Analyzing the Dynamics of Hate Speech on Online Platforms

Dhwani Jakhaniya[1] and Maunendra Sankar Desarkar[1]

Indian Institute of Technology Hyderabad, 502285, India
dhwani.patel2599@gmail.com, maunendra@cse.iith.ac.in

Abstract. In the rapidly evolving landscape of online microblogging platforms, hate speech has emerged as a particularly troubling issue. Alarmingly, numerous countries have seen a sharp increase in hate crimes driven by malicious hate campaigns. While the detection of hate speech has gained attention as a major research field, the complexities of its genesis and spread across online social networks remain largely unexamined. In this study, we focus on developing a benchmark dataset for hate and abusive speech, related to various aspects of Indian politics, and religious topics. The dataset comprises of a vast user base of politically active individuals, enabling us to capture the nuances of hate speech propagation within this context. Additionally, we employ advanced topic modeling techniques to analyze the data, uncovering the underlying themes and user reactions associated with the data. This detailed analysis reveals how people respond to these topics, the temporal activation of discussions, and the comparative trends between different hashtags. This study offers valuable insights for policymakers, social media companies, and researchers, helping them develop better strategies to reduce hate speech and promote positive online interactions.

Keywords: Hate speech · Topic modelling · NLP

1 INTRODUCTION

Social media platforms have transformed the ways of our interactions and news consumption, by allowing users to freely express themselves with emotive language during debates on any topics. Unfortunately, these emotional exchanges can sometimes target specific community groups, fostering negativity or discrimination against them. These communities can be defined by various identities, such as race, religion, gender, or sexual orientation, leading to multiple levels of discrimination. The spread of hate speech can incite violence, reinforce prejudices, and undermine societal harmony. Identifying and understanding the dynamics of hate speech on these platforms is therefore a critical area of research.

This paper focuses on hate speech detection and the analysis of its dynamics on Twitter (renamed as “X" now), a platform widely used for any discussions - including political and religious discussions too. Specifically, we have collected data using Twitter API for four different hashtags related to Indian politics and

religion. These hashtags were chosen due to their relevance and the high volume of discourse they generate, helping us to study hate speech dynamics.

The primary goals of this research are two-fold: first, to detect hate speech in the collected data using various learning-based models, and second, to employ topic modeling techniques to uncover underlying themes and user reactions associated with these hashtags.

2 RELATED WORK

The existing work on hate speech in social media can be categorized broadly into two different threads: hate speech classification, and modeling hate speech diffusion. Here we discuss about work from literature on these themes.

Hate Speech Detection: Waseem and Hovy [1] collected a dataset comprising 130,000 tweets encompassing seventeen distinct terms or phrases that were considered to be hateful. Within this dataset, a total of 16,849 tweets were labeled as racism, sexism, or neither. Many tweets classified as sexist came from specific TV shows, while the racist tweets pertain to Judaism and Islam. To address potential bias in the initial dataset, the authors relabeled a subset of tweets from the original dataset, and also annotated a fresh sample of tweets by feminist and anti-racism activists having domain expertise. Davidson et al. [2] collected tweets using Hatebase, a crowdsourced dictionary of hate speech terms. Crowdworkers were told to base their judgments on the overall tweet and its inferred context, rather than specific objectionable words or phrases, to prevent false positives. This dataset of 24,783 tweets was annotated as hate speech, offensive language, or neither. HateXplain [9], is an innovative benchmark hate speech dataset where each post in the dataset is carefully annotated from three different angles, offering a thorough understanding of the nature of hate. Annotation begins with identifying posts as hate, offensive, or regular speech - providing a basic knowledge of the type and intensity of language used in each post. They also look deeper into the target community, which has been affected by hate speech or inappropriate language in the message, to obtain insight into the many social groups affected by hate speech occurrences. Furthermore, it proposed the annotation of rationales, which are specific sections of the post that are used to determine if a post is hateful, offensive, or normal.

In [8], the authors show how Latent Dirichlet Allocation (LDA) may be applied to analyze online content that is shared across social media platforms by extremist communities, including Facebook, Gab, Telegram, and VK. They demonstrated that the use of simple unsupervised topic model architecture, such as LDA, can yield important insights into the online hate ecosystem, including the style of narratives and how different platforms are used.

Modeling Hate Speech Diffusion: The work in [6] used temporal analytics to record snapshots of Gab.com over time and investigate how hate speech evolves in this setting. To simulate the growth of hate speech, they used the DeGroot model, a framework commonly used in social network analysis to investigate information spread and opinion dynamics. In [5], the authors propose

a graph based approach for modeling hate speech diffusion. They observe that text based approach work better, but the proposed grpah based approach can also find complementary information regarding the spread of hate content. In [4], the authors study the role of *echo chambers*, where groups of users participate in creating large cascades of hate content.

In this work, we create a dataset containing social media posts mostly on topics related to sentiments or incidents related to the Indian Population. After initial manual annotation, we rely on semi-supervised approach to get the binary hate/non-hate labels for instances in the dataset. We then use these labels for analyzing the spread of hate for the underlying topics/discussion themes. We also try to analyze and reason about the dynamics of hate spread for the themes. We further try to identify sub-topics in each of these discussion themes, to analyze the nature of the hate inside a discussion theme containing large hateful content.

3 DATA COLLECTION

The data set used for our study was collected from Twitter Using Twitter's official API, within the duration of Feb 2021 to Oct 2022. The resulting dataset comprises 268,838 tweets, collected using some of the most trending hashtags during this period, specifically `#Gyanvapi, #Andhbhakt, #IndianMuslims`, and `#GoBackModi`. These hashtags were selected as they were used in large number of posts during the said time frame, and were also controversial and contained split opinions. For each instance of the dataset, the following pieces of information were obtained: `author_id, author_description, username, text, count of author_followers, author_tweets, author_location, retweets, created_at, likes`.

Filtering and annotation: We first performed standard preprocessing on the data involving the removal of duplicates, noise, and stop words, as well as word lemmatization, among other techniques. After preprocessing, the dataset contained 102,649 rows, each maintaining the original 12 columns mentioned above. We manually annotated 600 tweets for each hashtag, categorizing them as NON-HATE (0) or HATE (1). This annotation process involved a thorough exploration of the data and multiple iterations to ensure the accuracy and reliability of the labels assigned.

Need for such a dataset: The publicly available datasets typically focus on content that is from outside India. The dataset created as part of this research is primarily focused on Indian context. More detailed information on this dataset, broken down by hashtag, is provided in Table 1.

4 EXPERIMENTAL STUDY AND RESULTS

In this section, we describe the models used for classifying our hate speech dataset. We began by finetuning pre-trained models to optimize their performance for hate speech detection. To augment the data, we employed a semi-supervised approach, generating pseudo labels with the trained model. Final

Table 1. Dataset Description

Hashtags	Tweets	Labelled Tweets	Classes
#Andhbhakt	24462	600	Non-Hate : 76.16%, Hate : 23.83%
#GoBackModi	18671	600	Non-Hate : 87.66%, Hate : 12.33%
#GyanVapi	20697	600	Non-Hate : 93.5%, Hate : 6.50%
#Indian_Muslims	57254	600	Non-Hate : 66.42%, Hate : 33.58%

labels were determined using a voting algorithm that combined predictions from multiple models for improved reliability. We also conducted time series analysis to explore the temporal patterns of hate speech and applied the Latent Dirichlet Allocation (LDA) algorithm to detect underlying topics in hateful tweets.

4.1 Hate Speech Classification Models

To classify hate speech, we utilized three state-of-the-art transformer-based models: BERT, Distil BERT, and XLM-Roberta. These models have demonstrated significant effectiveness in natural language processing tasks. We can see that this dataset is imbalanced from Table 1. We use oversampling to balance the dataset before training. The hyperparameters that were set to common values across the methods are: MAX_LEN = 512, TRAIN_BATCH_SIZE = 16, EPOCHS = 5, LEARNING_RATE = 1e-05, VAL_BATCH_SIZE = 8.Table 2 presents the performance of each of these models for the hate speech classification task, using 5-fold cross-validation.

Table 2. Classification Models Performance

Model	Classes	Accuracy	F1_score	Precision	Recall
BERT	0	0.9182	0.91	0.98	0.85
	1		0.92	0.87	0.99
DistilBERT	0	0.8664	0.92	0.90	0.96
	1		0.60	0.67	0.55
XLM-RoBERTa	0	0.9376	0.94	0.94	0.93
	1		0.94	0.93	0.94

Given the multilingual nature of our dataset, which includes tweets in various languages, we have found that the XLM-RoBERTa model achieves the highest accuracy and F1_Score among all models tested. XLM-RoBERTa, a robustly optimized BERT model pre-trained on a large-scale multilingual corpus, is particularly well-suited for multilingual contexts, making it an ideal choice for our hate speech detection task. BERT emerges as the second-best method.

4.2 Time series Analysis for Hate Diffusion

The dataset contains a feature called 'created_at', which records the timestamp of each tweet. This feature enables a detailed time series analysis to uncover

trends and patterns in hate speech over time. Our dataset contains 2400 labeled examples (tweet text) which is very less to perform Time series analysis. So to make use of all tweets we used semi-supervised approach for data augmentation. First, we pre-processed all unlabeled data by removing links, special characters, and blank cells to ensure consistency and cleanliness. Next, we generated pseudo labels for the unlabeled tweets using our trained hate speech detection models. We selected tweets with prediction probabilities greater than 0.85 and less than 0.20, indicating high confidence in their classification, and incorporated these into our training dataset. With this expanded dataset, now consisting of 89,000 samples, we retrained our models to enhance their robustness. We then implemented a voting algorithm across BERT, XLM-RoBERTa, and DistilBERT models to generate the final labels for all hashtags. The accuracy on the combined dataset for XLM-Roberta, BERT, and DistilBert was 97.55%, 96.3%, and 92.7%, respectively. This approach allowed us to utilize a significantly larger dataset, improving the accuracy and reliability of our time series analysis on the temporal trends of hate speech.

The time span for the hashtags analyzed in our study varied. The hashtag #Andhbhakt was tracked from May 2021 to May 2022, covering a full year of activity. The hashtag #GoBackModi was analyzed from January 2021 to March 2021, a period before the Assembly elections in multiple Indian states, and showed periodic spikes correlating with specific political rallies, speeches, or elections. The hashtag #GyanVapi, tracked from May 2021 to May 2022, exhibited higher instances of hate speech during certain periods, corresponding to news coverage or legal proceedings related to the Gyanvapi case that concerned with the right to worship by people from different religions in different parts of a religious establishment - Gyanvapi mosque. Finally, the hashtag #IndianMuslims was monitored from June 2021 to October 2022, revealing a consistently high number of hate tweets indicative of ongoing social or political tensions involving the Indian Muslim community.

Figure 1 shows the counts of hate posts for the different hashtags, over a normlized time range. We observed that the hashtag #**Andhbhakt** is common over the time and maintains a steady count[1]. These posts are from people posting frustration and opposing individuals or groups supporting the central government or the prime minister. So, if we see daily conversation is active (No specific hike in hate on particular day) over the time in the context of political discussions or social media debates and number of tweets are less compared to other hashtags.

#**GoBackModi** hashtag displayed periodic spikes in activity, which correlated with specific political rallies, speeches, or election events. The spikes indicate heightened public sentiment and opposition towards Indian Prime Minister Mr. Narendra Modi during these times. The analysis of the hashtag #GoBackModi reveals how political events can trigger surges in hate speech, reflecting

[1] It may not be visible from the plot as the y-axis-range is set to the highest count across the four hashtags. The daily counts for the hashtag #AndhBhakt is much lesser than that peak, but is steady.

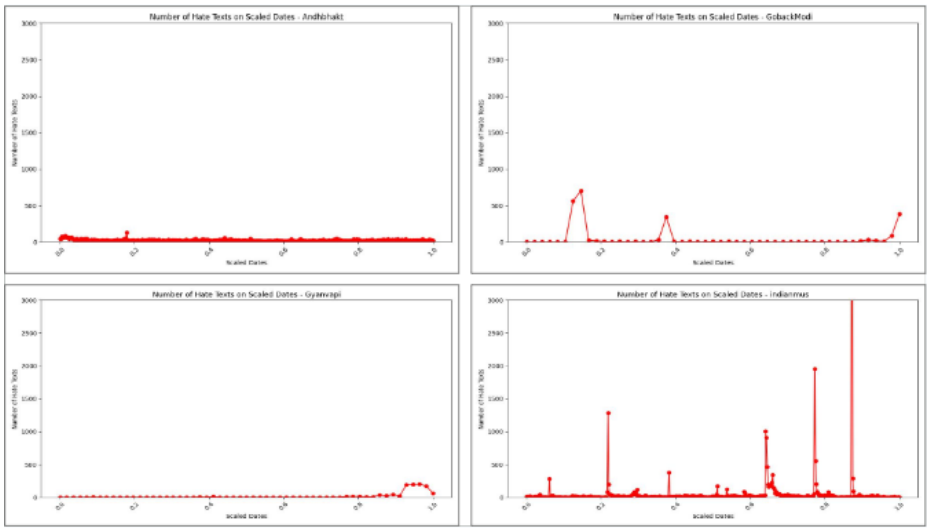

Fig. 1. Comparison of number of Hate Tweets for Different Hashtags over time in normalized 0-1 scale. From top-left in clockwise manner, we show the plots for the hashtags #AndhBhakt, #GoBackModi, #IndianMuslim, and #Gyanvapi.

the polarized nature of political discourse on social media. This hashtag served as a focal point for expressing dissent and criticism, capturing the dynamic and contentious atmosphere of the pre-election period. The reasons for spike in hate is mentioned in table 3.

Table 3. Incidents Leading to a Hike in Hate Tweets - #GoBackModi

Date	**No. of Hate Tweets**	**Reason for hike in hate**
2021-02-13 2021-02-14	561 702	Prime Minister Narendra Modi's planned visit to Tamil Nadu on February 14, 2021, ahead of the upcoming elections. Many Tamilians urged him not to visit, reflecting local opposition and ongoing campaigns related to farmers' rights and dissatisfaction with the Government's agricultural policies.[ref]
2021-02-25	342	Prime Minister Modi's visit to Tamil Nadu sparked the #GoBackModi trend due to political opposition, grievances over the NEET exam and Cauvery water issue, and general discontent with central government policies.[ref]
2021-03-30	383	PM Modi to address election rally in Puducherry on March 30 to canvass votes for the NDA.[ref]

#**Gyanvapi** hashtag displayed spikes in hate speech, particularly during time May 2022 when this survey was resumed and was conducted in this period of time. The reasons for spike in hate during May 2022 is mentioned in table 4.

For the #**IndianMuslims** hashtag, we see timeline of different Islam-related incidents overlap with the spikes in the hate speech. The number of people who posted hate content are too high compared to other hashtags making this hashtag is very prominent in India. Table 9 lists the possible reasons for spike in hate.

Table 4. Incidents Leading to a Hike in Hate Tweets - #Gyanvapi

Date	**No. of Hate Tweets**	**Reason for hike in hate**
2022-04-11	308	The survey resumed and was conducted for 2 days, with findings submitted to the court by May 17. The Supreme Court transferred the case to a district judge and appointed a senior judicial officer to handle it. ref
2022-04-12	277	
2022-06-12	497	
2022-07-29	271	

Table 5. Incidents Leading to a Hike in Hate Tweets - #IndianMuslims

Date	**No. of Hate Tweets**	**Reason for hike in hate**
2021-09-24	1285	Hindu Reporter is beating a Muslim protester in presence of police.[ref]
2022-04-11 2022-04-12	1006 906	Clashes between Hindus and Muslims during a religious festival prompted police in India to impose a curfew in one town and ban gatherings of more than four people in affected parts of Gujarat.[ref]
2022-06-12	1958	Muslims in India have taken to the streets to protest against anti-Islamic comments made by two members of Prime Minister Narendra Modi's Hindu nationalist Bhartiya Janata Party (BJP).[ref]
2022-07-29	6366	Murder of Muslim in Karnataka, primarily blaming BJP leaders and Sangh Parivar for inciting violence and discrimination against Muslims. Accusing the BJP government of discrimination against Muslims and failing to protect their lives.[ref]

Here we can see major activity in #IndianMuslim because it is very sensitive topic in india resulting in high hate tweets compared to other hashtags. Although #Gyanvapi also involves Hindu-Muslim issues, but it is less sensitive and primarily related to ongoing legal proceedings, resulting in fewer hate tweets compared to #IndianMuslim. The hashtag #GoBackModi saw spikes in hate tweets just before the elections, especially from South India (Tamil Nadu, Karnataka), and whenever PM Modi visited these regions. The hashtag #Andhbhakt is common

over time, daily conversations are active without specific spikes in hate speech, and the overall number of tweets is lower compared to other hashtags.

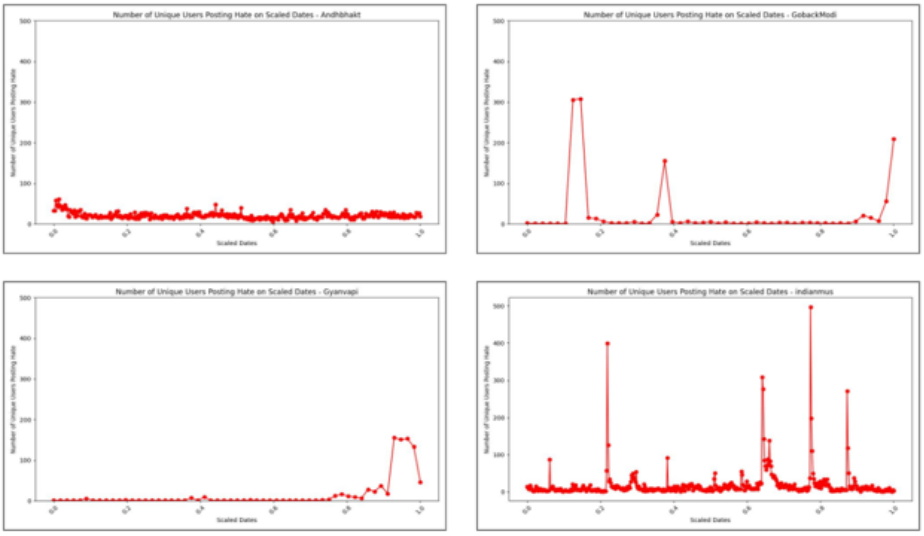

Fig. 2. Comparison of No. of Unique users posting hate tweets for Different Hashtags. Time is shown in x-axis, in normalized 0-1 scale. From top-left in clockwise manner, we show the plots for #AndhBhakt, #GoBackModi, #IndianMuslim, and #Gyanvapi.

Fig 2 shows Number of Unique users who posted hate tweets on each day for all hashtags. We can see that the hashtag #IndianMuslims attracts more unique users than the others. Daily engagement is higher for #Andhbhakt and #IndianMuslims, as these are common and consistently trending topics. Engagement for #IndianMuslims is particularly high due to its sensitive nature and spikes related to relevant events. Hashtags #GoBackModi and #Gyanvapi show higher user engagement at specific times of the day or on certain days of the week. Analysis can also reveal patterns suggesting automated activity, such as regular interval posting or high-frequency posting by users.

4.3 Topic modelling

To uncover the underlying topics in hateful tweets, we utilized topic modeling, specifically using the latent Dirichlet allocation (LDA) algorithm.

Firstly the text data underwent thorough preprocessing, including text cleaning (removing punctuation and stop words, converting text to lowercase, and applying stemming or lemmatization), tokenization into n-grams, and vectorization using the TF-IDF technique. For the LDA model training, we set hyperparameters such as chunksize = 2000, passes = 10, alpha = 'auto', iterations =

200, random_state = 100, and varied the number of topics for each hashtag to optimize the model's performance.

#Andhbhakt: We obtained the most common words, according to LDA, used in hateful comments for #Andhbhakt. The top-30 words for this topic were found out to be: *andhbhakt, bjp, people, aur, know, govt, log, hain, bjps_problem_is, leader, modi, ye, vote, tum, bhakts, state, question, toh, supporter, nation, farmer, country, party, sab, watch, money, word, ban, understand, twitter.*

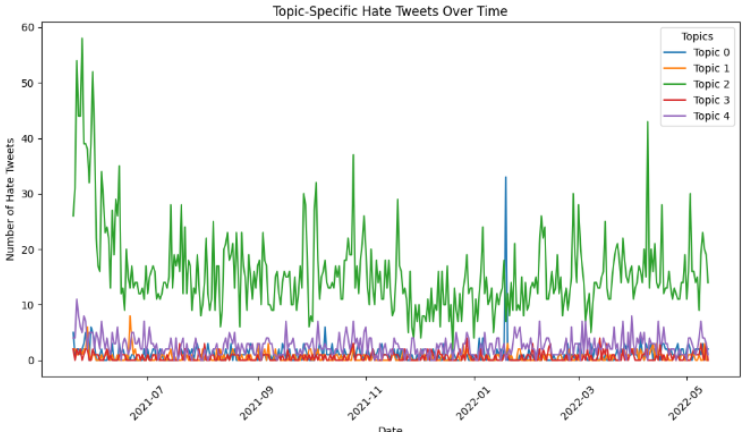

Fig. 3. Topic specific hate tweets over time - #Andhbhat

Using Intertopic Distance Map, we found that the adequate number of topics for the hashtag to be five. After determining the optimal number of topics, we plotted topic-specific hate tweets over time to identify which topics generate more hate tweets and to understand their temporal distribution in Fig 3.

Table 6. Generated topics and topic labels - #Andhbhat

Topic No.	Keywords of generated topic	Topic label
0	bjp, govt, leader, vote, bhakts, state, question, party, supporter, nation	Political Discussions and Criticism
1	farmer, watch, law, doctor, koi, cow, bhi, tag, remove, trust	Farmers and Agriculture-Related Issues
2	andhbhakt, people, know, modi, support, country, word, understand, twitter, want	Public Opinion and Social Media Commentary
3	bjps_problem_is, help, family, spot, sale, ve, bhakti, hell, din, brainwash	Criticism of BJP and Brainwashing Allegations
4	aur, log, hain, ye, tum, toh, sab, vaccine, thi, karne	General Public Sentiments and Miscellaneous Topics in transliterated/romanized hindi

We can see that Topic 2 generated the maximum number of hate tweets. This suggests that this topic is centered around political and ideological discourse, particularly involving supporters and opponents of Prime Minister Narendra Modi. The high frequency of hate tweets in this topic indicates significant polarization and hostility in discussions related to these themes. This insight highlights the contentious nature of political and religious debates on social media platforms. **#GoBackModi:** We obtained the most common words, according to LDA, used in comments for #GoBackModi.

The top-30 words for this topic were found out to be: *trend, modi, gobackmodi, farmer, government pm, bjp, want, tag, today, tweet, people, tamilnadu, know, pomonemodi, state, happen, country, tamil_nadu, let, twitter, support, tamil, visit, govt, stand, campaign, welcome, vote, tn.*

Using Intertopic Distance Map, we found that the adequate number of topics for the hashtag to be 9. After determining the optimal number of topics, we plotted topic-specific hate tweets over time to identify which topics generate more hate tweets and to understand their temporal distribution in Fig 4.

From Fig 4 we can see **Topic 8**, labeled Anti-Modi Sentiment, has the highest frequency of hate tweets, indicating many negative tweets target Prime Minister Modi. **Topic 7,** is the next most frequent source of hate tweets. This suggests that there is substantial discontent and anger directed towards the government's handling of farmer-related issues. Criticism is likely focused on policies perceived as detrimental to farmers. **Topic 5**, is the third most frequent source of hate tweets. This indicates that there is also a notable amount of negative sentiment directed towards the broader BJP leadership and their policies, not just Modi specifically. So we can say that these sentiments are notably present when PM Modi visits Tamil Nadu.

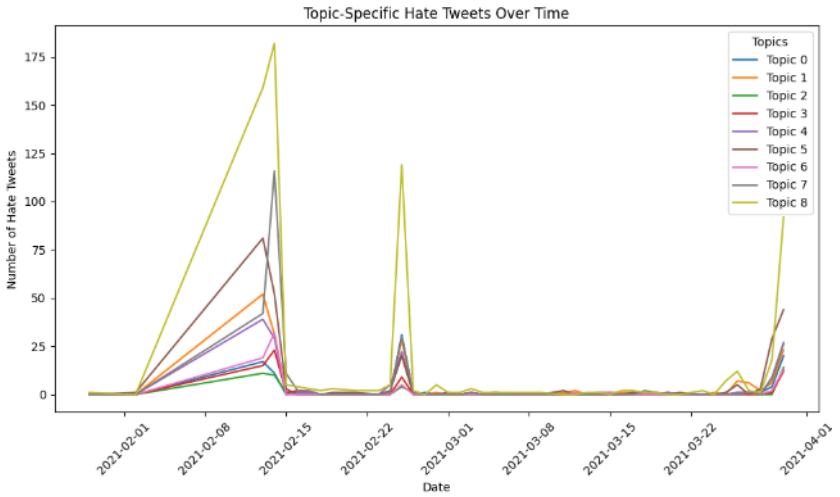

Fig. 4. Topic specific hate tweets over time - #GoBackModi

Table 7. Generated topics and topic labels - #GoBackModi

Topic No.	Keywords of generated topic	Topic label
0	happen, leader, party, bjp, voice, nation, hear, heart, travel, begin	Political Developments
1	stand, land, farmer, gobackmodimodi, forget, way, periyar, tamilnadu, hate, gobackmodi	Farmer Protests and Regional Sentiments
2	expect, pmofindia, roll, floor, supportfarmer, antifarmerbjp, tamilnadu, pm, remain, wish	Government Actions and Expectations
3	state, pttvonlinenew, sell, petrol, budget, miracle, talk, price, road, vote	State Politics and Economic Issues
4	trend, tweet, tag, gobackmodi, visit, today, twitter, let, want, support	Trending on social media
5	modi, people, pm, go_back, bjp, govt, welcome, gobackmodi, try, like	Modi and BJP Reception
6	bjp, pomonemodi, campaign, tamil, farmersdont, agriculture, set, gobackmoditalk, vote, cut	BJP Campaign and Agriculture Policies
7	government, farmer, change, year, medium, shame, gobackmodi, rahulgandhi, farmerprotest, issue	Government and Fanner Relations
8	gobackmodi, people, tamil_nadu, love, modijobdo, narendramodi, wait, destroy, miss, country	Anti-Modi Sentiment

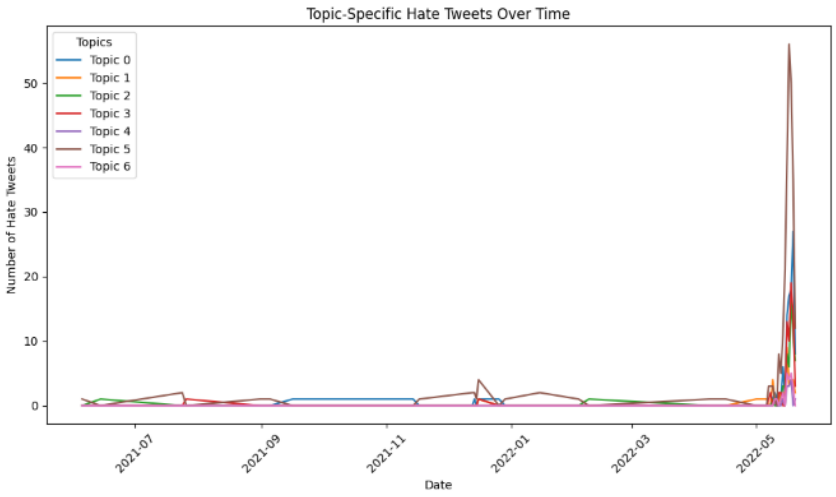

Fig. 5. Topic specific hate tweets over time - #Gyanvapi

#Gyanvapi: We obtained the most common words, according to LDA, used in comments for #Gyanvapi. The top-30 words for this topic were found out to

be: *survey, court, mosque, temple, order, shivle, build, gyanvapi, claim, case, masjid, shivling, videography, hindu, asi, demolish, truth, today, report, destroy, mybtstrack, kgfchapter, dharmaveer, thearchie, tripura, vishwanath, hear, aurangzeb, know, complex.*

Table 8. Generated topics and topic labels - #Gyanvapi

Topic No.	Keywords of generated topic	Topic label
0	face, amp, wait, nandi, gyanvapi, law, basement, change, mandir, question	Legal and Cultural Issues
1	mosque, survey, court, order, case, masjid, videography, today, report, dispute	Court Orders and Surveys
2	hindu, vishwanath, babri, worship, gyanvapi, case, masjid, justice, foot, right	Religious Rights and Legal Battles
3	shivle, masjid, claim, shivling, gyanvapimasjid, fountain, complete, hindu, babamilgaye, muslim	Claims and Religious Artifacts
4	asi, mybtstrack, kgfchapter, dharmaveer, thearchie, tripura, gyanvapi, channel, biplabkumardeb, conclude	Archaeological Findings and Protests
5	gyanvapi, temple, mosque, build, hindu, demolish, destroy, truth, know, Aurangzeb	Historical and Religious Conflicts
6	set, safety, spread, share, breakingnew, election, dismiss, yogiadityanath, save, comment	Political and Social Movements

Using Intertopic Distance Map, we found that the adequate number of topics for the hashtag to be 7. After determining the optimal number of topics, we plotted topic-specific hate tweets over time to identify which topics generate more hate tweets and to understand their temporal distribution in Fig 5.

From Fig 5 we can see **Topic 5**, has the highest frequency of hate tweets. There's a debate about whether the Gyanvapi mosque and a site in Mathura were originally Hindu temples destroyed by Muslim rulers, potentially Aurangzeb during the Mughal era. Hindus might be seeking to reclaim these sites based on their history. **Topic 0**, is the next most frequent source of hate tweets which dispute over Gyanvapi site in India. Hindus seek to reclaim land or change law to allow worship at mosque. Overall we can say that This hate posts suggests a highly charged online environment surrounding the Gyanvapi mosque dispute. There's a mix of calls for discussion, religious references, and potentially violent threats.

#IndianMuslims: We obtained the most common words, according to LDA, used in comments for #IndianMuslims. The top-30 words for this topic were found out to be: *Muslim, woman, activist, life, stop, hindu, state, police, kill, leader, people, support, country, blm, violence, assamhorror, deal, lead, hate, attack, crime, matter, bjp, justice, religion, happen, government, family, rss, arrest.*

Using Intertopic Distance Map, we found that the adequate number of topics for the hashtag to be 15. After determining the optimal number of topics, we plotted topic-specific hate tweets over time to identify which topics generate more hate tweets and to understand their temporal distribution in Fig 6.

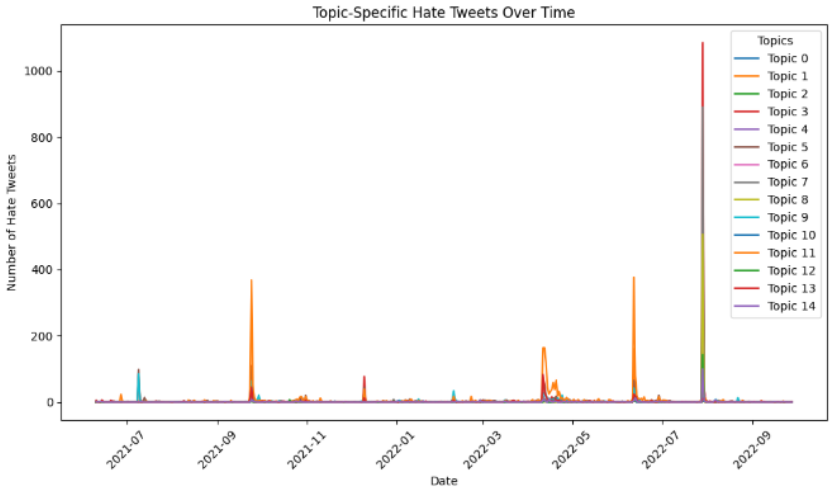

Fig. 6. Topic specific hate tweets over time - #IndianMuslims

From Fig 6 we can see **Topic 3, 7 and 8**, has the highest frequency of hate tweets on 2022-07-29 because Mohammed Fazil, a 23-year-old man, was brutally murdered on July 28, 2022. This seems to be deeply concerned with the impact of hate crimes (Fazil murder case) and discrimination on the Muslim community, the role of political and social actors, and the community's response to these challenges. **Topic 7**, is the next most frequent source of hate tweets on 2022-07-29 which indicates a notable concern with government actions and state violence, reflecting public outrage and the contentious nature of political activism. **Topic 8**, is the next most frequent source of hate tweets on 2022-07-29 which indicates Muslim women and girls facing racism, violence, and discrimination in India, highlighting issues of democracy, secularism, and human rights, with references to historical injustices, terrorist labels, and advocacy for change by figures like political leaders. **Topic 11**, is the next most frequent source of hate tweets which suggests a focus on incidents in Assam involving attacks (like acid attack, alleged violence by police forces etc.) , and also with arrests and condemnation of human rights abuses. The overall analysis shows that a large proportion of hate tweets focus on the discrimination and violence faced by the citizens, especially the Muslim community in India.

The topic modeling of hateful tweets across these hashtags reveals a significant level of division and anger, particularly in political and religious contexts. Political polarization is prominent, with high levels of hate speech directed to-

Table 9. Generated topics and topic labels - #IndianMuslims

Topic No.	Keywords of generated topic	Topic label
0	Stop, crime, justice, start, pray, politic, kill, member, pay, situation	Crime and Justice
1	religion, protest, watch, people, truth, group, incident, fail, oppression, discrimination	Religious Protests and Discrimination
2	life, support, government, bjp, protect, victim, care, society, play, state	Government and Society
3	muslim, kill, hate, matter, live, target, fascism, murder, faith, innocent	Fazil murder case
4	lead, medium, law, voice, raise, threaten, hand, order, state, run	Legal and Political Advocacy
5	country, deal, let, stand, hindutvaterror, trend, right, humanity, people, islamophobia	National Issues and Rights
6	hindu, rss, save, death, bjp, youth, power, goon, kill, election	Hindu Nationalism and Violence
7	activist, state, leader, violence, happen, student, murder, bjp, government, house	Violence and Activism
8	fight, terror, girl, spread, love, life, break, lose, week, notice	Terror and Social Issues
9	woman, muslimgirl, face, man, indianmuslim, action, destroy, democracy, word, brutally	Muslim women and girls facing racism
10	family, humanright, jail, try, child, rise, defend, imagine, narendramodi, change	Family and Human Rights
11	police, assamhorror, attack, arrest, assam, condemn, unhumanright, amp, shame, mosque	Police and Religious Conflict
12	blm,speak, tytlive, silence, atrocity, injustice, realface, join, ignore, ground	Black Lives Matter and Global Atrocities
13	genocide, today, know, want, human_right, terrorist, terrorism, way, hatred, community	Genocide and Terrorism
14	people, blood, home, rssterrorist, continue, die, praise, sullideal, article, eye	Continued Violence and Terrorism

wards political figures, especially Prime Minister Narendra Modi and the BJP. Hashtags like #Andhbhakt and #GoBackModi show intense political and ideological opposition. The #Andhbhakt hashtag with users frequently expressing strong anti-follower sentiments towards political figures and their supporters, while #GoBackModi Tweets express dissatisfaction and hostility toward Modi, reflecting regional discontent and opposition to his presence and policies. The sentiment is especially strong during his visits, indicating regional political dynamics. Religious tensions are also evident, particularly in historical and religious disputes related to the Gyanvapi mosque. The participation in discussions spikes around specific incidents, unlike other hashtags where there is more regular activity. The #IndianMuslims hashtag highlights large and continuous discussions citing discrimination and violence against minority communities in India. Many tweets focus on issues such as hate crimes, societal discrimination, and violence

against minority communities. The discussions often highlight incidents of violence and government actions, reflecting broader societal issues and human rights concerns. Tweets under this hashtag underscore the systemic challenges faced by minorities in India.

These findings reveal the contentious nature of online discourse in India, with significant implications for social and political cohesion. Some issues remain perpetual points of discussions (having some base intensities), and become really bursty when certain incidents spark further discussion on these themes (extraneous effects). On the other hand, there are certain topics that become active discussion point when some triggering event happens in the society. Having said that, there is a significant portion of social media content that is clean and not hateful, as indicated by the ratio of non-hate and hate posts in the dataset, which is a positive impact of social media, and should not be forgotten.

5 CONCLUSION

In this study, we examine the detection and dynamics of hate speech on Twitter, particularly focusing on political and religious discussions in the context of Indian hashtags. Social media platforms have undeniably transformed our modes of interaction and news consumption, providing a space for free expression but also creating avenues for the spread of hate speech. Our research highlights the critical importance of understanding these dynamics to mitigate the negative impacts on societal harmony. We employed various machine learning models to detect hate speech within the collected data and utilized advanced topic modeling techniques like Latent Dirichlet Allocation (LDA) to uncover underlying themes and user reactions associated with the selected hashtags. The results demonstrated the efficacy of these models in identifying and classifying hate speech accurately. Through topic modeling, we were able to extract significant insights from a large volume of unstructured data, providing a comprehensive understanding of the context in which hate speech occurs, the temporal patterns of discussions, and the comparative trends between different hashtags.

The findings from our study underscore the complex interplay between user behavior, content characteristics, and the broader network dynamics that influence the propagation of hate speech on social media platforms. Future work could extend this research to other social media platforms and explore additional machine learning techniques to further enhance hate speech detection and analysis. Ultimately, understanding and addressing the dynamics of hate speech is essential for maintaining societal harmony and protecting the diverse communities that engage in online discourse.

References

1. Zeerak Waseem and Dirk Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In Proceedings of the NAACL student research workshop, pages 88–93, 2016.

2. Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In Proceedings of the international AAAI conference on web and social media, volume 11, pages 512–515, 2017.
3. Jennifer Golbeck, Zahra Ashktorab, Rashad O Banjo, Alexandra Berlinger, Siddharth Bhagwan, Cody Buntain, Paul Cheakalos, Alicia A Geller, Quint Gergory, Rajesh Kumar Gnanasekaran, et al. A large labeled corpus for online harassment research. In Proceedings of the 2017 ACM on web science conference, pages 229–233, 2017.
4. Goel, V., Sahnan, D., Dutta, S., Bandhakavi, A., and Chakraborty, T. (2023). Hatemongers ride on echo chambers to escalate hate speech diffusion. PNAS nexus, 2(3)
5. Beatty, Matthew. "Graph-based methods to detect hate speech diffusion on Twitter." 2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). IEEE, 2020.
6. Mathew, B., Dutt, R., Goyal, P., and Mukherjee, A. (2019, June). Spread of hate speech in online social media. In Proceedings of the 10th ACM conference on web science (pp. 173-182).
7. Jae Yeon Kim, Carlos Ortiz, Sarah Nam, Sarah Santiago, and Vivek Datta. Intersectional bias in hate speech and abusive language datasets. arXiv preprint arXiv:2005.05921, 2020.
8. Sear, R., Restrepo, N. J., Lupu, Y., and Johnson, N. F. (2022). Dynamic Topic Modeling Reveals Variations in Online Hate Narratives. In Lecture notes in networks and systems (pp. 564–578). https://doi.org/10.1007/978-3-031-10464-0_38
9. Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. Hatexplain: A benchmark dataset for explainable hate speech detection. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, pages 14867– 14875, 2021.
10. V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. R. Pardo, P. Rosso, and M. Sanguinetti, "SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter," in *Proceedings of the 13th International Workshop on Semantic Evaluation*, 2019.
11. S. Kumar, R. West, and J. Leskovec, "Disinformation on the Web: Impact, Characteristics, and Detection of Wikipedia Hoaxes," in *Proceedings of the 25th International Conference on World Wide Web*, 2020.