Election-Driven Trends in Reddit Hate Speech: A Clustering Approach to Target Analysis Before and During Trump Campaign Periods in USA

REBECCA ANDREI, University of Twente, The Netherlands

The intersection of political campaigns and online discourse is critical for understanding the emergence of hate speech. This paper examines how Donald Trump's 2024–2025 presidential campaign influenced hate speech patterns on Reddit, a major platform for political discussion. Using 55M Reddit posts collected before and during Trump's campaign, we apply Natural Language Processing (NLP) techniques to analyze the dynamics of hate speech over time.

Hate speech classification is performed with a fine-tuned Bidirectional Encoder Representations from Transformers (BERT) model, which captures deep contextual meaning and detects subtle, implicit forms of hate speech common in political conversations. Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) is then employed to identify latent groupings of hate speech targets. HDBSCAN is chosen over traditional DBSCAN for its ability to handle varying cluster densities, noisy data, and high-dimensional embeddings produced by BERT. Cluster boundaries are further refined using an algorithm that separates overlapping clusters to delineate target categories clearly.

This approach offers empirical insights into how political campaigns may shape toxic online behavior and broader trends in social polarization, informing academic research and policy discussions on digital discourse, political communication, and platform moderation.

Additional Key Words and Phrases: hate speech, Reddit, clustering, BERT, HDBSCAN, U.S. election campaign, comparative study

1 INTRODUCTION

The relationship between political campaigns and online dialogue has been a subject of intense academic interest [12]. Political campaigns actively shape public opinion and discourse around key issues. Online platforms like Reddit amplify this influence by spreading campaign messages rapidly and interactively [4]. Furthermore, it enables online users to spread their stance on political implications, such as concerns related to rights and freedoms, misinformation, and even expressions of hate and prejudice. Potential consequences of spreading hate speech online during political campaigns include escalating social division [10], increased harassment or discrimination [13], and even violent incidents [1]. Studying this relationship helps researchers understand how political rhetoric translates into public discourse.

In particular, the influence of presidential campaigns on the nature of hate speech has brought attention to social polarization [7]. In this context, social polarization translates into the adoption of an us-versus-them mentality, where individuals strongly align with one political side and view others as threats or enemies, allowing hate speech to often occur as a means to segregate society into opposing identity-based groups.

This study examines trends in hate speech on Reddit during two key periods: before and during Donald Trump's 2024-2025 election campaign. Specifically, we will look at two key equally long periods surrounding the political campaign:

- Before campaign: 1st of January 2024
- During campaign: 1st of November 2024

This study examines the shifts in hate speech targets over time using Natural Language Processing (NLP) techniques, specifically Bidirectional Encoder Representations from Transformers (BERT)[1] for classification. Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN)[2] is used for clustering the targets of hate speech.

2 MOTIVATION

Given the societal segregation caused by political debates [8], understanding how political campaigns influence hate speech online is crucial for highlighting harmful trends in online dialogue. Prior studies [3, 7, 12] have largely focused on general trends of hate speech, or different countries, without analyzing the specific targets of such speech during the recent USA campaign. This research seeks to fill that gap by identifying and comparing the main targets of hate speech before and during Trump's 2024-2025 campaign, offering insights into how political climates can shift public hostility toward different groups.

3 PROBLEM STATEMENT

3.1 Problem Description

Despite a growing body of research on hate speech and political events, there is still a limited understanding of how hate speech targets evolve throughout political cycles. It is consequential to find insights into the shifting of the volume of each target cluster, pre- and during the political campaign, and, specifically, how such a salient political event influences the spread of hate speech.

3.2 Research Questions

This problem raises the central inquiry guiding this study: *How does a major political event, such as Donald Trump's 2024-2025 campaign, influence the spread and targeting of hate speech on Reddit?*

To address this overarching question, we will explore the following sub-questions:

(1) How does the volume of hate speech change before and during the Trump 2024 campaign?
(2) How does the proportion of each target cluster of hate speech shift before and during the Trump campaign, based on clustering analysis?

4 OBJECTIVES

The objective of this paper is to classify Reddit posts and comments as hate speech (label 0), offensive language (label 1), and neutral

Author's address: Rebecca Andrei, r.andrei-1@student.utwente.nl, University of Twente, Enschede, The Netherlands, 7500AE.

[1] https://huggingface.co/docs/transformers/en/model_doc/bert
[2] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.HDBSCAN.html

(label 2) using the BERT model. After filtering and segregating the hate speech content, the next step is to cluster the instances to identify primary targets in two periods (pre-campaign and during-campaign). The clustering approach is finalized using HDBSCAN. The results of both clustering processes are observed and compared, as we will focus on the volume of the hate speech, proportions of each cluster pre- and during-campaign, shifting in targets, and plot any observed trends.

5 TECHNICAL BACKGROUND

Reddit is an online discourse platform known for relatively unfiltered dialogue and provides a rich dataset for studying online behavior. Recent advancements in NLP [2], especially Transformer-based models like BERT, have significantly boosted the performance of text classification tasks such as hate speech detection. BERT is a linguistic model developed by a group of Google scientists. Unlike traditional models (e.g., Bag-of-Words, TF-IDF, or even Word2Vec), BERT reads text **bidirectionally**, meaning it understands each word in the context of the full sentence rather than just its neighbors. In 2024, Shasank et. al [9] have proved that using BERT combined with another classifier improves the model's performance, and, to find clusters of targets, they have searched manually, using a Regular Expression (RegEx), for keywords that can be categorized as offensive terms.

For the clustering component, this study uses HDBSCAN instead of traditional DBSCAN [3] or RegEx. HDBSCAN extends DBSCAN by constructing a hierarchy of clusters that can adapt to varying densities, an essential feature for high-dimensional BERT embeddings. It also reduces sensitivity to hyperparameter tuning by eliminating the need for a fixed distance threshold (epsilon), relying instead on a more interpretable min_cluster_size. This flexibility enables more nuanced discovery of shifting hate speech targets—ranging from concentrated slurs to diffuse ideological language—making HDBSCAN a more robust tool for uncovering latent structure in Reddit discourse.

6 METHODOLOGY

The process followed in this study for the analysis of the hate speech targets is depicted in Fig.1. Each process box is explained in the following subsections.

6.1 Datasets

The raw datasets are sourced from academic torrents[4] and collectively amount to approximately 500GB of Reddit data. The data is divided temporally into two periods: before the election campaign (January, February, and March) and during the campaign (November, December, and January). Each period comprises roughly 240GB of data. Upon downloading, the datasets followed a standardized folder structure (e.g., November > Submissions/Comments > .zip files). We do not differentiate between comments and submissions.

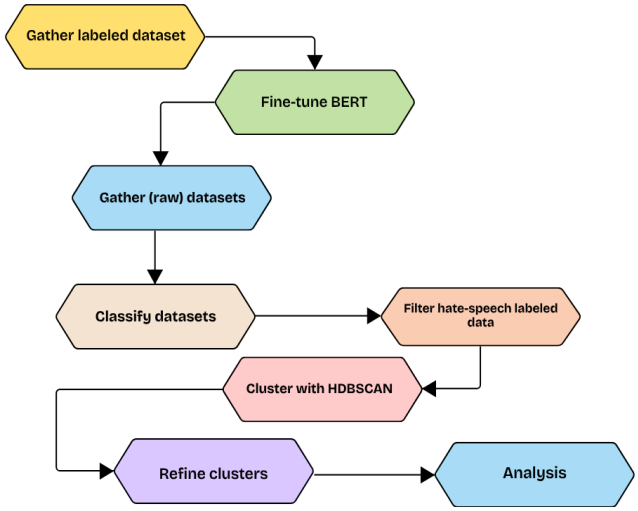

Fig. 1. Methodology Flowchart

6.2 Preprocessing

Each file is parsed and converted into .jsonl format, following a cleanup of unnecessary fields (e.g., "*_meta*", *"all_awardings*", "*approved_at_utc*", "*archived*", "*author*", metadata about author, timestamps data, "*subreddit_name_prefixed*", etc). Table 9 demonstrates the preprocessing step of one JSON-line object, before and after the cleanup. Subsequently, the data is split into 1.5GB chunks for compatibility with GitLab upload restrictions. Splitting the files into 1.5GB chunks did not affect either the data continuity or the analysis, since the remainder is also retained and processed as a final chunk. This preprocessing step results in around 160 processed files (chunks) for each time period. Due to time constraints (*see Limitations section*), only 6 files for each period are sequentially selected (i.e., not randomly) from the start of the respective time. Hence, the chosen periods for this study are January (pre-campaign) and November (during-campaign), which result in **55M** Reddit posts and comments altogether.

6.3 Fine-tuning

A pretrained BERT model (bert-base-uncased) is fine-tuned on a labeled dataset obtained from HuggingFace[5]. This labeled dataset originally indicated class imbalance, with significantly fewer instances labeled as hate speech and offensive (Class 0 and 1, respectively). To mitigate the imbalance, upsampling is performed on both Class 0 (Hate Speech) and Class 1 (Offensive) to equalize class distributions. Upsampling (also known as oversampling) is a data processing and optimization technique that addresses class imbalance in a dataset by adding data. Upsampling adds data by using original samples

[3] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html
[4] https://academictorrents.com/browse.php?search=reddit&c6=1
[5] https://huggingface.co/datasets/tdavidson/hate_speech_offensive/blob/main/data/train-00000-of-00001.parquet

from minority classes until all classes are equal in size [6]. The balanced data set is divided into 90% training and 10% testing sets. To fine-tune the model (and to even deploy the model later), we need to load the Tokenizer[6].

6.3.1 Tokenizer. Because the model can not work with actual words, the Tokenizer transforms the input text from the dataset into a sequence of numbers. For example, the input text *"a b c d"* will be transformed into ["a", "b", "c", "d"]. Each word is then mapped to an Integer from the model's vocabulary, say ["a", "b", "c", "d"] →[0, 1, 2, 3]. The model also needs to understand the beginning and the end (or separation), so it creates special tokens ("[CLS]" and "[SEP]", respectively). Therefore, the sequence becomes: ["[CLS]", 0, 1, 2, 3, "[SEP]"] →[101, 0, 1, 2, 3, 102]. The Tokenizer follows the above-mentioned steps for the entire dataset and feed the result to the model.

6.3.2 Classification Report. To assess the model's performance, we use the Sklearn metrics library[7], which allows us to evaluate the quality of the predictions. This library is part of the Scikit-learn library[8]. The metrics needed for asserting the classification report are *Precision*, *Recall*, and *F1-score*. To determine these metrics, we will first look at the skeleton of a confusion matrix in Table 1, where we find the True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN).

Actual \Predicted	**0**	**1**	**2**
0	TP_{00}	FN_{01}	FN_{02}
1	FP_{10}	TP_{11}	FN_{12}
2	FP_{20}	FP_{21}	TP_{22}

Table 1. Confusion Matrix Skeleton for 3-Class Classification

The corresponding values for TP, TN, FP, FN are depicted in Table 2.

Actual \Predicted	**0**	**1**	**2**
0	1899	0	0
1	105	1771	45
2	3	5	1929

Table 2. Confusion Matrix for 3-Class Classification with corresponding values

The **Precision**[5] tells us of all Reddit content the model predicted as X (label 0, 1, or 2), how many are indeed X. Precision is computed as:

$$Precision = \frac{TP}{TP + FP}$$

Recall[5] computes of all actual X-labeled content how many the model correctly identifies, with the following formula:

$$Recall = \frac{TP}{TP + FN}$$

[6]https://github.com/huggingface/transformers/blob/v4.52.3/src/transformers/models/bert/tokenization_bert.py#L51
[7]https://scikit-learn.org/stable/api/sklearn.metrics.html
[8]https://scikit-learn.org/stable/

Lastly, we need to assert the **harmonic mean**[5] (**F1-score**), meaning a balance between Precision and Recall. The following formula is used:

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Table 3 presents the performance metrics of the classifier with their corresponding values.

Label	**Precision**	**Recall**	**F1-Score**
0 (Hate Speech)	0.95	1.00	0.97
1 (Offensive)	1.00	0.92	0.96
2 (Neutral)	0.98	1.00	0.99

Table 3. Classification report for the model evaluation

6.4 Classification

The fine-tuned classifier is deployed on the 12 preprocessed data chunks (6 from each campaign period). Due to computational and time constraints, the classification process has been executed in parallel: one terminal processing the pre-campaign data and another for the during-campaign data. Each chunk required approximately 24 hours for classification.

6.5 Filter hate speech labeled data

A Python script is used to filter out only hate speech-labeled data, necessary in the next phase. As provided in Table 5, we recognize 86763 hate-speech labeled Reddit posts in the "Before Campaign" period, and 93541, respectively, in the "During Campaign" period.

6.6 Clustering

To analyze thematic structures within the detected hate speech, the HDBSCAN algorithm is deployed on the filtered Reddit data. HDBSCAN automatically groups similar targets of hate speech. It finds patterns by grouping closely related items that are packed closely (dense regions) and ignores outliers, helping us discover natural clusters without needing to choose the number of clusters in advance. The outliers are treated as noise and left out. In our case, we use HDBSCAN to group together similar targets of hate speech. More specifically, if certain groups are mentioned in similar ways, the algorithm will place them in the same cluster. This helps us better understand which types of targets are being talked about in similar hateful ways, without us having to guess how many such groups exist. We deploy HDBSCAN to group similar entities based on their embeddings, using **min_cluster_size = 10**. This means any group smaller than 10 entities is not considered a real cluster but treated as noise. This helps to avoid creating clusters that are too small to be meaningful, making sure the groups we find are reliable and easier to understand. We use the default value for **min_samples**, which is set equal to **min_cluster_size**, helping to balance cluster stability and noise sensitivity. This means that a point must have at least 10 nearby neighbors to be considered a core point, further reducing the chance of including noise in clusters. Additionally, the algorithm uses the **Euclidean distance metric**. This measures how similar (or different) the embeddings (word vectors) are from each other, which is well-suited for the dense, continuous vector

Table 4. Most used words in each cluster, with word counts

Ideology	#	**Community Group**	#	**Gender (Men)**	#	**Sexual Orientation**	#
racist	98	group	337	men	854	gay	315
white	53	gay	220	white	355	queer	158
maga	66	groups	122	guys	322	group	111
women	44	community	253	black	149	women	104
feminist	23	movement	55	trans	42		
homophobic	10						
Nationality	#	**Race/Ethnicity**	#	**Gender (Women)**	#	**Religion**	#
asian	115	white	635	women	714	islam	170
white	30	black	323	milf	214	judaism	89
american	77	asian	141	trans	132		
arab	64	latina	117	white	150		
indian	75			black	82		
muslim	80			girls	490		
				chubby	25		

representations produced by the embedding model. Together, these settings ensure that the clusters reflect meaningful groupings in the data while minimizing the influence of outliers. The clustering algorithm identifies 6 clusters from the pre-campaign data and 3 clusters from the during-campaign data.

However, this method indicates limitations in handling overlapping hate speech categories. For example, a phrase such as "cis white straight men" targets multiple identity axes (gender, sexual orientation, and race). To address this, a post-processing algorithm is developed to decompose such phrases into their essential target categories and assign hate speech labels to each relevant identity axis. These new categories are shown in Table 6. As represented in the category mapping, some of the groups contain slurs. The ethical justification for displaying and filtering upon these terms is that the 0-labeled (hate speech) filtered content includes as targets these specific words.

	Before Campaign	**During Campaign**
Hate Speech count	86763	93541

Table 5. Number of hate speech instances before & during Trump's campaign

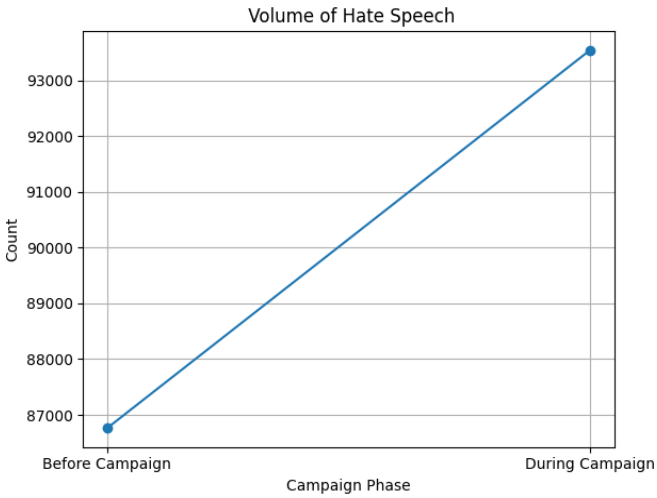

Fig. 2. Increase in hate speech volume between campaign phases

7 RESULTS

RQ1: Volume Change in Hate Speech

To investigate how the volume of hate speech changes during these two key periods (before and during the campaign), we analyze the hate speech instances in Table 5. By computing the difference across the two time periods, the data shows a clear increase: from 86,763 instances before the campaign to 93,541 during it.

Fig. 2 shows the linear increase between these two points. This represents a 7.81% increase, amounting to 6,778 additional instances during the campaign period. This increase suggests a measurable amplification in hate speech volume in line with the early stages of the campaign cycle. To evaluate whether the observed increase in hate speech from the **pre-campaign** to the **during-campaign** period is statistically significant, we apply a **Chi-Square Test of Independence**. The following data are analyzed, using formulas from [11]:

Hypotheses

- H_0: There is no association between time period and hate speech frequency.

Semantic Category	Keywords
Gender woman	women, girls, milf, feminist, cis woman, trans woman
Gender men	men, guys, cis men, trans men
Sexual orientation	gay, queer, lgbt, bisexual, lesbian, trans, fag
Race ethnicity	black, white, asian, latino, brown, african, european, mizrahi, palestinian, jew, hispanic, native, indian
Religion	muslim, jew, jewish, christian, islamic, zionist
Ideology	maga, racist, liberal, conservatism, feminist, misogynist, homophobic
Community group	community, group, movement, chat, groupchat
other	government, support group, organization

Table 6. Refined categories used to split the conglomerated clusters

Period	Hate Speech	Non-Hate Speech	Total Posts
Before	86,763	27,581,505	27,668,268
During	93,541	27,270,890	27,364,431
Total	180,304	54,852,395	55,032,699

Table 7. Contingency table of Reddit posts before and during the campaign

- H_1: There is an association (i.e., a real increase in hate speech during the campaign).

Expected Values

In Table 7, we find the necessary values for computing E_{ij} (the expected frequency for a specific cell in a contingency table), O_{ij} (the observed frequency, i.e., the actual count of occurrences from the table), and ultimately X^2 (Chi-Square test statistic). The expected frequency for each cell is calculated using the following: (See Appendix Detailed Chi-Square Computation for step-by-step computations)

$$E_{ij} = \frac{(\text{Row Total}) \times (\text{Column Total})}{\text{Grand Total}}$$

Chi-Square Test Statistic

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 336.19$$

Degrees of Freedom

$$df = 1$$

p-Value

$$p < 0.0001$$

Since the test yields $\chi^2 = 336.19$ with $df = 1$ and $p - value < .0001$ at $p < .05$, we reject the null hypothesis. There is a statistically significant increase in hate speech during the campaign period, indicating that the 7.8% rise is unlikely to be due to random variation.

RQ2: Shifts in Clusters of Targets of Hate Speech

We perform a cluster analysis to categorize hate speech instances according to the targeted group. The category frequencies from periods before-campaign and during-campaign are shown in Fig. 3, and Fig. 4, respectively. We notice that the frequencies of hate speech targets not only change, but increase in **during-campaign** period.

Furthermore, the analysis yields ten clusters, of which eight are relevant. Table 4 provides a clear definition of the composition of each target. The most frequent hate-speech targets from **before campaign** and **during campaign** periods are shown in Table 4, where each column represents the cluster, and columns named **'#'** represent the sum of word counts in both periods.

In Table 8, we see the numerical differences between the target categories before and during campaign phases, where columns marked by '#' show the number of entities contained in each cluster. In particular, an interesting finding is that **Ideology** target category (described by: racism, MAGA, feminism, etc., see Table 4) experienced an increase of **134.21%**. This suggests that more hate speech has been pointed towards different races, in particular European ancestry, MAGA supporters, and feminism and women. The categories **community_group** (depicted as general

Category	# - Before	# - During	Difference	% Change
ideology	76	178	+102	**134.21%**
community_group	306	487	+181	59.15%
gender_men	1050	1641	+591	56.29%
sexual_orientation	419	617	+198	47.26%
nationality	520	747	+227	43.65%
race_ethnicity	753	1079	+326	43.29%
gender_women	633	879	+246	38.86%
religion	155	191	+36	23.23%
uncategorized	573	1152	+579	**101.05%**
other	48	104	+56	**116.67%**

Table 8. Change in frequency and percentage of hate speech target clusters before and during the Trump 2024 campaign.

group(s), lgbtq+ community, and social movements) and **gender_men** (category described as biological and transgender men, European and African-American ancestry men, and "guys", *See Table 4*) experienced a dramatic proportional increase as well, over 50%.

8 STATE OF THE ART

The detection and analysis of hate speech have seen significant advancements with the rise of deep learning and NLP technologies. Early approaches relied on Bag-of-Words (BoW), n-gram models, and handcrafted lexicons, which struggled with generalization and nuance. More recent works have adopted Transformer-based architectures such as BERT and RoBERTa, which excel in capturing contextual semantics and have demonstrated superior performance in hate speech detection across multiple languages and platforms [2, 9]. While many studies have focused on classifying content as hateful or non-hateful, fewer have investigated the specific targets of hate speech. ElSherief et al. (2018) and Shasank et al. (2024) took a step in this direction by using keyword-based categorization to investigate identity-based targets. However, such rule-based methods lack scalability and adaptability to evolving language patterns.

8.1 Contributions

This paper builds upon this foundation by employing **HDBSCAN**, a clustering algorithm better suited for non-uniform densities, and integrating it with **BERT**-based embeddings. Moreover, this study reveals a **temporal analysis** of hate speech targets, before and during a major political event, showing a noticeable shift in both volume and clustering behavior. In addition, the **labeled dataset** provides reliability, given the model's performance metrics shown in Table 3. This high performance has been achieved by fine-tuning the uncased BERT model on a **rich and balanced** hate speech dataset. The accuracy of the results is reinforced by the application of **HDBSCAN** and a novel post-processing algorithm, avoiding manual feature engineering. Lastly, the results themselves act as empirical evidence of political polarization's impact on online behavior, offering implications for moderation strategies during campaign cycles.

9 LIMITATIONS

This study is subject to a limited window of time, ten weeks specifically. Furthermore, the available hardware does not meet the computational requirements for the planned datasets. For these reasons, the approximately 160 chunks from each dataset are reduced to only 6 chunks per period (pre- and during-campaign). For the same reason, the BERT model is fine-tuned with a dataset of approximately 60.000 labeled Reddit posts.

A larger dataset would have provided better training diversity, which would have led to more meaningful segmentation in clustering analysis, subsequently leading to more reliable results.

10 FUTURE WORK

Future research could address the limitations of this study (*See Limitations Section*) by expanding the dataset, integrating multi-modal data, enhancing computational resources (hardware), and applying more complex clustering. Expanding the dataset by including more months before and during the campaign would enhance the results of this study, enabling long-term trend analysis. Multi-modal data, such as images and memes, which are prevalent in Reddit posts, may carry implicit hate speech that is not readily apparent through text. Furthermore, higher hardware standards are necessary to handle the larger and improved datasets mentioned above. For example, the time calculated necessary to complete the initially planned $\tilde{3}20$ file chunks is 160 days, given that the servers maintain the same availability. Lastly, future improvements can be achieved by deploying a more complex clustering analysis, given that a majority of the clusters exhibit limitations when analyzing overlapping groups.

11 CONCLUSIONS

This study explores how hate speech on Reddit evolves in response to significant political events, focusing on the 2024–2025 U.S. presidential campaign of Donald Trump. Using a fine-tuned BERT model for hate speech detection on 55M Reddit posts and comments and HDBSCAN for clustering, we identify both quantitative and qualitative changes in toxic discourse across two key time-frames: before and during the campaign. The results indicate a 7.81% increase in hate speech volume during the campaign period, with specific target groups, such as those related to ideology, community affiliation, and gender, experiencing increased aggression. By decomposing overlapping clusters, we are able to uncover nuanced trends that would have otherwise been obscured. This decomposition reveals that hate speech not only increases in volume but also evolves in structure, often intensifying toward certain identity-based categories. These insights contribute to a broader understanding of how political climates can influence toxic digital behavior and provide a foundation for future work in computational social science and platform moderation.

REFERENCES

[1] Carlos Arcila Calderón, Patricia Sánchez Holgado, Jesús Gómez, Marcos Barbosa, Haodong Qi, Alberto Matilla, Pilar Amado, Alejandro Guzmán, Daniel López-Matías, and Tomás Fernández-Villazala. 2024. From online hate speech to offline hate crime: The role of inflammatory language in forecasting violence against migrant and LGBT communities. *Humanities and Social Sciences Communications* 11, 1 (2024), 1369. https://doi.org/10.1057/s41599-024-03899-1

[2] K. R. Chowdhary. 2020. *Natural Language Processing.* Springer India, New Delhi, 603–649. https://doi.org/10.1007/978-81-322-3972-7_19

[3] Mai ElSherief, Vivek Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth Belding. 2018. Hate Lingo: A Target-Based Linguistic Analysis of Hate Speech in Social Media. *Proceedings of the International AAAI Conference on Web and Social Media* 12, 1 (2018). https://doi.org/10.1609/icwsm.v12i1.15041

[4] J. D. Gallacher. 2021. *Online intergroup conflict: How the dynamics of online communication drive extremism and violence between groups.*

[5] Google Developers. 2025. Classification: Accuracy, recall, precision, and related metrics. https://developers.google.com/machine-learning/crash-course/classification/accuracy-precision-recall. Accessed: 2025-06-16.

[6] IBM. 2025. What is upsampling? https://www.ibm.com/think/topics/upsampling. Accessed: 2025-06-03.

[7] Farhan Ahmad Jafri, Kritesh Rauniyar, Surendrabikram Thapa, Mohammad Aman Siddiqui, Matloob Khushi, and Usman Naseem. 2024. CHUNAV: Analyzing Hindi Hate Speech and Targeted Groups in Indian Election Discourse. *ACM Transactions on Asian and Low-Resource Language Information Processing* (2024). https://doi.org/10.1145/3665245

[8] W. Ben McCartney, John Orellana-Li, and Calvin Zhang. 2024. Political Polarization Affects Households' Financial Decisions: Evidence from Home Sales. *The Journal of Finance* 79, 2 (2024), 795–841. https://doi.org/10.1111/jofi.13315 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/jofi.13315

[9] Shasank Sekhar Pandey, Alberto Garcia-Robledo, and Mahboobeh Zangiabady. 2024. Decoding Online Hate in the United States: A BERT-CNN Analysis of

36 Million Tweets from 2020 to 2022. In *18th IEEE International Conference on Semantic Computing (ICSC2024)*. IEEE, United States, 329–334. https://doi.org/10.1109/ICSC59802.2024.00059

[10] Alexandra A. Siegel. 2020. Online hate speech. *Social Media and Democracy: The State of the Field, Prospects for Reform* (2020), 56–88.

[11] Ronald J. Tallarida and Rodney B. Murray. 1987. *Chi-Square Test.* Springer New York, New York, NY, 140–142. https://doi.org/10.1007/978-1-4612-4974-0_43

[12] Zorica Trajkova and Silvana Neshkovska. 2018. Online hate propaganda during election period: The case of Macedonia. *Lodz Papers in Pragmatics* 14, 2 (2018), 309–334. https://doi.org/10.1515/lpp-2018-0015

[13] Oana Ștefăniță and Diana-Maria Buf. 2021. Hate Speech in Social Media and Its Effects on the LGBT Community: A Review of the Current Research. *Romanian Journal of Communication and Public Relations* 23, 1 (2021), 47–55. https://doi.org/10.21018/rjcpr.2021.1.322

Appendix

Detailed Chi-Square Computation

$$E_{ij} = \frac{(\text{Row Total}) \times (\text{Column Total})}{\text{Grand Total}}$$

$$E_{\text{Before, Hate}} = \frac{27{,}668{,}268 \times 180{,}304}{55{,}032{,}699} \approx 90{,}649.7$$

$$E_{\text{During, Hate}} = \frac{27{,}364{,}431 \times 180{,}304}{55{,}032{,}699} \approx 89{,}654.3$$

$$E_{\text{Before, Non-hate}} = \frac{27{,}668{,}268 \times 54{,}852{,}395}{55{,}032{,}699} \approx 27{,}577{,}618.3$$

$$E_{\text{During, Non-hate}} = \frac{27{,}364{,}431 \times 54{,}852{,}395}{55{,}032{,}699} \approx 27{,}274{,}776.7$$

Degrees of Freedom

$$df = (2-1)(2-1) = 1$$

Chi-Square Statistic

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

$$\begin{aligned}\chi^2 &= \frac{(86{,}763 - 90{,}649.7)^2}{90{,}649.7} + \frac{(93{,}541 - 89{,}654.3)^2}{89{,}628.8} + \frac{(27{,}581{,}505 - 27{,}577{,}618.3)^2}{27{,}577{,}618.3} + \frac{(27{,}270{,}890 - 27{,}274{,}776.7)^2}{27{,}274{,}776.7} \\ &\approx 166.6 + 168.5 + 0.54 + 0.55 \\ &\approx 336.19\end{aligned}$$

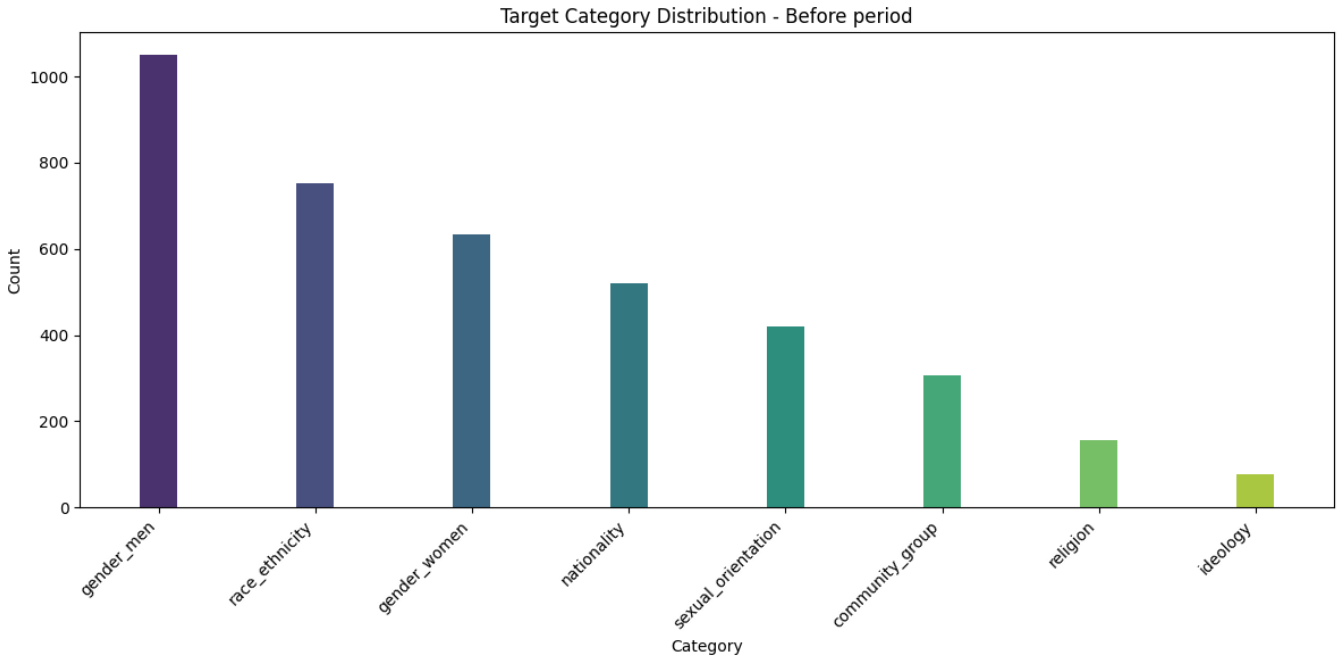

Fig. 3. Ordered frequency of hate speech targets in **Before** campaign phase

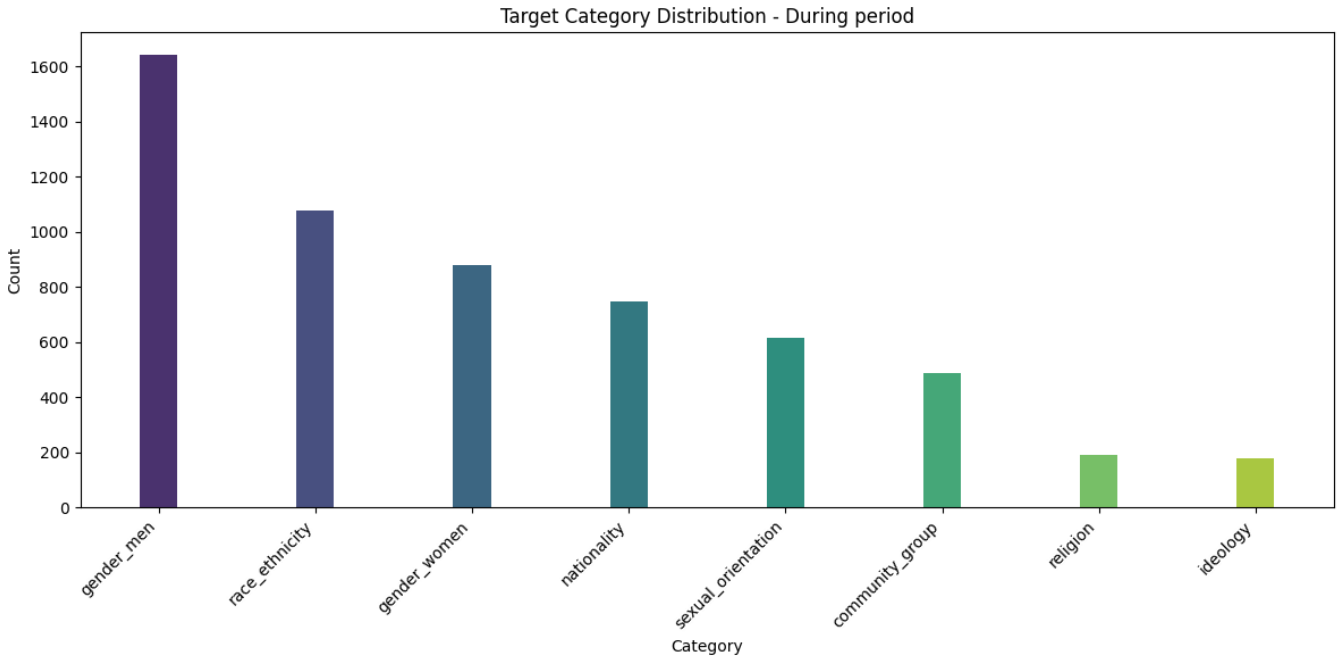

Fig. 4. Ordered frequency of hate speech targets **During** campaign phase

Before	After
"_meta": "retrieved_2nd_on": 0, "all_awardings": [], "approved_at_utc": 0, "approved_by": null, "archived": false, "associated_award": null, "author": "abcd", "author_flair_background_color": "", "author_flair_css_class": null, "author_flair_template_id": null, "author_flair_text": null, "author_flair_text_color": "abcd", "author_is_blocked": false, "awarders": [], "banned_at_utc": null, "banned_by": null, "body": "abcd", "can_gild": false, "can_mod_post": false, "collapsed": true, "collapsed_reason": null, "collapsed_reason_code": "", "comment_type": null, "controversiality": 0, "created": 0, "created_utc": 0, "distinguished": null, "downs": 0, "edited": false, "gilded": 0, "gildings": , "id": "abcd", "is_submitter": false, "likes": null, "link_id": "abcd", "locked": false, "mod_note": null, "mod_reason_by": null, "mod_reason_title": null, "mod_reports": [], "name": "abcd", "no_follow": true, "num_reports": 0, "parent_id": "abcd", "permalink": "abcd", "removal_reason": null, "replies": "", "report_reasons": [], "retrieved_on": 0, "saved": false, "score": 0, "score_hidden": false, "send_replies": true, "stickied": false, "subreddit": "abcd", "subreddit_id": "abcd", "subreddit_name_prefixed": "abcd", "subreddit_type": "abcd", "top_awarded_type": null, "total_awards_received": 0, "treatment_tags": [], "unrepliable_reason": null, "updated_on": 0, "ups": 0, "user_reports": []	"id": 0, "author": "abcd", "selftext": "abcd", "title": "abcd", "body": "abcd"

Table 9. Retrieved sample Reddit post before and after preprocessing. The actual content is replaced by *dummy data*