions on Twitter during Venezuela Political Crisis. *In ACM International Conference Proceeding Series*. Association for Computing Machinery, 2021.

[92] **Howard, P. N.** and **B. Kollanyi** (2016). Bots,# strongerin, and# brexit: computational propaganda during the uk-eu referendum. *arXiv preprint arXiv:1606.06356*.

[93] **Hristakieva, K.**, **S. Cresci**, **G. Da San Martino**, **M. Conti**, and **P. Nakov**, The spread of propaganda by coordinated communities on social media. *In 14th ACM Web Science Conference 2022*, WebSci '22. Association for Computing Machinery, New York, NY, USA, 2022.

[94] **Hristakieva, K.**, **S. Cresci**, **G. Da San Martino**, **M. Conti**, and **P. Nakov**, The spread of propaganda by coordinated communities on social media. *In 14th ACM Web Science Conference 2022*. 2022.

[95] **Hristakieva, K.**, **S. Cresci**, **G. Da San Martino**, **M. Conti**, and **P. Nakov**, The Spread of Propaganda by Coordinated Communities on Social Media. Association for Computing Machinery (ACM), 2022.

[96] **Hürriyetoğlu, A.**, **O. Mutlu**, **F. Duruşan**, **O. Uca**, **A. S. Gürel**, **B. Radford**, **Y. Dai**, **H. Hettiarachchi**, **N. Stoehr**, **T. Nomoto**, **M. Slavcheva**, **F. Vargas**, **A. Javid**, **F. Beyhan**, and **E. Yörük**, Extended multilingual protest news detection - shared task 1, case 2021 and 2022. *In Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2022)*. Association for Computational Linguistics (ACL), online, 2022.

[97] **Hürriyetoğlu, A.**, **O. Mutlu**, **F. F. Liza**, **E. Yörük**, **R. Kumar**, and **S. Ratan**, Multilingual protest news detection - shared task 1, case 2021. *In Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction*

of Socio-political Events from Text (CASE 2021). Association for Computational Linguistics (ACL), online, 2021.

[98] **Hürriyetoğlu, A.**, **H. Tanev**, **V. Zavarella**, **R. Yeniterzi**, **O. Mutlu**, and **E. Yörük**, Challenges and applications of automated extraction of socio-political events from text (case 2022): Workshop and shared task report. *In Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2022)*. Association for Computational Linguistics (ACL), online, 2022.

[99] **Hussain, M. N.**, **K. K. Bandeli**, **S. Al-khateeb**, and **N. Agarwal**, Analyzing shift in narratives regarding migrants in europe via blogosphere. 2018.

[100] **IANS** (2019). Bhopal: Hundreds join BJP's pro CAA-NRC rally. *freepressjournal*.

[101] **IANS** (2019). Anti-CAA protests continue to hit rail, road traffic in West Bengal. *freepressjournal*.

[102] **IANS** (2019). Delhi's Shaheen Bagh rings in new year with anti-Citizenship Act slogans. *The Hindu*.

[103] **India Today** (2020). Sushant Singh Rajput dies by suicide at 34 in Mumbai. [Online; accessed 14-June-2020].

[104] **Jackson, S. J.** and **S. Banaszczyk** (2016). Digital standpoints: Debating gendered violence and racial exclusions in the feminist counterpublic. *Journal of Communication Inquiry*, **40**(4), 391–407.

[105] **Jain, A. K.**, **M. N. Murty**, and **P. J. Flynn** (1999). Data clustering: a review. *ACM computing surveys (CSUR)*, **31**(3), 264–323.

[106] **Jakesch, M.**, **K. Garimella**, **D. Eckles**, and **M. Naaman** (2021). Trend alert: A cross-platform organization manipulated twitter trends in the indian general election. *Proc. ACM Hum.-Comput. Interact.*, **5**(CSCW2).

[107] **Kapoor, A.**, **M. Dhawan**, **A. Goel**, **A. T H**, **A. Bhatnagar**, **V. Agrawal**, **A. Agrawal**, **A. Bhattacharya**, **P. Kumaraguru**, and **A. Modi**, HLDC: Hindi legal documents corpus. *In Findings of the Association for Computational Linguistics: ACL 2022*. Association for Computational Linguistics, Dublin, Ireland, 2022.

[108] **Karkın, N.**, **N. Yavuz**, **İ. Parlak**, and **Ö. Ö. İkiz**, Twitter use by politicians during social uprisings: an analysis of gezi park protests in turkey. *In Proceedings of the 16th Annual International Conference on Digital Government Research*. 2015.

[109] **Khatua, A.** and **A. Khatua**, Leave or remain? deciphering brexit deliberations on twitter. *In 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*. 2016.

[110] **Korolov, R.**, **D. Lu**, **J. Wang**, **G. Zhou**, **C. Bonial**, **C. Voss**, **L. Kaplan**, **W. Wallace**, **J. Han**, and **H. Ji**, On predicting social unrest using social media. *In 2016 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM)*. IEEE, 2016.

[111] **Krasnowska-Kieraś, K.** and **A. Wróblewska**, Empirical linguistic study of sentence embeddings. *In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019.

[112] **Kudo, T.** and **J. Richardson**, Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 2018.

[113] **Kuo, C.-C.** and **K.-Y. Chen** (2022). Toward zero-shot and zero-resource multilingual question answering. *IEEE Access*.

[114] **Leetaru, K.** and **P. A. Schrodt**, Gdelt: Global data on events, location, and tone, 1979–2012. *In ISA annual convention*, volume 2. Citeseer, 2013.

[115] **Li, Q.**, **H. Ji**, and **L. Huang**, Joint event extraction via structured prediction with global features. *In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2013.

[116] **Liu, F.**, **D. Ford**, **C. Parnin**, and **L. Dabbish**, Selfies as social movements: Influences on participation and perceived impact on stereotypes. volume 1. Association for Computing Machinery, New York, NY, USA, 2017.

[117] **Liu, Q. H.**, **F. M. Lü**, **Q. Zhang**, **M. Tang**, and **T. Zhou** (2018). Impacts of opinion leaders on social contagions. *Chaos*, **28**(5).

[118] **Lotan, G.**, **E. Graeff**, **M. Ananny**, **D. Gaffney**, **I. Pearce**, *et al.* (2011). The arab spring : the revolutions were tweeted: Information flows during the 2011 tunisian and egyptian revolutions. *International journal of communication*, **5**, 31.

[119] **Luceri, L.**, **A. Badawy**, **A. Deb**, and **E. Ferrara**, Red bots do it better: Comparative analysis of social bot partisan behavior. *In The Web Conference 2019 - Companion of the World Wide Web Conference, WWW 2019*. Association for Computing Machinery, Inc, 2019.

[120] **Luceri, L.**, **S. Giordano**, and **E. Ferrara** (2020). Detecting Troll Behavior via Inverse Reinforcement Learning: A Case Study of Russian Trolls in the 2016 US Election. Technical report.

[121] **Luo, X.**, **S. Kato**, **A. Obata**, **B. Ahsan**, **R. Okada**, and **T. Nakanishi**, A joint scene text recognition and visual appearance model for protest issue classification. *In Proceedings of the 4th ACM Workshop on Intelligent Cross-Data Analysis and Retrieval*, ICDAR '23. Association for Computing Machinery, New York, NY, USA, 2023. ISBN 9798400701863.

[122] **Lynn, V.**, **S. Giorgi**, **N. Balasubramanian**, and **H. A. Schwartz**, Tweet classification without the tweet: An empirical examination of user versus document attributes. *In Proceedings of the Third Workshop on Natural Language Processing and Computational Social Science*. 2019.

[123] **Magdy, W.**, **K. Darwish**, **N. Abokhodair**, **A. Rahimi**, and **T. Baldwin**, # isisisnotislam or# deportallmuslims? predicting unspoken views. *In Proceedings of the 8th ACM Conference on Web Science*. 2016.

[124] **Mahapatra, S.** and **K. Garimella**, Digital public activism and the redefinition of citizenship: The movement against the citizenship (amendment) act of india. *In Weizenbaum Conference 2021: Democracy in Flux–Order, Dynamics and Voices in Digital Public Spheres*. DEU, 2021.

[125] **Maity, K.** and **S. Saha**, A multi-task model for sentiment aided cyberbullying detection in code-mixed indian languages. *In* **T. Mantoro**, **M. Lee**, **M. A. Ayu**, **K. W. Wong**, and **A. N. Hidayanto** (eds.), *Neural Information Processing*. Springer International Publishing, Cham, 2021.

[126] **Marcoux, T.** and **N. Agarwal**, Narrative trends of covid-19 misinformation. *In Text2Story@ ECIR*. 2021.

[127] **Marwell, G.** and **P. Oliver**, *The critical mass in collective action*. Cambridge University Press, 1993.

[128] **Mathew, B.**, **R. Dutt**, **P. Goyal**, and **A. Mukherjee**, Spread of hate speech in online social media. *In Proceedings of the 10th ACM conference on web science*. 2019.

[129] **Mathew, B.**, **A. Illendula**, **P. Saha**, **S. Sarkar**, **P. Goyal**, and **A. Mukherjee** (2020). Hate begets hate: A temporal study of hate speech. *Proc. ACM Hum.-Comput. Interact.*, **4**(CSCW2).

[130] **Mathur, P.**, **R. Sawhney**, **M. Ayyar**, and **R. Shah** (2019). Did you offend me? Classification of Offensive Tweets in Hinglish Language, 138–148.

[131] **McCarthy, J. D.** and **M. N. Zald** (1977). Resource mobilization and social movements: A partial theory. *American journal of sociology*, **82**(6), 1212–1241.

[132] **McInnes, L.** and **J. Healy**, Accelerated hierarchical density based clustering. *In 2017 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2017.

[133] **McInnes, L.**, **J. Healy**, and **J. Melville** (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

[134] **Misra, I.**, **A. Shrivastava**, **A. Gupta**, and **M. Hebert** (2016). Cross-stitch networks for multi-task learning, 3994–4003.

[135] **Mohammad, S.** and **P. Turney**, Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. *In Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*. Association for Computational Linguistics, Los Angeles, CA, 2010. URL `https://aclanthology.org/W10-0204`.

[136] **Murakami, A.** and **R. Raymond**, Support or oppose? classifying positions in online debates from reply activities and opinion expressions. *In Coling 2010: Posters*. 2010.

[137] **Muthiah, S.**, **P. Butler**, **R. P. Khandpur**, **P. Saraf**, **N. Self**, **A. Rozovskaya**, **L. Zhao**, **J. Cadena**, **C. T. Lu**, **A. Vullikanti**, **A. Marathe**, **K. Summers**,

G. Katz, **A. Doyle**, **J. Arredondo**, **D. K. Gupta**, **D. Mares**, and **N. Ramakrishnan** (2016). EMBERS at 4 years: Experiences operating an open source indicators forecasting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, **13-17-Augu**, 205–214.

[138] **Muthiah, S.**, **P. Butler**, **R. P. Khandpur**, **P. Saraf**, **N. Self**, **A. Rozovskaya**, **L. Zhao**, **J. Cadena**, **C.-T. Lu**, **A. Vullikanti**, *et al.*, Embers at 4 years: Experiences operating an open source indicators forecasting system. *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016.

[139] **Muthiah, S.**, **B. Huang**, **J. Arredondo**, **D. Mares**, **L. Getoor**, **G. Katz**, and **N. Ramakrishnan** (2015). Planned protest modeling in news and social media. *Proceedings of the National Conference on Artificial Intelligence*, **5**, 3920–3927.

[140] **Neha, K.**, **V. Agrawal**, **A. B. Buduru**, and **P. Kumaraguru**, The pursuit of being heard: An unsupervised approach to narrative detection in online protest. *In 2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. 2022.

[141] **Neha, K.**, **V. Agrawal**, **V. Kumar**, **T. Mohan**, **A. Chopra**, **A. B. Buduru**, **R. Sharma**, and **P. Kumaraguru**, A tale of two sides: Study of protesters and counter-protesters on #citizenshipamendmentact campaign on twitter. *In 14th ACM Web Science Conference 2022*, WebSci '22. Association for Computing Machinery, New York, NY, USA, 2022.

[142] **Neha, K.**, **T. Mohan**, **A. B. Buduru**, and **P. Kumaraguru**, Truth and travesty intertwined: a case study of# ssr counterpublic campaign. *In Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. 2021.

[143] **Newman, M. E.** and **M. Girvan** (2004). Finding and evaluating community structure in networks. *Physical review E*, **69**(2), 026113.

[144] **Ng, L. H. X.** and **K. M. Carley**, Online Coordination: Methods and Comparative Case Studies of Coordinated Groups across Four Events in the United States. Association for Computing Machinery (ACM), 2022.

[145] **Ng, L. H. X.** and **K. M. Carley**, Online coordination: methods and comparative case studies of coordinated groups across four events in the united states. *In 14th ACM Web Science Conference 2022*. 2022.

[146] **Nizzoli, L.**, **S. Tardelli**, **M. Avvenuti**, **S. Cresci**, and **M. Tesconi**, Coordinated behavior on social media in 2019 uk general election. *In Proceedings of the International AAAI Conference on Web and Social Media*, volume 15. 2021.

[147] **Ortiz, S. N.**, **L. N. Forrest**, **T. J. Fisher**, **M. Hughes**, and **A. R. Smith** (2019). Changes in internet suicide search volumes following celebrity suicides. *Cyberpsychology, Behavior, and Social Networking*, **22**(6), 373–380.

[148] **Pacheco, D.**, **A. Flammini**, and **F. Menczer**, Unveiling Coordinated Groups behind White Helmets Disinformation. *In The Web Conference 2020 - Companion of the World Wide Web Conference, WWW 2020*. Association for Computing Machinery, 2020.

[149] **Pacheco, D.**, **P.-M. Hui**, **C. Torres-Lugo**, **B. T. Truong**, **A. Flammini**, and **F. Menczer** (2020). Uncovering Coordinated Networks on Social Media: Methods and Case Studies. URL `http://arxiv.org/abs/2001.05658`.

[150] **Pacheco, D.**, **P.-M. Hui**, **C. Torres-Lugo**, **B. T. Truong**, **A. Flammini**, and **F. Menczer**, Uncovering coordinated networks on social media: Methods and case studies. *In Proceedings of the International AAAI Conference on Web and Social Media*, volume 15. 2021.

[151] **Panda, A.**, **R. Kommiya Mothilal**, **M. Choudhury**, **K. Bali**, and **J. 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), 1–14.

[152] **Pei, S.**, **L. Muchnik**, **J. S. Andrade, Jr**, **Z. Zheng**, and **H. A. Makse** (2014). Searching for superspreaders of information in real-world social media. *Scientific reports*, **4**(1), 5547.

[153] **Persily, N.** (2017). The 2016 us election: Can democracy survive the internet? *Journal of democracy*, **28**(2), 63–76.

[154] **Pond, P.** and **J. Lewis** (2019). Riots and Twitter: connective politics, social media and framing discourses in the digital public sphere. *Information Communication and Society*, **22**(2), 213–231.

[155] **Qian, J.**, **M. ElSherief**, **E. Belding**, and **W. Y. Wang**, Leveraging intra-user and inter-user representation learning for automated hate speech detection. *In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. 2018.

[156] **Raleigh, C.**, **A. Linke**, **H. Hegre**, and **J. Karlsen** (2010). Introducing acled: an armed conflict location and event dataset: special data feature. *Journal of peace research*, **47**(5), 651–660.

[157] **Ramakrishnan, N.**, **P. Butler**, **S. Muthiah**, **N. Self**, **R. Khandpur**, **P. Saraf**, **W. Wang**, **J. Cadena**, **A. Vullikanti**, **G. Korkmaz**, **C. Kuhlman**, **A. Marathe**, **L. Zhao**, **T. Hua**, **F. Chen**, **C. T. Lu**, **B. Huang**, **A. Srinivasan**, **K. Trinh**, **L. Getoor**, **G. Katz**, **A. Doyle**, **C. Ackermann**, **I. Zavorin**, **J. Ford**, **K. Summers**, **Y. Fayed**, **J. Arredondo**, **D. Gupta**, and **D. Mares** (2014). 'Beating the news' with EMBERS: Forecasting civil unrest using open source indicators. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (July 2015), 1799–1808.

[158] **Ranade, P.**, **S. Dey**, **A. Joshi**, and **T. Finin** (2022). Computational Understanding of Narratives: A Survey.

[159] **Rashed, A.**, **M. Kutlu**, **K. Darwish**, **T. Elsayed**, and **C. Bayrak**, Embeddings-based clustering for target specific stances: The case of a polarized turkey. *In Proceedings of the International AAAI Conference on Web and Social Media*, volume 15. 2021.

[160] **Raynauld, V.**, **E. Richez**, and **K. B. Morris** (2018). Canada is #idlenomore: exploring dynamics of indigenous political and civic protest in the twitterverse. *Information, Communication & Society*, **21**(4), 626–642.

[161] **Rehurek, R.**, **P. Sojka**, *et al.* (2011). Gensim—statistical semantics in python. *Retrieved from genism. org*.

[162] **Reichert, F.**, **A. T. N. Au**, and **A. J. Fiedler** (2024). How collective demands strengthen sympathy for normative and non-normative protest action: The example of hong kong's anti-extradition law amendment bill protests. *Sociology Compass*, **18**(1), e13169.

[163] **Reimers, N.** and **I. Gurevych** (2020). Making monolingual sentence embeddings multilingual using knowledge distillation, 4512–4525.

[164] **Rezapour, R.**, **P. Ferronato**, and **J. Diesner**, How do moral values difer in Tweets on social movements? *In Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*. Association for Computing Machinery, 2019.

[165] **Ribeiro, M. H.**, **P. H. Calais**, **Y. A. Santos**, **V. A. F. Almeida**, and **W. Meira** (2018). Characterizing and Detecting Hateful Users on Twitter. Technical report.

[166] **Rogers, A.**, **O. Kovaleva**, and **A. Rumshisky** (2019). Calls to Action on Social Media: Detection, Social Impact, and Censorship Potential, 36–44.

[167] **Rogers, A.**, **O. Kovaleva**, and **A. Rumshisky** (2019). Calls to action on social media: Potential for censorship and social impact. *EMNLP-IJCNLP 2019*, 36.

[168] **Ruder, S.** (2017). An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

[169] **Saha, P.**, **B. Mathew**, **K. Garimella**, and **A. Mukherjee**, "short is the road that leads from fear to hate": Fear speech in indian whatsapp groups. *In Proceedings of the Web Conference 2021*, WWW '21. Association for Computing Machinery, New York, NY, USA, 2021.

[170] **Salminen, J.**, **H. Almerekhi**, **M. Milenković**, **S.-g. Jung**, **J. An**, **H. Kwak**, and **B. J. Jansen**, Anatomy of online hate: developing a taxonomy and machine learning models for identifying and classifying hate in online news media. *In Twelfth International AAAI Conference on Web and Social Media*. 2018.

[171] **Sane, S. R.**, **S. Tripathi**, **K. R. Sane**, and **R. Mamidi**, Stance detection in code-mixed Hindi-English social media data using multi-task learning. *In Proceedings of the Tenth Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Association for Computational Linguistics, Minneapolis, USA, 2019.

[172] **Santini, R. M.**, **D. Salles**, and **G. Tucci** (2021). When machine behavior targets future voters : The use of social bots to test narratives for political campaigns in Brazil. *International Journal of Communication*, **15**, 1220–1243.

[173] **Saroj, A.** and **S. Pal** (2020). An Indian Language Social Media Collection for Hate and Offensive Speech. Technical report. URL `https://www.aclweb.org/anthology/S19-2007/`.

[174] **Schmidt, A.** and **M. Wiegand**, A survey on hate speech detection using natural language processing. *In Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, April 3, 2017, Valencia, Spain*. Association for Computational Linguistics, 2019.

[175] **Schrodt, P. A.**, **J. Beieler**, and **M. Idris**, Three'sa charm?: Open event data coding with el: Diablo, petrarch, and the open event data alliance. *In ISA Annual Convention*. Citeseer, 2014.

[176] **Schubert, E.**, **J. Sander**, **M. Ester**, **H. P. Kriegel**, and **X. Xu** (2017). Dbscan revisited, revisited: why and how you should (still) use dbscan. *ACM Transactions on Database Systems (TODS)*, **42**(3), 1–21.

[177] **Serrano, M. Á.**, **M. Boguná**, and **A. Vespignani** (2009). Extracting the multiscale backbone of complex weighted networks. *Proceedings of the national academy of sciences*, **106**(16), 6483–6488.

[178] **Service, E. N.** (2019). Government open to suggestions to implement Citizenship Act, says MHA amid raging protests. *newindianexpress*.

[179] **Shao, C.**, **G. L. Ciampaglia**, **O. Varol**, **K.-C. Yang**, **A. Flammini**, and **F. Menczer** (2018). The spread of low-credibility content by social bots. *Nature communications*, **9**(1), 1–9.

[180] **Sharma, K.**, **Y. Zhang**, **E. Ferrara**, and **Y. Liu**, Identifying coordinated accounts on social media through hidden influence and group behaviours. *In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 2021.

[181] **Shen, F.**, **C. Xia**, and **M. Skoric** (2020). Examining the roles of social media and alternative media in social movement participation: A study of hong kong's umbrella movement. *Telematics and Informatics*, **47**, 101303.

[182] **Shevtsov, A.**, **C. Tzagkarakis**, **D. Antonakaki**, and **S. Ioannidis** (2022). Identification of Twitter Bots Based on an Explainable Machine Learning Framework: The US 2020 Elections Case Study. Technical report.

[183] **Simmons, E.** (2014). Grievances do matter in mobilization. *Theory and Society*, **43**(5), 513–546.

[184] **Sinpeng, A.** (2021). Hashtag activism: social media and the #freeyouth protests in thailand. *Critical Asian Studies*, **53**(2), 192–205.

[185] **Søgaard, A.** and **Y. Goldberg**, Deep multi-task learning with low level tasks supervised at lower layers. *In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Berlin, Germany, 2016.

[186] **Solovev, K.** and **N. Pröllochs**, Hate speech in the political discourse on social media: Disparities across parties, gender, and ethnicity. *In Proceedings of the ACM Web Conference 2022*, WWW '22. Association for Computing Machinery, New York, NY, USA, 2022.

[187] **Soni, J.** (2019). CAA protest row: Section 144 imposed in UP Mau's Hajipura Chowk area. *freepressjournal*.

[188] **Spiro, E.** and **Y.-Y. Ahn** (eds.), *Predicting Online Extremism,Content Adopters, and Interaction Reciprocity*, volume 10047 of *Lecture Notes in Computer Science*. Springer International Publishing, Cham, 2016.

[189] **Sree Hari, K.**, **D. Aravind**, **A. Singh**, and **B. Das**, Detecting Propaganda in Trending Twitter Topics in India—A Metric Driven Approach. 2021, 657–671.

[190] **Starbird, K.** and **L. Palen**, (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*, CSCW '12. Association for Computing Machinery, New York, NY, USA, 2012.

[191] **Stella, M.**, **E. Ferrara**, and **M. De Domenico** (2018). Bots increase exposure to negative and inflammatory content in online social systems. *Proceedings of the National Academy of Sciences*, **115**(49), 12435–12440.

[192] **Theocharis, Y.**, **W. Lowe**, **J. W. Van Deth**, and **G. García-Albacete** (2015). Using twitter to mobilize protest action: online mobilization patterns and action repertoires in the occupy wall street, indignados, and aganaktismenoi movements. *Information, Communication & Society*, **18**(2), 202–220.

[193] **Times, H.** (2020). Sushant's death embroiled in a web of theories. [Online; accessed 03-August-2020].

[194] **Tufekci, Z.**, Big questions for social media big data: Representativeness, validity and other methodological pitfalls. *In Eighth international AAAI conference on weblogs and social media*. 2014.

[195] **Ueda, M.**, **K. Mori**, **T. Matsubayashi**, and **Y. Sawada** (2017). Tweeting celebrity suicides: Users' reaction to prominent suicide deaths on Twitter and subsequent increases in actual suicides. *Social Science & Medicine*, **189**, 158–166.

[196] **Uyheng, J.**, **D. Bellutta**, and **K. M. Carley** (2022). Bots Amplify and Redirect Hate Speech in Online Discourse About Racism During the COVID-19 Pandemic. *Social Media + Society*, **8**(3), 205630512211047.

[197] **Uyheng, J.** and **K. M. Carley**, Computational analysis of bot activity in the asia-pacific: A comparative study of four national elections. *In Proceedings of the International AAAI Conference on Web and Social Media*, volume 15. 2021.

[198] **Valecha, R.**, **S. K. Srinivasan**, **T. Volety**, **K. H. Kwon**, **M. Agrawal**, and **H. R. Rao** (2021). Fake news sharing: An investigation of threat and coping cues in the context of the zika virus. *Digital Threats*, **2**(2). URL `https://doi.org/10.1145/3410025`.

[199] **Van der Maaten, L.** and **G. Hinton** (2008). Visualizing data using t-sne. *Journal of machine learning research*, **9**(11).

[200] **Van Stekelenburg, J.** and **B. Klandermans** (2013). The social psychology of protest. *Current Sociology*, **61**(5-6), 886–905.

[201] **Vargas, L.**, **P. Emami**, and **P. Traynor**, On the detection of disinformation campaign activity with network analysis. *In Proceedings of the 2020 ACM SIGSAC Conference on Cloud Computing Security Workshop*. 2020.

[202] **Varol, O.**, **E. Ferrara**, **C. L. Ogan**, **F. Menczer**, and **A. Flammini** (2014). Evolution of online user behavior during a social upheaval. *WebSci 2014 - Proceedings of the 2014 ACM Web Science Conference*, (i), 81–90.

[203] **Varol, O.**, **E. Ferrara**, **C. L. Ogan**, **F. Menczer**, and **A. Flammini**, Evolution of online user behavior during a social upheaval. *In Proceedings of the 2014 ACM conference on Web science*. 2014.

[204] **Verleysen, M.** *et al.* (2003). Learning high-dimensional data. *Nato Science Series Sub Series III Computer And Systems Sciences*, **186**, 141–162.

[205] **Veyseh, A. P. B.**, **V. Lai**, **F. Dernoncourt**, and **T. H. Nguyen**, Unleash gpt-2 power for event detection. *In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2021.

[206] **Wang, D.**, **M. T. Amin**, **S. Li**, **T. Abdelzaher**, **L. Kaplan**, **S. Gu**, **C. Pan**, **H. Liu**, **C. C. Aggarwal**, **R. Ganti**, *et al.*, Using humans as sensors: an estimation-theoretic perspective. *In IPSN-14 proceedings of the 13th international symposium on information processing in sensor networks*. IEEE, 2014.

[207] **Wang, R.** and **K.-H. Chu** (2019). Networked publics and the organizing of collective action on Twitter: Examining the #Freebassel campaign. *Convergence*, **25**(3), 393–408.

[208] **Wang, R.** and **A. Zhou** (2021). Hashtag activism and connective action: A case study of #HongKongPoliceBrutality. *Telematics and Informatics*, **61**.

[209] **Wang, R.** and **A. Zhou** (2021). Hashtag activism and connective action: A case study of# hongkongpolicebrutality. *Telematics and Informatics*, **61**, 101600.

[210] **Watch, H. R.** (2019). India: Citizenship Bill Discriminates Against Muslims. *Human Right Watch*.

[211] **Web, F.** (2019). CAA protests: Why are students protesting? Here's all you need to know.

[212] **Wei, K.**, **Y.-R. Lin**, and **M. Yan**, Examining protest as an intervention to reduce online prejudice: A case study of prejudice against immigrants. *In Proceedings of The Web Conference 2020*, WWW '20. Association for Computing Machinery, New York, NY, USA, 2020.

[213] **Williams, M. L.**, **P. Burnap**, **A. Javed**, **H. Liu**, and **S. Ozalp** (2020). Hate in the machine: Anti-black and anti-muslim social media posts as predictors of offline racially and religiously aggravated crime. *The British Journal of Criminology*, **60**(1), 93–117.

[214] **Wu, H. H.**, **R. J. Gallagher**, **T. Alshaabi**, **J. L. Adams**, **J. R. Minot**, **M. V. Arnold**, **B. F. Welles**, **R. Harp**, **P. S. Dodds**, and **C. M. Danforth** (2023). Say their names: Resurgence in the collective attention toward black victims of fatal police violence following the death of george floyd. *PLOS ONE*, **18**(1), 1–26.

[215] **Xiao, Z.**, **W. Song**, **H. Xu**, **Z. Ren**, and **Y. Sun**, Timme: Twitter ideology-detection via multi-task multi-relational embedding. *In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2020.

[216] **Xiong, Y.**, **M. Cho**, and **B. Boatwright** (2019). Hashtag activism and message frames among social movement organizations: Semantic network analysis and thematic analysis of Twitter during the #MeToo movement. *Public Relations Review*, **45**(1), 10–23.

[217] **Xu, W. W.** (2020). Mapping Connective Actions in the Global Alt-Right and Antifa Counterpublics. *International Journal of Communication*, **14**(0).

[218] **Yang, K.-C.**, **O. Varol**, **P.-M. Hui**, and **F. Menczer** (2020). Scalable and generalizable social bot detection through data selection. *Proceedings of the AAAI Conference on Artificial Intelligence*, **34**(01), 1096–1103.

[219] **Yang, K.-C.**, **O. Varol**, **P.-M. Hui**, and **F. Menczer** (2020). Scalable and Generalizable Social Bot Detection through Data Selection. Technical report.

[220] **Yaqub, U.**, **S. A. Chun**, **V. Atluri**, and **J. Vaidya** (2017). Analysis of political discourse on twitter in the context of the 2016 US presidential elections. *Government Information Quarterly*, **34**(4), 613–626.

[221] **Yardi, S.** and **D. Boyd** (2010). Dynamic Debates: An Analysis of Group Polarization Over Time on Twitter. *Bulletin of Science, Technology & Society*, **30**(5), 316–327.

[222] **Zhang, Y.**, **F. Guo**, **J. Shen**, and **J. Han**, Unsupervised key event detection from massive text corpora. *In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22. Association for Computing Machinery, New York, NY, USA, 2022.

[223] **Zhang, Z.**, *From media hype to twitter storm: news explosions and their impact on issues, crises and public opinion*, volume 36. 2019.

[224] **Zhang, Z.**, **D. Robinson**, and **J. Tepper**, Detecting hate speech on twitter using a convolution-gru based deep neural network. *In European semantic web conference*. Springer, 2018.

[225] **Zia, H. B.**, **I. Castro**, **A. Zubiaga**, and **G. Tyson**, Improving zero-shot cross-lingual hate speech detection with pseudo-label fine-tuning of transformer language models. *In Proceedings of the International AAAI Conference on Web and Social Media*, volume 16. 2022.