CHUNAV: Analyzing Hindi Hate Speech and Targeted Groups in Indian Election Discourse

FARHAN AHMAD JAFRI*, Jamia Millia Islamia, New Delhi, India
KRITESH RAUNIYAR*, Delhi Technological University, Delhi, India
SURENDRABIKRAM THAPA*†, Department of Computer Science, Virginia Tech, Blacksburg, United States
MOHAMMAD AMAN SIDDIQUI, Jamia Millia Islamia, New Delhi, India
MATLOOB KHUSHI, Department of Computer Science, Brunel University London, London, United Kingdom of Great Britain and Northern Ireland
USMAN NASEEM, School of Computing, Macquarie University, Sydney, Australia

In the ever-evolving landscape of online discourse and political dialogue, the rise of hate speech poses a significant challenge to maintaining a respectful and inclusive digital environment. The context becomes particularly complex when considering the Hindi language—a low-resource language with limited available data. To address this pressing concern, we introduce the **CHUNAV** dataset—a collection of 11,457 Hindi tweets gathered during assembly elections in various states. **CHUNAV** is purpose-built for hate speech categorization and the identification of target groups. The dataset is a valuable resource for exploring hate speech within the distinctive socio-political context of Indian elections. The tweets within **CHUNAV** have been meticulously categorized into "Hate" and "Non-Hate" labels, and further subdivided to pinpoint the specific targets of hate speech, including "Individual", "Organization", and "Community" labels (as shown in Figure 1). Furthermore, this paper presents multiple benchmark models for hate speech detection, along with an innovative ensemble and oversampling-based method. The paper also delves into the results of topic modeling, all aimed at effectively addressing hate speech and target identification in the Hindi language. This contribution seeks to advance the field of hate speech analysis and foster a safer and more inclusive online space within the distinctive realm of Indian Assembly Elections.

CCS Concepts: • **Information systems** → *Web searching and information discovery*; **Information retrieval**; • **Applied computing** → Law, social and behavioral sciences.

Additional Key Words and Phrases: Hate Speech, Natural Language Processing, Indian Election, Topic Modeling, Ensemble Methods

*These authors contributed equally to this research and are arranged in alphabetical order.
†Corresponding author: surendrabikram@vt.edu

Authors' Contact Information: Farhan Ahmad Jafri, Jamia Millia Islamia, New Delhi, Delhi, India; e-mail: farhanjafri88888@gmail.com; Kritesh Rauniyar, Delhi Technological University, Delhi, Delhi, India; e-mail: rauniyark11@gmail.com; Surendrabikram Thapa, Department of Computer Science, Virginia Tech, Blacksburg, Virginia, United States; e-mail: surendrabikram@vt.edu; Mohammad Aman Siddiqui, Jamia Millia Islamia, New Delhi, Delhi, India; e-mail: smohdaman9@gmail.com; Matloob Khushi, Department of Computer Science, Brunel University London, London, United Kingdom of Great Britain and Northern Ireland; e-mail: matloob.khushi@brunel.ac.uk; Usman Naseem, School of Computing, Macquarie University, Sydney, New South Wales, Australia; e-mail: usman.naseem@mq.edu.au.

ACM 2375-4702/2024/5-ART
https://doi.org/10.1145/3665245

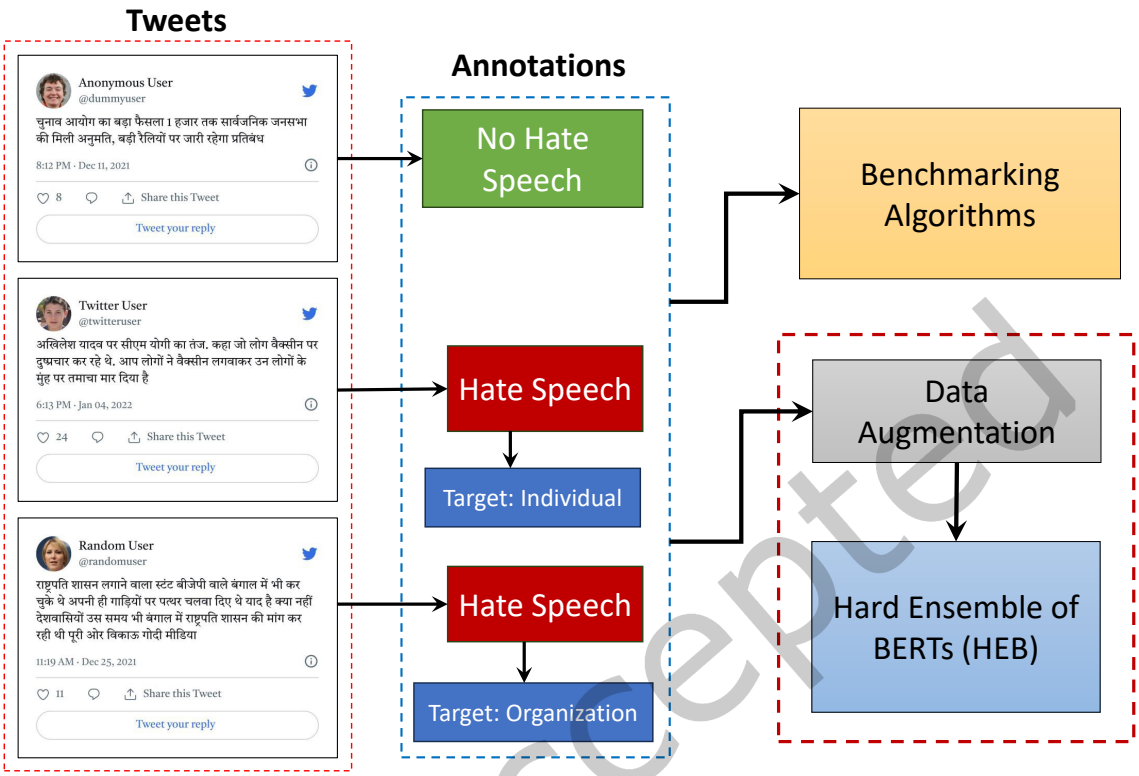

Fig. 1. In this paper, we annotate 11,457 tweets and benchmark them using different algorithms for hate speech and their target classification. We also propose a hard ensemble of BERT for the classification of hate speech and its targets.

1 INTRODUCTION

Since the advent of the Internet in the 1980s, the Internet has been one of the most unifying inventions of the 21st century [26]. With the boom in access to the Internet in the last three decades, social media has become the go-to place for the exchange of information and ideas [22]. Social media platforms have eradicated communication barriers and created a decentralized channel where people are allowed to engage in discourses in a democratic fashion, as defined by Amedie [4]. Factors such as widespread access and penetration in the everyday lifestyle of the masses and social media have an enormous effect on the sociopolitical environment and culture [51]. Platforms such as Twitter and Facebook have been an enormous part of this modern-era discourse. The growing emphasis on free speech on these platforms in recent years has led people of all backgrounds and political ideologies to find a safe environment on this platform to broadcast their beliefs and opinions [34, 36].

The intangible nature of the concepts of politics and personality, along with a tool such as social platforms, bring out a personal touch on ideas [25]. This personalization sometimes leads to disagreement between ideological opposites. A healthy discord may be a solution for many issues, but in an online environment with the mask of anonymity, there is always room for unethical conduct. Hate speech is defined by Djuric et al. [20] as *"abusive language targeted on a specific group of individuals regarding their characteristics such as religion, race, ethnicity, gender, etc"*. Hate speech is one of the significant problems with dealing with social media. Hate speeches on social media can range from direct personal attacks, name-calling, racism, homophobia, rape threats, etc. to indirect sarcastic comments [49, 61]. These hate attacks also come from different sources, such as hate comments that can be administered from personal accounts where the perpetrator can be identified, or they can come as

orchestrated attacks from unidentifiable troll accounts. The increase of hate speech in the digital age raises ethical questions for both the online platforms and the users [63]. Free speech is an ethical dilemma that plays an important role in the issue of hate speech. As social media evolves as a platform for online discourse, intensive work needs to be done on defining boundaries for free speech and making sure it does not transform into hate speech.

To make sure hate speeches are not tolerated, major social media companies have strict legal policies for the content that is allowed on the platforms. However, enforcing such policies in multiple languages and regional contexts is a difficult task. With the issue of resource shortages in the case of human moderation and keyword matching in the vast content, an automated artificial intelligence methodology needs to be perfected. An artificial intelligence approach can help solve basic problems such as improving speed and scale. Similar model architectures can be trained on multiple languages and their representative geographical connotations, helping to expand this as a solution to a global problem. Artificial models can help reduce human errors as opposed to human moderators, and they can also act as a helping hand by screening the content upfront and involving human support in grey areas. The development in the architecture of natural language processing models has enabled them to understand the complex contexts of the textual contents. Involving artificial intelligence along with human moderation can provide a robust approach to the problem [12, 23, 29].

In a diverse country like India, where society encompasses a wide array of sub-cultures and religious practices, the political machinery uses these differences very strategically and closely. Social media intake has seen a boom in India, according to the Internet and Mobile Association of India (IAMAI) report for 2022. The non-Internet population has become a minority for the first time. With an estimated 900 million Internet users in 2025[1], social media is changing the traditional mass communication channels for information distribution to more personalized social media feeds. The option of broadcasting information to the public has changed for political parties as well, from centralized channels such as newspapers to decentralized social media platforms. At times of elections in India, there has always been increased interaction between the political parties and the general public. From highlighting agendas to attacks between competing parties, elections are a sensitive time for balancing the line between discourse and unethical territory. During the 2014 parliamentary elections, political parties have extensively used social media platforms to engage with the voters, as explained by Narasimhamurthy [44]. Political activities on social media have not only come from political parties officially but also from anonymous accounts, where the risks of hate speeches and fake news grow.

With the manipulation of algorithms, botnets, and fake accounts, digital propaganda is a useful weapon to shape public opinion online, as described by Neyazi and Ahmed [46]. Hate propaganda is an old tactic used during election season, and during this time, online hate speeches are also spiked. From criticism of the opposition to name-calling and, in some cases, blatant swearing and hate in India, hate speeches tend to widen an already present gap within communities, fostering generational hate and, in some cases, communal violence and hate crimes; hence, this is a very sensitive issue that needs to be controlled in an online environment. Apart from social issues between communities, on a personal basis, hate speeches have affected people on a behavioral, social, and normative level, as demonstrated by Bilewicz et al.[10].

Hindi, the most widely spoken language in India, is spoken by a significant majority of the population, accounting for 43.6 percent, according to the 2011 Census. In contrast, English is the primary language for only 0.02 percent of the population[2]. The vast difference shows Hindi is a prominent language in rural and urban India. However, despite its widespread use, Hindi faces challenges in terms of resources, especially in the realm of natural language processing (NLP). The development of NLP models that can effectively accommodate the diverse range of regional languages worldwide poses a considerable barrier to mitigating hate speech online. To

[1]https://www.iamai.in/media

[2]https://censusindia.gov.in/nada/index.php/catalog/42561

address this issue and effectively moderate the vast amount of content on digital platforms, the deployment of language models is crucial. While English enjoys a wealth of resources and attention in the development of NLP models, major regional languages like Hindi often lag behind in this regard. Hindi, written in the Devanagari script, boasts a complex structure comprising 14 vowels and 33 consonants. Despite its rich significance, Hindi remains a low-resource language in the NLP community. Furthermore, Hindi is not a monolithic language; it encompasses a variety of regional dialects across different states in India. To accurately address local concerns and effectively combat hate speech, it is imperative to include languages like Hindi in hate speech analysis. Models primarily trained in English often fall short in mapping the nuances of geographical contexts and linguistic diversity that are so crucial in understanding and combating hate speech. In a context where the resource is limited, the analysis of hate speech to a target level is almost non existent. Incorporating low-resource languages like Hindi into NLP research and development efforts not only aids in increasing local engagement and collaboration but also ensures that hate speech moderation tools are more representative and effective across diverse linguistic and cultural landscapes.

In this paper, the **CHUNAV** dataset introduced, which serves as a valuable resource for addressing the critical issue of hate speech analysis in the Hindi language context. **CHUNAV** comprises 11,457 tweets in Hindi, collected during the state elections of Uttarakhand, Goa, Punjab, and Uttar Pradesh. These tweets have been meticulously labeled into two categories: "Hate" and "Non-Hate". Additionally, we have further categorized the hate speech targets into "Individual", "Organization", and "Community" allowing for a precise identification of the specific targets of hate speech. Our contributions in this work encompass the following key elements:

- In this paper, we propose **CHUNAV**, a dataset for **C**ategorizing **H**ate Speech and **UN**veiling **A**ssociated **V**ictim Groups (targets) in discourse related to Indian Assembly Election. We devise a robust annotation schema for manually annotating the dataset.
- Performed a detailed analysis of the corpus using various topic modeling techniques to gain more insights into the election discourse.
- Trained and provided benchmark models, which can be employed to gauge the effectiveness of hate speech detection and target identification in the Hindi language.
- We introduce a novel approach that combines ensemble methods and oversampling techniques to effectively address the challenge of hate speech detection and target identification in the Hindi language. This method outperforms the baselines for tackling hate speech on digital platforms, especially within the context of the Indian Assembly Elections.

These contributions collectively aim to advance the field of hate speech analysis and target identification in the Hindi language, with specific relevance to discourse related to Indian Assembly Elections, while providing a benchmark for evaluating the performance of hate speech detection models.

2 RELATED WORKS

Numerous studies focusing on computational methods for political discourse analysis and detecting hate speech in social media have been published in recent years. We go over important research in hate speech identification in the ensuing subsections.

2.1 Semantic Web and Political Discourse

The Internet is a vast platform where various forms of semantics and discussions take place. Among these, political discourse plays a significant role, reflecting a wide spectrum of debates and exchanges. Political discussions on the Internet can range widely, from intense and aggressive debates to educational and productive exchanges. They can be divided into several categories, such as antagonistic discourse, which is marked by resistance and disagreement, and deliberative discourse, which promotes reasoned discussion and consensus [14]. It's worth

noting that this diversity of political discourse online has implications for the broader understanding of political communication and engagement in the digital age. Different forms of political discourse, from antagonistic debates to deliberative exchanges, shape the way individuals interact, learn, and form their political opinions in the virtual space. Therefore, examining and analyzing the dynamics of online political discussions is crucial in comprehending the evolving landscape of political participation and discourse in the age of the Internet [14].

Numerous studies have been undertaken to better understand the political discourse on the Internet. Linguists have traditionally used political discourse analysis (PDA) to analyze political discourse and speeches. While these manual methods offer accuracy, they are less practical when dealing with the vast volumes of user-generated content that flood the Internet. Thus, researchers have widely been using automated tools and techniques for the analysis of political discourse [62]. For instance, Calderón et al. [11] used topic modeling approaches to understand the discourse and hate speech against immigrants on Twitter. Similarly, Bilbao-Jayo and Almeida [9] did a comprehensive analysis of political tweets across seven policy domains and their respective sub-domains, providing insights into the online discourse landscape. They also assessed the abilities of various Natural Language Processing (NLP) models to identify tweets belonging to different domains and sub-domains. Furthermore, Torregrosa et al. [65], through various qualitative and quantitative analyses, showed that social media platforms like Twitter serve as a *sentiment thermometer* to understand various sentiments surrounding different political ideologies. These studies collectively underscore the importance of analyzing social media content to gain insights into the evolving landscape of political discourse and public sentiment.

In this complex landscape of political discourse on the Internet, hate speech presents moral and societal issues in political debate and on the web. It is defined by insulting, biased, or abusive language Bhandari et al. [7]. It can sabotage constructive dialogue, promote division, and worsen the atmosphere on the Internet [68]. Therefore, there are efforts to identify and combat hate speech to promote more respectful and productive discourse. This leads to the need for advanced methods and technologies for hate speech detection, particularly in the context of different languages. With this background, the next subsection delves into the previous studies and challenges of hate speech detection in various languages.

2.2 Hate Speech Detection in High-Resource Languages / Other Languages

Hate speech detection on the Internet is an essential research area, focusing on both hate detection and hate classification [37]. Hate detection typically involves working with datasets where samples are categorized as either hate speech or non-hate speech, while hate classification goes beyond this binary distinction to consider various aspects of hate speech, including its nature, intensity, and direction [6, 70]. Researchers have made significant contributions in this domain, developing datasets and methodologies to tackle hate speech detection challenges [35].

One notable effort in this regard is the creation of a multilingual hate speech dataset by Ousidhoum et al. [48]. This dataset includes data in English, French, and Arabic, allowing for the analysis of hate speech across different languages. Their research demonstrated the effectiveness of deep learning models, which outperformed traditional Bag-of-Words (BOW)-based methods in multi-label classification tasks. Another approach involved crowd-sourcing a lexicon of hate speech terms to compile a collection of tweets, as seen in the work by Davidson et al. [18]. The tweets were categorized into three groups based on their content: offensive language, hate speech, and non-hate speech. This study shed light on the categorization differences between sexist, racist, and homophobic content. Similarly, Waseem and Hovy [68] created a dataset by annotating 16,914 tweets containing sexist and racist content. These examples are not hierarchically placed, and the sexist and racist labels are independent of each other.

Moreover, datasets have been created by analyzing content from various sources. de Gibert et al. [19] generated a sentence-level dataset from the white supremacist forum Stormfront, providing annotations for 9,916 sentences

with binary classes of hate and non-hate. This dataset served as a foundation for developing machine learning models, including SVMs, CNNs, and LSTMs. The HateXplain benchmark dataset, introduced by Mathew et al. [38], contains 20,148 annotated data points with labels such as hate, offensive, or normal. It also identifies target communities and the text portions (rationales) supporting the categorization.

Furthermore, researchers have explored cyberbullying and offensive terms, conducting manual annotations on datasets, as exemplified by Reynolds et al. [53]. They identified offensive terms commonly used in cyberbullying posts and categorized them based on their derogatory connotations, calculating the frequency and weighted average severity of these terms.

Additionally, researchers have focused on specific regions, languages, and contexts. Thapa et al. [62] released the NEHATE dataset, comprising 13,505 Nepali tweets manually reviewed to identify instances of hate speech in the context of conversations about local elections in Nepal. The tweets are categorized based on their intended targets, including community, individuals, and organizations. In the context of the Indonesian language, Alfina et al. [3] introduced a hate speech detection dataset with a specific focus on religious aspects. They employed various classification models, including Naïve Bayes, Support Vector Machine, Bayesian Logistic Regression, and ensemble methods, to analyze this dataset. Notably, they utilized word n-gram features in conjunction with the Random Forest and Decision Tree algorithm.

Finally, researchers have acknowledged the unique challenges posed by hate speech detection in languages like Hindi. These challenges stem from linguistic complexity, dialects, script variants, and a lack of linguistic resources and annotated data. Researchers, such as Velankar et al. [67], have highlighted the difficulties in developing hate speech detection programs for Hindi.

2.3 Hate Speech Detection in Hindi Language

Hate speech detection in the Hindi language poses unique challenges due to limited linguistic resources and research in comparison to high-resource languages like English. The scarcity of data and research in the Hindi language hampers the development of effective hate speech detection models. Nevertheless, there are notable efforts and datasets that contribute to this emerging field.

For instance, Safi Samghabadi et al. [56] developed an end-to-end deep learning approach using data collected and annotated in three languages: Hindi, Bengali, and English. The data focus on aggression and misogyny, with both labels having multi-class labels to denote the severity of the contents present in the text. This research demonstrates the potential for cross-lingual hate speech detection, even in low-resource languages. Additionally, Bhardwaj et al. [8] has annotated Hindi language data collected from Twitter into different labels based on the content in them, including fake news, hate speech, offense, and defamation. These labels are independent of each other, and the data is gathered from various sources. For evaluation, they use m-BERT for classification, demonstrating the applicability of multilingual models in hate speech detection.

Researchers like Velankar et al. [67] have explored various deep-learning classification models to detect hate speech in Hindi and Marathi Twitter data. Models such as CNN, BERT, LSTM, and BiLTSM are used, and a hierarchical approach for multi-class classification is proposed. This research highlights the adaptability of advanced NLP techniques to low-resource languages like Hindi. Furthermore, Rahul et al. [50] experimented with different combinations of models, including CNN, Bi-LSTM, and GRU, on Hindi-English code-mixed language. Their approach involved supplying these models with character embeddings and utilizing attention layers to enhance their performance. This research addresses the challenges of multilingual and code-mixed content in hate speech detection.

It's essential to acknowledge that while NLP has made significant progress in the English language, challenges such as a lack of pre-training data, unequal resource availability, and constrained processing power have hindered its development in Hindi [32]. One major obstacle is the insufficiently large corpus available for the Hindi

Table 1. Summary of related datasets

Works	Year	Data Source	Language	Objective	Size	Sub-classes/Targets	Context
de Gibert et al. [19]	2018	Website	English	Cyberbullying	9,916	✗	General discourse
Mollas et al. [41]	2022	YouTube and Reddit	English	Hate speech	998	Race, Disability, Sexual Orientation, Violence, Generalised vs Directed, National Origin, Religion, and Gender	General discourse
Zampieri et al. [71]	2019	Twitter	English	Offensive Language	14,100	Group, Individual, and Other	Social media discourse
Sitaula et al. [58]	2021	Twitter	Nepali	Sentiment Analysis	33,247	✗	COVID 19
Mossie and Wang [43]	2018	Facebook	Amharic	Hate Speech	6,120	✗	General discourse
Arshad et al. [5]	2023	Twitter	Urdu	Hate Speech	7,800	✗	Religious Hate
CHUNAV (Ours)	2024	Twitter	Hindi	Hate speech	11,457	Individual, Organization, Community	Election in India

language. Thus, realizing this need, we release **CHUNAV**, an annotated dataset containing over 11,457 tweets. We believe that this dataset serves as a crucial resource for advancing Hindi-language NLP, particularly in the context of hate speech detection.

In the context of related hate speech detection datasets, Table 1 provides a comparative summary of hate speech datasets in different languages, shedding light on the current state of the landscape and highlighting the need for further research and resources in low-resource languages like Hindi.

3 DATASET

In this section, we describe the methods of dataset collection, annotation criteria, and statistics.

3.1 Dataset Collection

A comprehensive analysis was conducted by systematically collecting tweets through the Twitter API[3], spanning the period from November 1, 2021, to March 9, 2022. This time frame was chosen to coincide with India's state election campaign, and the investigation aimed to assess the prevalence and impact of hate speech in political discourse during this critical phase. It's worth noting that the official election results were declared on March 10, 2022, after all votes had been registered and counted. To gather the dataset for this analysis, the Twitter API was employed. Various hashtags, including #indianelections, #goaelections, #manipurelections, #UPelections, #UttarakhandElections, and #PunjabElections, both in their plural and non-plural forms, were utilized as search criteria to collect tweets related to the state elections. This approach ensured a comprehensive and diverse dataset that covered a wide range of election-related discussions on Twitter. Fig. 2 shows the wordcloud of the complete dataset collected.

[3]https://developer.twitter.com/en/docs/twitter-api

Fig. 2. Wordcloud for the complete dataset

3.2 Filtering Criteria

We established criteria to filter the tweets for our dataset, and one of them was to save tweets that mostly comprised Hindi. The work retained tweets that had a few non-Hindi terms or phrases like *share*, *retweet*, etc. and removed tweets that were judged to be uninformative, such as spam or ads, as well as tweets that only utilized election-related hashtags for the purpose of spamming. As a result of these criteria and the data refinement process, final dataset containing 11,457 tweets was prepared. Each tweet was carefully labeled and text-annotated, ensuring that the dataset was well-prepared for the analysis of hate speech in the context of the Indian state elections. Additionally, we disregarded tweets that lacked context about the assembly election and that may have been impacted by local circumstances since they could generate uncertainty and make it more difficult to accurately classify hate speech.

3.3 Annotation Process

To ensure precise and consistent annotation of the tweets, we developed a comprehensive annotation scheme using a two-phase annotation technique. The annotation process was conducted by a group of three annotators, all of whom possessed a deep understanding of the Hindi language and the political context in India. In the initial phase of annotation, the annotators collectively annotated 200 tweets, focusing on the identification of hate speech within the tweets and the identification of the specific targets of that hate speech. While this initial phase was informative, some ambiguity in the annotation instructions became apparent. With the revised instructions in place, a second phase of annotations was carried out, encompassing an additional 100 tweets. Despite these clarifications, certain inconsistencies in the annotations persisted. To address these inconsistencies, the annotators engaged in a discussion with expert annotators during a dedicated meeting. This collaborative effort led to the refinement of the annotation instructions, making them clearer and more accurate for the subsequent stages of the annotation process. By iteratively improving the instructions and open communication among the annotators, we tried to ensure that the annotation of hate speech and its targets in the dataset was as precise

and consistent as possible. This meticulous approach was crucial in generating a high-quality dataset for our analysis.

3.4 Annotation for Hate Speech

The data was annotated for two primary labels: *Hate Speech* and *Non-Hate Speech*, to classify the tweets. To achieve this, guidelines were given to the annotators to label the tweets. If there was uncertainty among annotators about the appropriate label for a tweet, it was tagged as 'Non-Informative' and subsequently removed from the dataset. The annotation guidelines are outlined below.

Hate Speech: In the context of Indian election campaign, the tweets that frequently aimed at particular groups due to their political beliefs or associations and conveyed hostility or aggression towards them were labeled as *hate speech*. It also involved employing satirical content to spread harmful messages with the purpose of belittling or degrading a specific political group or individual. The text included hateful topics, such as direct insult, discriminatory language based on sexual orientation or race, or targeting of minority groups. The annotators particularly looked for the following when annotating for the presence of hate speech.

- **Specific dialect:** During Indian elections, hate speech frequently focused on specific groups due to their political convictions or associations. This involved using language that belittled, diminished or devalued a particular political crowd or individual.
- **Conflict and violence:** Hate speech frequently conveys aggression directed at a specific group or an individual. This encompassed the use of language that supports violence or conflict towards particular political parties or individuals.
- **Malicious satirical content:** Hate speech during Indian elections employed satires to spread destructive messages aimed to insult, humiliate, or dehumanize a certain political party or person.

Additionally, it is crucial to acknowledge that both satirical and sarcastic posts can serve as modes for conveying hate speech, and they can pose challenges when it comes to identification. Hate speech can be masked with sarcasm and satire to make it more covert and difficult to identify. Satirical posts can also be used to convey hate speech in a way that is meant to be ironic or amusing but still has the potential to be offensive. Annotators were provided examples of such tweets. This helped us to get accurate annotations for the presence of hate speech.

No Hate Speech: A non-hateful tweet presents events or the viewpoints of others in an impartial manner and does not contain any offensive or hateful material. The attributes that distinguish non-hate speech were established within the annotation guidelines. The tweet was annotated as non-hateful if it contained legitimate criticism towards political persons or parties. We also labeled tweets as non-hateful if they contained information that is truthful and educational with the absence of antagonism, false information, or fake news. Additionally, non-hate speech during the Indian election campaign refrained from targeting specific groups of individuals based on their political beliefs or affiliations. These guidelines were furnished to annotators to assist them in recognizing and categorizing tweets as non-hate speech. To ensure clarity in the guidelines, we elaborated on the key attributes of non-hate speech as follows:

- **Helpful critique:** Non-hateful discourse frequently encompassed a constructive critique of political figures and parties. This type of discourse also involved assessments of political occurrences and developments.
- **Insightful and reliable:** Objective and informative content is frequently observed during political events. This content often comprises news and analyses related to political agenda, providing a well-rounded perspective on the campaign. Such tweets are deemed non-hateful.

3.5 Named Entity Recognition (NER) for Hate Speech Tweets

We started using Named Entity Recognition (NER) techniques in our effort to identify specific targets within hate speech situations. NER, a core Natural Language Processing (NLP) technique, is intended to identify and categorize entities within a given text corpus, such as names of people, groups, and places. However, it became apparent during our initial investigation that NER alone was not consistently delivering accurate results in identifying the intended targets of hate speech incidents.

For instance, as illustrated in Fig. 3 (a), the term 'कमल' refers to the Bharatiya Janata Party of India[4] but is also a commonly used name in India. Thus, when experimenting with named entity recognition, we found it to be misclassified as a person entity. Similarly, Fig. 3 (b) showcases situations where multiple NER elements can complicate the accurate identification of targets, and certain words were not identified correctly. Additionally, lesser-known political parties or less frequent names faced issues with proper identification, as seen in Fig. 3 (c), where 'रालोसपा' is a lesser-known party, and NER algorithms identified multiple words with incorrect entity classes. Thus, we deemed that NER was not suitable for our analysis.

Fig. 3. Some examples of tweets that underscore problems with identification of targets of hate speech with NER algorithms

To address these limitations, we opted for a manual annotation approach. In this method, annotated a dataset of hate speech samples to categorize the targets into different classifications. These classes fall into three main categories: INDIVIDUAL, which refers to specific individuals targeted; ORGANIZATION, which includes targeted institutions; and COMMUNITY, which denotes instances in which hate speech targets larger socioeconomic groupings or communities. This manual technique significantly improved the precision and specificity of target identification, enabling a more thorough exploration of hate speech dynamics.

3.6 Annotation for Targets of Hate Speech

The process of annotating hate speech within our dataset involved further categorizing it into three distinct subgroups: *Individual*, *Organization*, and *Community*. These categories were defined and annotated based on the following guidelines:

- **Individual:** Within the scope of our dataset, an individual is defined as a self-reliant person engaged in political activities. This category encompassed figures such as politicians, political contenders, activists, journalists, and anyone participating in political discussions or having a vested interest in the election. Notably, some of the prominently mentioned personalities in our dataset are Yogi, Sidhu, and Akhilesh Yadav.

[4]https://www.bjp.org

- **Organization:** In the context of Indian elections, an organization is a structured group of people established to achieve specific political objectives. Examples of such organizations in our context include political parties like the Congress or the BJP, as they are groups that advocate for particular social or policy causes.
- **Community:** In the context of our dataset related to Indian elections, a "community" is defined as a collective of people who hold similar convictions or attributes, such as sharing the same social class, caste, faith, religious groups, background, etc.

These guidelines provided a clear framework for annotating and categorizing the targets of hate speech, enabling a more comprehensive understanding of the different dimensions of hate speech within the political discourse during the election campaign. Table 2 presents some examples of non-hateful speech, hate speech, and targets of hate speech.

Table 2. Table represents the few examples from the CHUNAV Dataset

Tasks	Labels	Examples	Translation
Hate Speech Detection	Hate Speech	गोवा की गुनहगार कांग्रेसभारत की आजादी के बाद भी गोवा को 15 साल तक गुलामी झेलनी पड़ी, क्योंकि कांग्रेस के प्रधानमंत्री नेहरू जी को अपने इमेज की चिंता थी। 15 साल तक गोवा पर जुल्म की जिम्मेदार है कांग्रेस।	Congress is the culprit of Goa. Even after India's independence, Goa had to suffer slavery for 15 years because Congress Prime Minister Nehru ji was worried about his image. Congress is responsible for atrocities on Goa for 15 years
	No Hate Speech	Delhi : चुनाव आयोग का बड़ा फैसला 1 हजार तक सार्वजनिक जनसभा की मिली अनुमति, बड़ी रैलियों पर जारी रहेगा प्रतिबंध	Big decision of Election Commission, permission given for public gathering up to 1 thousand, ban on big rallies will continue
Targets of Hate Speech Detection	Individual	अखिलेश यादव पर सीएम योगी का तंज. कहा जो लोग वैक्सीन पर दुष्प्रचार कर रहे थे. आप लोगों ने वैक्सीन लगवाकर उन लोगों के मुंह पर तमाचा मार दिया है	CM Yogi's taunt on Akhilesh Yadav. Said those who were spreading false propaganda on the vaccine. You people have slapped those people on the face by getting them vaccinated.
	Organization	राष्ट्रपति शासन लगाने वाला स्टंट बीजेपी वाले बंगाल में भी कर चुके थे अपनी ही गाड़ियों पर पत्थर चलवा दिए थे याद है क्या नहीं देशवासियों उस समय भी बंगाल में राष्ट्रपति शासन की मांग कर रही थी पूरी ओर विकाऊ गोदी मीडिया	BJP people had also done the stunt of imposing President's rule in Bengal and had stoned their own vehicles. Do you remember, countrymen, even at that time the entire Vikau Godi media was demanding President's rule in Bengal.
	Community	प्रधानमंत्री मोदीजी ने तमिलनाडु में 11 नए मेडिकल कालेज का वीडियो कांफ्रेंस द्वारा उद्घाटन किया। उत्तराखंडी को कालेज नहीं चाहिए उसे तो चुनाव के दौरान 11 बोतल मिल जाए उसी से खुश हो जाएगा। नशा नही, रोजगार दो । जाग पहाड़ी जाग ! जाग उत्तराखंडी जाग !	Prime Minister Modiji inaugurated 11 new medical colleges in Tamil Nadu through video conference. If an Uttarakhandi does not want a college, he will be happy if he gets 11 bottles during the elections. Don't give drugs, give employment. Wake up hill, wake up! Wake up Uttarakhandi!

3.7 Dataset Statistics and Analysis

We assessed the quality of our annotation using Fleiss' Kappa score [62]. The Kappa score (κ) for hate speech vs non-hate speech annotation was 0.75 whereas the kappa score for target identification was around 0.69. Our new **CHUNAV** dataset had 11,457 tweets, with 970 (8.46%) classified as "Hate Speech" and 10,487 (91.53%) marked as "No Hate". Furthermore, the "Hate Speech" is divided into three targets: "Individual" having 400 (3.49%), "Organisation" having 415 (3.62%) and "Community" having 154 (1.34%) tweets (Table 3). The statistics in the dataset represent a real-world scenario in which the majority of posts are neutral and just a minority involve

hate speech. Table 4, shows the state-wise distribution of political tweets. From the Table 4, Punjab and Uttar-Pradesh have the highest hate speech ratio in tweets. Fig. 4, on the other hand, shows the histogram of number of words and characters used in tweets with targeted hate speech.

Table 3. Dataset Statistics for "CHUNAV" dataset. The values in parentheses are average characters per tweet (Avg. Char) and average words per tweet (Avg. words) are calculated after preprocessing of text.

Task	Labels	#Tweets	Avg. Char	Avg. words
Hate Speech Detection	Hate	970	158.37 (145.02)	27.39 (25.37)
	Non-Hate	10487	150.06 (136.70)	24.51 (22.69)
Identification of Targets	Individual	400	174.91 (141.70)	28.44 (25.68)
	Organisation	415	175.36 (141.38)	28.36 (25.41)
	Community	154	164.24 (135.52)	26.23 (23.52)

Table 4. State-wise Label and Target Distribution

States	Labels		Targets		
	Hate	Non-Hate	Individual	Organisation	Community
Punjab	259	1752	147	72	40
Uttar Pradesh	246	2301	82	121	43
Goa	197	1996	68	93	36
Uttarakhand	144	2080	62	60	22
Manipur	55	884	14	34	7

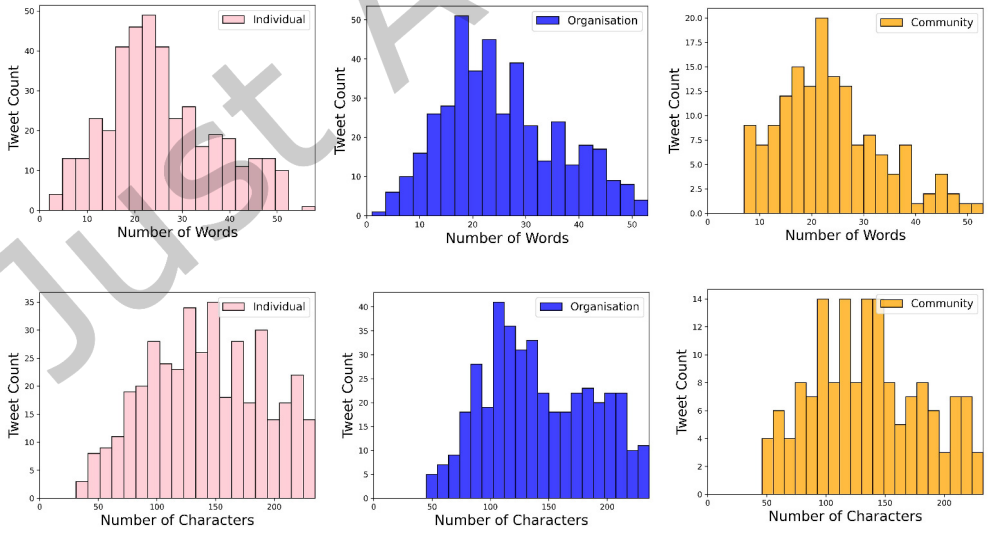

Fig. 4. Histogram of the number of words and number of characters used in Targets of Hate Speech tweets

3.8 Exploratory Data Analysis

As a component of our exploratory data analysis, we identify the top 10 words within our dataset as well as the top 10 words associated with various classes and sub-classes. To accomplish this, TF-IDF (Term Frequency-Inverse Document Frequency) was employed to extract significant words based on their importance weights. The statistical method known as TF-IDF plays a vital role in assessing the significance of a word within a collection of documents. The TF-IDF score comprises two components: firstly, the TF (Term Frequency) element, which indicates how often a term occurs within a particular document, and secondly, the IDF (Inverse Document Frequency) element, which highlights the word's prevalence or rarity across the entire document set. To compute the TF-IDF score, multiplication of the TF (Term Frequency) score and the IDF (Inverse Document Frequency) score.

$$tf\ idf(t, d, D) = tf(t, d).idf(t, D) \tag{1}$$

where,

$$tf(t, d) = log(1 + freq(t, d)) \tag{2}$$

$$tf(t, d) = log(\frac{N}{count(d \in D : t \in d)}) \tag{3}$$

Simple explanation of the formula:

$$TF = log(\frac{C}{D}) \tag{4}$$

$$IDF = log(\frac{E}{F}) \tag{5}$$

$$TF - IDF = TF * IDF \tag{6}$$

where,
C = Frequency of the specific word's appearance within the document.
D = Overall word count within the document.
E = Total number of documents within the corpus.
F = Word is present in multiple documents within the corpus.

From Table 5, it is evident that words चुनाव (Election), मतदान (Vote), पंजाब (Punjab), and कांग्रेस (Congress) hold substantial importance across most of the tasks. In Table 5, each individual word is paired with its corresponding translation and an accompanying TF-IDF score. Table 5 presents data pertaining to three categories: *All Posts* extracted from the dataset, *No-Hate Speech Posts*, and *Hate Speech Posts*. Table 6 displays the top-10 words associated with the classes denoting the 'Targets of Hate Speech', categorized into three groups: *Individual*, *Community*, and *Organization*. Fig. 2 offers a straightforward representation of the words present in our dataset through the use of a word cloud.

Table 5. Top-10 words that occur most frequently in the entire dataset, as well as within each category of Hate Speech and Non-Hate Speech.

All Posts			No Hate Speech Posts			Hate Speech Posts		
Words	**Translation**	**TF-IDF**	**Words**	**Translation**	**TF-IDF**	**Words**	**Translation**	**TF-IDF**
चुनाव	Election	0.1448	चुनाव	Election	0.1487	कांग्रेस	Congress	0.0991
मतदान	Vote	0.0916	मतदान	Vote	0.0915	पंजाब	Punjab	0.0838
कांग्रेस	Congress	0.0810	गोवा	Village	0.0826	भाजपा	BJP	0.0831
गोवा	Village	0.0799	कांग्रेस	Congress	0.0825	चुनाव	Election	0.0678
पंजाब	Punjab	0.0776	विधानसभा	Assembly	0.0801	वोट	Vote	0.0660
विधानसभा	Assembly	0.0772	पंजाब	Pujab	0.0785	सरकार	Government	0.0657
उत्तराखंड	Uttarakhand	0.0616	उत्तराखंड	Uttarakhand	0.0667	पार्टी	Party	0.0619
पार्टी	Party	0.0616	पार्टी	Party	0.0625	केजरीवाल	Kejrival	0.0566
भाजपा	BJP	0.0597	भाजपा	BJP	0.0583	देश	Contry	0.0517
सिंह	Singh	0.0524	चरण	Phase	0.0491	गोवा	Village	0.0447

Table 6. Top-10 most frequent words in each target class. The TF-IDF scores are given for each word.

Target: Individual			Target: Community			Target: Organization		
Words	**Translation**	**TF-IDF**	**Words**	**Translation**	**TF-IDF**	**Words**	**Translation**	**TF-IDF**
केजरीवाल	Kejrival	0.1047	वोट	Vote	0.1264	कांग्रेस	Congress	0.1673
पंजाब	Punjab	0.0866	पंजाब	Punjab	0.0889	भाजपा	BJP	0.1489
मोदी	Modi	0.0818	चुनाव	Election	0.0744	बीजेपी	BJP	0.0961
अखिलेश	Akhilesh	0.0664	राम	Ram	0.0653	सरकार	Government	0.0895
कांग्रेस	Congress	0.0609	देश	Country	0.0635	पार्टी	Party	0.0846
चुनाव	Election	0.0555	मुस्लिम	Muslim	0.0616	पंजाब	Punjab	0.0749
सिद्धू	Sidhu	0.0536	लोग	People	0.0560	वोट	Vote	0.0707
जनता	People	0.0533	सिर्फ	Only	0.0466	चुनाव	Election	0.0698
सिंह	Singh	0.0513	पाकिस्तान	Pakistan	0.0433	जनता	People	0.0550
दिल्ली	Delhi	0.0482	जय	Victory	0.0279	गोवा	Village	0.0456

4 TOPIC MODELING

Topic modeling (TM) encompasses key concepts including 'words,' 'documents,' and 'corpora.' In this context, a 'word' serves as the foundational unit within discrete textual data, representing individual vocabulary elements that are uniquely indexed within a document. Meanwhile, the term 'document' pertains to an aggregation of N words. A corpus consists of a collection of M documents, and when we refer to multiple collections of this type, we use the term 'corpora' in the plural form. The term 'topic' pertains to the allocation of a predetermined vocabulary. In simpler terms, within a corpus, each document exhibits a unique distribution of the mentioned topics, determined by the specific terms it contains [45]. Over the past few years, there has been a significant focus within the machine learning (ML) and NLP domains on probabilistic graphical models. Among these statistical topic models, Latent Dirichlet Allocation (LDA) stands out as a robust framework for effectively characterizing and summarizing the content found in extensive collections of documents [39]. LDA has made a substantial mark on the fields of statistical ML and NLP, quickly establishing itself as one of the most favored techniques for probabilistic topic modeling within the realm of ML [64].

Provided a collection of text D containing M documents, where each document d contains N_d words, the process of Latent Dirichlet Allocation (LDA) [30] model D using the subsequent generative procedure:

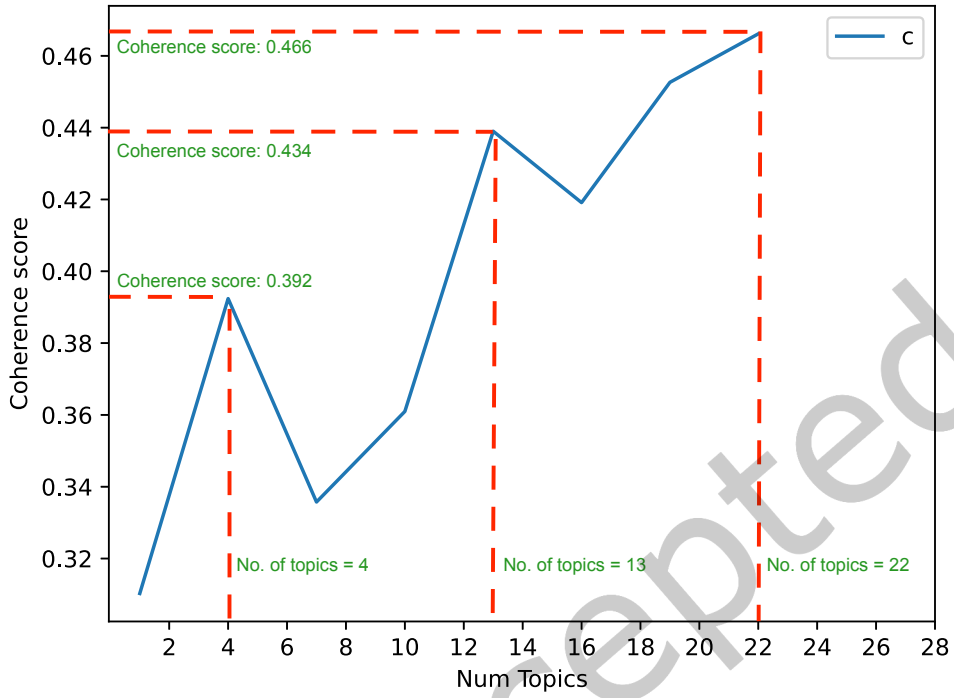

Fig. 5. Relationship between the number of topics and the coherence score

(1) Select a topic t (where t is an element of the set $\{1, \ldots, T\}$) by employing a multinomial distribution φ_t derived from a Dirichlet distribution with the parameter β.
(2) Select a multinomial distribution θ_d for document d (where d belongs to the set $\{1, \ldots, M\}$) using a Dirichlet distribution characterized by the parameter α.
(3) For a term w_n (where n ranges from 1 to N_d) within document d,
   (a) From the distribution θ_d, choose a topic z_n.
   (b) Pick a term w_n from the distribution φ_{zn}.

In the generative process described above, the words contained within documents are considered observed variables, while the remaining elements encompass latent variables (φ and θ) and hyperparameters (α and β). The computation and acquisition of the probability for the observed data D within a corpus are accomplished as follows in equation 7.

$$p(D|\alpha,\beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha)\left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn}|\theta_d)p(w_{dn}|z_{dn},\beta)\right)d\theta_d \tag{7}$$

In our dataset, the process of topic modeling involves several distinct phases aimed at achieving precise outcomes. Our research employs the LDA approach to simulate topics and create clusters for each topic discussion. Initially, we begin with our raw, unannotated dataset and initiate a data cleaning step, which involves the removal of extraneous characters such as hashtags and mentions. Additionally, duplicate data was eliminated to enhance the data's relevance and optimize it for our model, eliminated stopwords. Subsequently, we tokenize

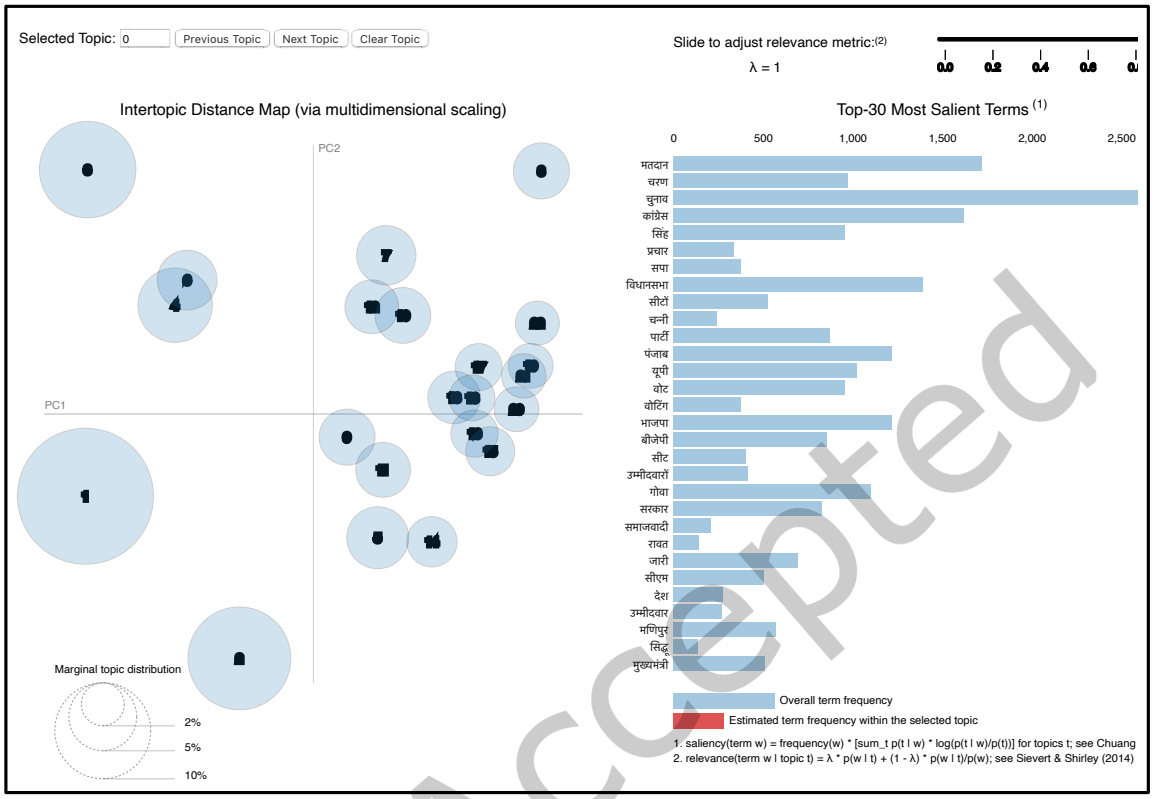

Fig. 6. Rejected TM (No. of topics = 22)

the data, fit it into the LDA model and carefully observed the results and fine-tune the number of topics to ensure relevance. Finally, visualized the outcomes and assign a descriptive topic label to each cluster.

We created topics after pre-processing the data and segmentation of the corpus into tokens. During this process, a meticulous examination of the clusters was conducted to ensure non-overlapping clusters. Our aim was to prevent the recurrence of the same words across two or more clusters and to evaluate the quality of the topic modeling, we employed coherence scores. The coherence score generally increases with the number of topics up to a point, indicating better topic quality and interpretability. However, beyond this optimal point, further increasing the number of topics may lead to a decrease in coherence, as topics may become too specific, overlapping, or less meaningful. These scores involve assessing the semantic proximity of high-scoring words within a topic's vocabulary. A higher coherence score indicates the generation of more accurate topics. Nonetheless, it is essential to note that high coherence scores do not always guarantee ideal results, as there is a possibility of clusters overlapping and still yielding a high coherence score.

In our analysis, as illustrated in Fig. 5, we observed the highest coherence score of 0.466 when utilizing 22 topics. Nevertheless, as depicted in Fig. 6, opting for 22 topics led to the creation of overlapping clusters. The subsequent choice for the number of topics could have been 13, as indicated in Fig. 5, where the coherence score exhibited a notable peak. Regrettably, the issue of overlapping clusters persisted even with 13 topics which is shown in Fig. 7. Following several iterations, our experiments ultimately revealed that employing 4 as the number of topics yielded well-defined clusters devoid of overlap with the coherence score of 0.392, as visualized

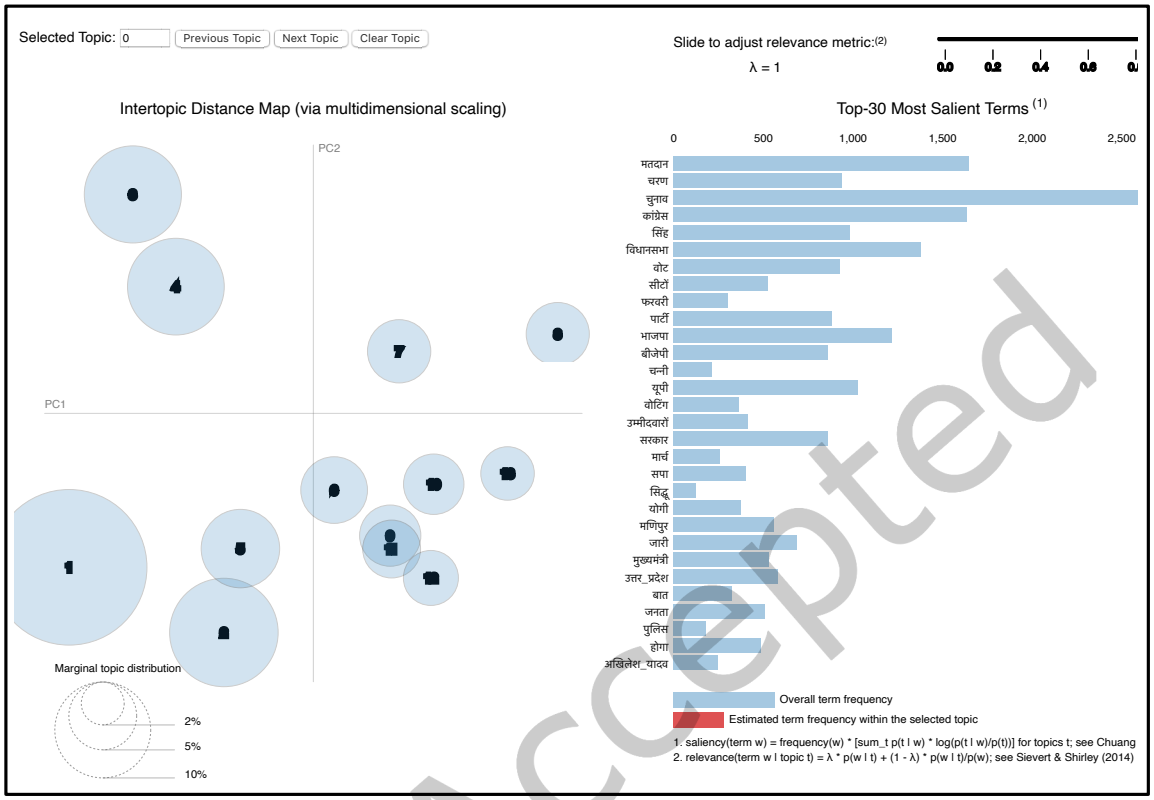

Fig. 7. Rejected TM (No. of topics = 13)

in Fig. 8. In the context of our LDA model, the hyperparameter α was configured to the 'auto' setting, and the hyperparameter β was retained at its default value by using the Gensim library [5]. The choice of hyperparameters, such as the number of topics and the Dirichlet prior parameters (α and β), significantly affects the probability distributions that the model learns. The optimum selection of these parameters is typically not fixed and might differ depending on the particular dataset and the objectives of the research. Choosing the correct number of topics is crucial; opting for too many can lead to fragmented and overlapping topics. The Dirichlet priors, alpha (α) and beta (β), regulate the sparsity of the topic distribution per document and the word distribution per topic, respectively. Usually, these hyperparameters are established using domain knowledge or identified by model selection methods. There is no generally ideal configuration for these parameters; the most suitable option typically depends on the context, including both the interpretability and the statistical characteristics of the model. We also conducted experiments involving perplexity, a metric that assesses the model's understanding of the data's underlying patterns and structure. However, we excluded perplexity from our analysis as it did not prove to be particularly useful in determining the optimal number of topics.

Table 7 provides a compilation of four topics, complete with the dominant words associated with each topic. Additionally, Fig. 9 offers a visual representation of these topics through wordclouds. Subsequently, relying on qualitative assessment, we proceed to elucidate and assign thematic labels to each of these four topics.

[5]https://radimrehurek.com/gensim/index.html

- **Indian Election and Places** - The first cluster of words encompasses terms related to the election process and significant geographical locations within India like विधानसभा (Legislative Assembly), सीएम (CM - Chief Minister), उत्तराखंड (Uttarakhand), पंजाब (Punjab).
- **Voting and Stages** - The second topic highlights the voting-related term and also represents various stages during the electoral process. Terms such as मतदान (Voting), चरण (Stage/Phase), सीटों (Seats), मतदाताओं (Voters) were used to depict fundamental elements of during the voting stage.
- **Indian Political Parties and Leaders** - The third cluster depicted the complex interplay involving political parties and their leaders. Words like सत्ता (Power), कांग्रेस (Congress), भाजपा (BJP), सिद्धू (Sidhu), अखिलेश यादव (Akhilesh Yadav) formed integral components of Indian political parties and prominent leaders.
- **Indian Politics and Governance** - The fourth topic generate the term related to structures, and governing bodies responsible for making decisions, passing laws, and ensuring a just and efficient operation of its functions. Few relevant words for this topic were सरकार (Government), वोट (Vote), विकास (Development), जनता (People), and देश (Country).

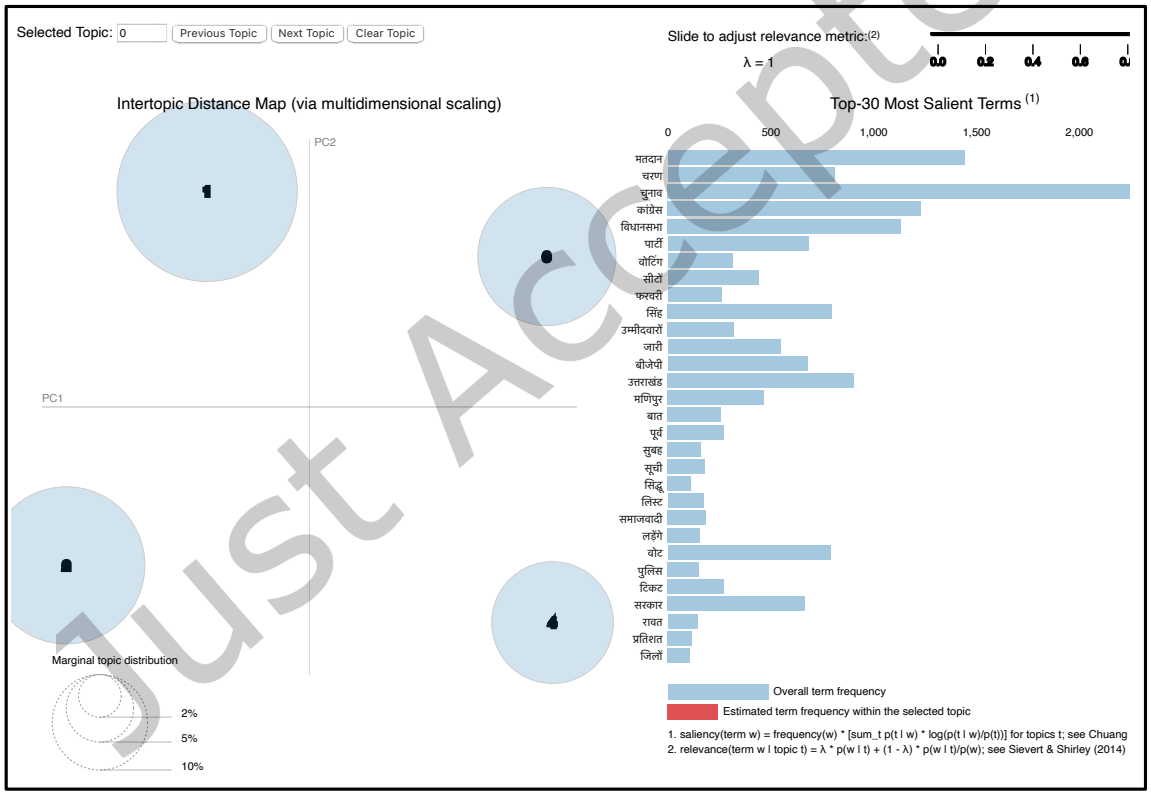

Fig. 8. Selected TM (No. of topics = 4)

5 METHODS FOR CLASSIFICATION OF HATE SPEECH AND TARGETS

In this section, we delve into various algorithms and methods for the classification of hate speech and its targets.

Table 7. Top-7 dominant words from each topic

Topic	Domain	Dominant Words
First topic	Indian Election and Places	चुनाव (Election), उम्मीदवारों (Candidates), नेता (Leader), विधानसभा (Legislative Assembly), सीएम (CM - Chief Minister), उत्तराखंड (Uttarakhand), पंजाब (Punjab)
Second topic	Voting and Stages	मतदान (Voting), चुनाव (Election), चरण (Stage/Phase), सीटों (Seats), आयोग (Commission), मतदाताओं (Voters), अपील (Appeal)
Third topic	Indian Political Parties and Leaders	पार्टी (Party), , सत्ता (Power), कांग्रेस (Congress), भाजपा (BJP), समाजवादी (Samajwadi), सिद्धू (Sidhu), अखिलेश यादव (Akhilesh Yadav)
Fourth topic	Indian Politics and Governance	सरकार (Government), वोट (Vote), विकास (Development), जनता (People), देश (Country), प्रधानमंत्री (Prime Minister), निर्देशन (Direction)

Indian Election and Places
Topic 1
(a)

Voting and Stages
Topic 2
(b)

Indian Political Parties and Leaders
Topic 3
(c)

Indian Politics and Governance
Topic 4
(d)

Fig. 9. Wordcloud for a particular topic.

5.1 Baseline Methods and Algorithms

We carried out experiments with various algorithms, ranging from conventional machine learning to cutting-edge transformer-based approaches, with the aim of building comprehensive baseline algorithms fit to detect hate speech and targets within instances of hate speech.

5.1.1 Machine Learning Algorithms. When it came to traditional **machine learning**, our method included a variety of well-known algorithms. We used Naive Bayes, Decision Trees, SVM and logistic regression. To generate features for these machine learning algorithms, TF-IDF vectorizer was used.

Naive Bayes: Naive Bayes [54] is a probabilistic classification algorithm that works based on Bayes' theorem. It's particularly effective for text classification tasks. Naive Bayes makes the *naive* assumption that features are independent of each other, which simplifies the calculation of probabilities.

Decision Tree: Decision Tree (DT) [16, 17, 55] is a hierarchical model that divides a dataset into smaller subsets based on features. These subsets are then categorized into classes based on the characteristics of the data. DT are known for their transparent and interpretable nature.

Support Vector Machines (SVM): Support Vector Machines [15] seek to find an optimal hyperplane that effectively separates data points belonging to different classes. It maximizes the margin between these data points, making SVMs suitable for classification tasks, especially in high-dimensional spaces.

Logistic Regression: Logistic Regression (LR) [60] is a statistical model used for binary classification. It estimates the probability of an input belonging to a particular category based on its features. LR is appreciated for its simplicity and interpretability.

5.1.2 Deep Learning Algorithms. In our approach, we also incorporated deep learning algorithms, specifically Bidirectional Long Short-Term Memory (BiLSTM) and Convolutional Neural Network (CNN)+BiLSTM. These neural network architectures are powerful tools for processing sequential and textual data.

BiLSTM: BiLSTM is a variant of the Long Short-Term Memory (LSTM) [27] neural network. It processes input data bi-directionally, allowing it to capture context and dependencies in both directions of a sequence. This is especially valuable in tasks that involve analyzing text or sequences of data.

CNN + BiLSTM: The CNN+BiLSTM architecture combines Convolutional Neural Networks (CNN) with Bidirectional Long Short-Term Memory (BiLSTM) networks. CNNs are effective in capturing local features and patterns in data, while BiLSTM is proficient in understanding sequential relationships. This combination leverages the strengths of both approaches for comprehensive data analysis [2].

In our implementation, we utilized the TensorFlow tokenizer to process text data and create features, enabling these deep learning models to analyze and extract meaningful information from the dataset.

5.1.3 Transformer-based Approaches. In our analysis, we also leveraged Transformer-based approaches, which have gained prominence in the field of natural language processing (NLP) due to their ability to capture long-range dependencies and understand contextual information. Specifically, we employed BERT-based models. A brief description of models is given below.

XLM-RoBERTa: This model is an XLM-Roberta-based one that was developed using data from 198 million tweets in various languages [13, 69]. It is a great choice in the field of multilingual understanding due to its broad, multilingual pretraining as a foundation. This model does an outstanding task of gathering up small contextual elements and is quick to understand how words work together to form phrases. Since there are so many different languages, this ability to capture contextual information is particularly impressive since it enables XLM-RoBERTa to perform wonderfully across a variety of linguistic varieties.

BERT (HAM - Hindi-Abusive-MuRIL): The Hindi-Abusive-MuRIL form of BERT, often known as HAM, stands for a targeted effort to identify and address abusive language in the context of Devanagari Hindi. This model explores on a variety of texts that covers the nuances of Hindi language usage [24].

BERT (HCMAM - Hindi-Codemixed-Abusive-MuRIL): BERT tackles the difficulty of spotting offensive words in the context of Hindi codemixed text. Codemixing, a combination of different languages inside a single sentence, presents special linguistic challenges. Hindi-Codemixed-Abusive-MuRIL, or BERT (HCMAM), arises from a base of training that spans Devanagari Hindi and other languages prevalent in the codemixed environment [52].

RoBERTa-Hasoc: A multi-class hate speech model called hate-roberta-hasoc-hindi was developed using the Hindi Hasoc Hate Speech Dataset 2021. The label mappings are None, Offensive, Hate, and Profane from 0 to 3. Its comprehensive exposure to full large-scale training and exploration of a huge and diverse text corpus relevant to hate speech supports its expertise.

Ensemble of BERTs: We also utilized an ensemble of BERT models to comprehensively understand the dataset and extract valuable insights. By combining multiple BERT models with diverse pre-training data and characteristics, we aimed to enhance the depth and breadth of our analysis. Ensemble learning involves aggregating the predictions of multiple models to make more informed decisions. In our case, the hard ensemble of

BERT (HEB) models allowed us to explore various aspects of the data, including different linguistic nuances, contextual understanding, and linguistic diversity. Details on the approach are found in section 5.3.

5.1.4 Modeling Choices for Algorithms. In our methodology, we aimed to explore a diverse range of algorithms and methods to comprehensively tackle the challenges of hate speech classification and target detection. Our selection of traditional machine learning algorithms, including Naive Bayes, Decision Trees, SVM, and logistic regression, was motivated by the rationale to establish baseline performance and understand the interpretability of simpler models in hate speech detection. Concurrently, we incorporated deep learning architectures such as BiLSTM and CNN+BiLSTM to delve into the capabilities of neural networks in capturing intricate textual patterns and dependencies, particularly relevant to the nature of hate speech. Furthermore, by leveraging transformer-based approaches like BERT variants (XLM-RoBERTa, BERT-HAM, and BERT-HCMAM), we aimed to explore the capabilities of state-of-the-art models in the contextual understanding of tweets to address the challenges posed by diverse linguistic contexts. Additionally, the ensemble of BERT models allowed us to explore the potential synergies between different pre-trained language models to enhance the robustness of our hate speech detection framework. Through this diverse experimental setup, we seek to explore more into the strengths and limitations of various algorithms, paving the way for future research directions in hate speech detection and mitigation particularly in the context of the Indian election.

5.2 Evaluation Metrics

In our analysis, we employed a range of evaluation metrics to assess the performance and effectiveness of our hate speech detection and target classification models. These metrics provided a quantitative understanding of the models' capabilities.

Accuracy is a fundamental evaluation metric that gauges the correctness of predictions made by our baseline algorithms for target identification within hate speech instances. It quantifies the ratio of correctly classified instances to the total instances in the dataset, providing an overall view of algorithmic performance.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

where,

$$TP : True\ Positive$$
$$TN : True\ Negative$$
$$FP : False\ Positive$$
$$FN : False\ Negative$$

Mean Absolute Error (MAE) contributes a granular assessment by calculating the average absolute difference between predicted and actual values across various target categories. This metric offers insights into the extent of prediction errors for each category, aiding in identifying areas where improvements are required.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|x_i - x| \tag{9}$$

where,

$$|x_i - x| = absolute\ error$$
$$n = number\ of\ error$$

Precision also known as Positive Predictive Value, measures the accuracy of the positive predictions made by a model. It is the ratio of correctly predicted positive instances to all instances predicted as positive.

$$Precision = \frac{TruePositive}{(TruePositive + FalsePositive)} \tag{10}$$

Recall also known as Sensitivity or True Positive Rate, measures the model's ability to capture all the positive instances in the dataset. It is the ratio of correctly predicted positive instances to all actual positive instances.

$$Precision = \frac{TruePositive}{(TruePositive + FalseNegative)} \tag{11}$$

F1-Score is the harmonic mean of precision and recall. It provides a balance between precision and recall, making it a valuable metric for models aiming to achieve a balance between true positives and false positives.

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{12}$$

5.3 Data Augmentation and Ensemble

In order to explore the ways to improve classification performance, we explored ensembling techniques along with augmentation techniques.

5.3.1 Data Augmentation Technique: Data augmentation is the process of artificially generating new training data from existing data [2]. In our data augmentation technique, we started with a dataset of 11,457 Hindi tweets, consisting of 970 hate speech tweets and 10,487 non-hate speech tweets, mirroring the real-world distribution. To enhance the diversity of our dataset, Google Translate API was employed to perform multilingual augmentation. For the hate speech tweets, we translated them into 10 different languages, including English, Nepali, Russian, Spanish, Japanese, Urdu, Tamil, German, Greek, and French. Subsequently, it was back-translated these tweets into Hindi. This process resulted in the generation of 9,700 (970×10) augmented instances in addition to the original 970 hate speech tweets totaling 10,659 hate speech tweets in the augmented dataset. Each original hate speech instance was expanded into 10 augmented instances in different languages.

Similarly, for the individual, organization, and community target classes, we had 400, 415, and 154 tweets, respectively. To address class imbalance, oversampling was done to the community target class to reach a total of 462 tweets. To augment these target class tweets, similar procedure was executed. We translated them into English and French and back-translated them, resulting in a final augmented set of 462 tweets for the community class. The data augmentation technique helped in significantly expanding our dataset while ensuring that the semantic meaning remained similar across the augmented instances. Table 8 shows the dataset statistics before and after augmentation.

Table 8. Dataset Statistics before and after augmentation for our dataset.

Task	Labels	Total Number of Tweets	
		Before Augmentation	After Augmentation
Hate Speech Detection	Hate	970	10659
	Non-Hate	10487	10487
Identification of Targets	Individual	400	400
	Organisation	415	415
	Community	154	462

5.3.2 Ensemble Techniques: The performance of predictive models can be significantly improved through ensembling techniques, which involve combining the predictions of multiple models [21, 57]. Ensembling methods can broadly be categorized into two types: soft ensemble and hard ensemble. In a soft ensemble, the individual models' outputs are aggregated using weighted averages, where the weights are determined by the confidence

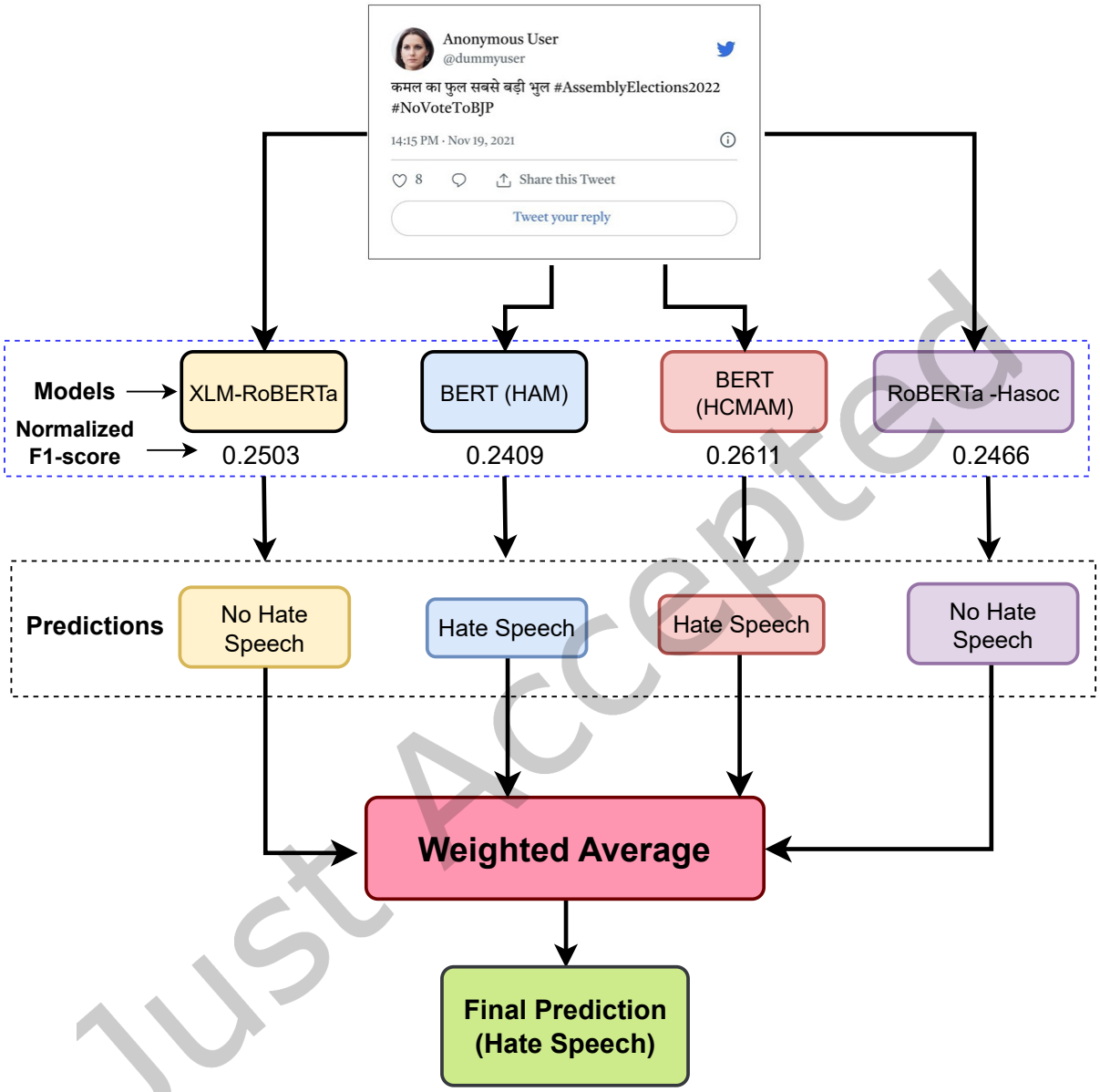

Fig. 10. Ensemble technique used in our experimentation. We leverage the strengths of four BERT-based models for enhanced performances.

or performance of each model. This approach is particularly useful when dealing with models that provide probability scores or confidence levels as part of their predictions. It aims to exploit the varying strengths of individual models to improve overall prediction accuracy. On the other hand, in a hard ensemble, the final prediction is

made by majority voting or taking the mode of the predictions from the individual models. This approach is beneficial when dealing with models that provide discrete class labels as output. It helps in combining the diverse perspectives of individual models and often results in robust predictions.

In our implementation, we opted for a hard ensemble of BERT-based models, including XLM-RoBERTa, BERT (HAM), BERT (HCMAM), and RoBERTa-Hasoc. We refer to this ensemble as *Hard Ensemble of BERT-Models (HEB)*. By aggregating the predictions of these models through majority voting, HEB provides a collective decision that leverages the diversity and strengths of each individual model. This ensemble approach aims to enhance the overall performance of our hate speech detection and target classification tasks, resulting in more accurate and robust predictions. Additionally, our decision to employ transformer-based models in this ensemble was based on our observations of lesser accuracy with traditional machine learning and deep learning algorithms as shown in Table 9 and Table 10. Recognizing the powerful capabilities of transformer models in capturing complex patterns and semantics within textual data, we strategically chose to utilize them in our ensemble to maximize predictive accuracy and robustness.

Each tweet produced four distinct output prediction, each originating from a distinct transformer-based model. These results from the four models were meticulously organized and saved in a CSV format. To analyze individual data or tweets, we utilized the mode operation to identify the most frequent category of prediction. This mode category was then stored in a CSV format, with the data or tweet ID remaining unchanged. To make it easier, when identifying Hate Speech, we represented the result as binary (0 for non-hate and 1 for hate). As for categorizing the Targets of Hate Speech, individual was assigned as 0, community as 1, and organization as 3.

We encountered a situation where all four models reached a **deadlock/tie**. For instance, among these four models, two predicted '1' for Hate, while the remaining two predicted '0' for Non-Hate. In this situation, two categories have the same weightage, so we define this situation as a deadlock/tie. Fig. 11 shows the different situations during the ensemble process.

Tweet ID	XLM-RoBERTa	BERT (HAM)	BERT (HCMAM)	RoBERTa-Hasoc	New Prediction (Ensemble)
15××××23	0	0	0	1	**0 (Non-Hate Speech)**
47××××69	1	0	1	1	**1 (Hate Speech)**
31××××54	1	1	0	0	**Deadlock/Tie**

Fig. 11. Different situations during the ensemble process (Binary classification)

Solution for deadlock/tie: To resolve this deadlock, we implemented a specific approach. For each tweet where a deadlock situation occurred, normalized F1-score (validation) was calculated for each algorithm and stored these scores in a list. The normalized F1-scores (validation) were used to determine the weightage of the algorithm pairs involved in the tie. We compared the final two weighted values, and the output of the selected pair with the highest value became the output of the hard ensemble.

Let's illustrate this with an example: In a situation where two algorithms predict a tweet as hateful and the other two as non-hateful, see Fig. 12, we store the normalized F1-scores of these algorithms for the validation set. In this example, we can see that XLM-RoBERTa and BERT (HAM) label the tweet as hateful, while BERT (HCMAM) and RoBERTa-Hasoc label it as non-hateful. Now, calculated the sum of normalized F1-scores for the pair XLM-RoBERTa and BERT (HAM) and the sum of normalized F1-scores for the pair BERT (HCMAM) and RoBERTa-Hasoc. After this calculation, we find that the latter pair has a higher sum than the former. Consequently, we assigned the label of the tweet as non-hateful (label predicted by the latter pair) based on this analysis. This process ensures that even in situations where all four models reach a tie, we can make a definitive prediction, enhancing the reliability and accuracy of our ensemble-based predictions. Fig. 12 visually represents the methodology employed to resolve deadlock situations.

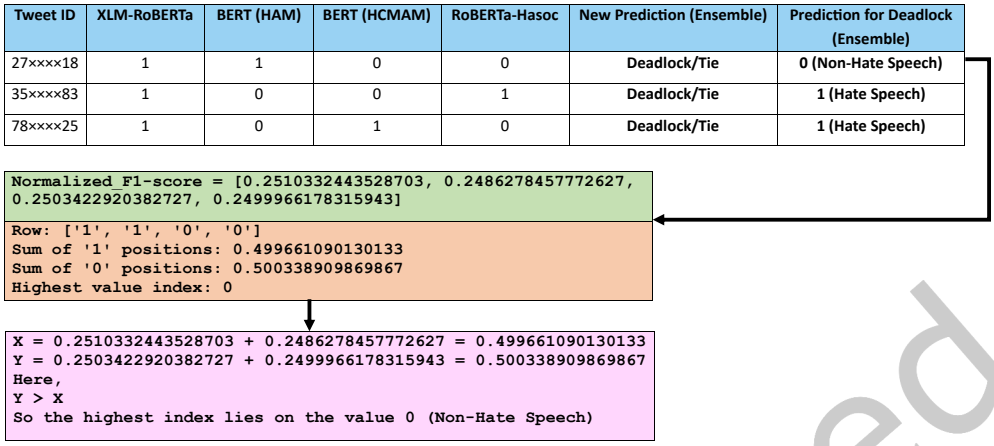

Fig. 12. Approach for the deadlock situation (Binary classification)

6 RESULTS AND ANALYSIS

In this section, we describe our classification results along with the interpretation. We also present the analysis of misclassified results.

6.1 Results for Hate Speech Detection

In the task of hate speech classification for both original and augmented data as shown in Table 9, results of our original dataset show that transformer-based models have a lot of potential. The baseline model 'RoBERTa-Hasoc' on the original dataset has achieved an F1-score of 0.725 and an accuracy of 0.923, which is the most accurate of the baseline models. Importantly, our HEB model performed better than 'RoBERTa-Hasoc,' increasing the F1-score by 1.10% on the original dataset. It can be observed that employing the augmented dataset resulted in a notable boost in performance and achieved very prominent results. The baseline model 'RoBERTa-Hasoc' on the augmented dataset achieved 31.72% more relative f1-score as compared to the original dataset. Also with an accuracy of 0.959 and an F1-score of 0.959, the 'Hard Ensemble of BERTs (HEB)' model turned out to be the best-performing model. Compared to the top-performing baseline model, this indicates a significant improvement in accuracy and F1-score particularly considering that in the task of hate speech detection, small increments are also paramount in curbing hate speech in social media.

6.2 Results for Target Detection

In the task of target detection for both original and augmented data as shown in Table 10, we observed that, with an F1-score of 0.652, BERT (HCMAM) outperformed all other baseline models for the original dataset. The "Hard Ensemble of BERTs (HEB)" demonstrated the best accuracy and F1-score, measuring 0.765 and 0.726, respectively, and performed rather well in the original data. Our research indicates that employing an augmented dataset leads to a considerable improvement in outcomes in both baseline models and 'Hard Ensemble of BERTs (HEB)'. The model with the greatest accuracy (0.868) and F1-score (0.866) among all the models we evaluated was the 'Hard Ensemble of BERTs (HEB)'. Additionally, the model "Hard Ensemble of BERTs (HEB)" outperformed all other models with a significantly lower MMAE score of 0.278 and 0.151 in both original and augmented datasets respectively.

Table 9. Performance on detection of hate speech by different algorithms

Model	Original Dataset			Augmented Dataset		
	Acc↑	MMAE↓	F1-Score↑	Acc↑	MMAE↓	F1-Score↑
Naive Bayes	0.766	0.410	0.589	0.887	0.107	0.892
Decision Tree	0.849	0.438	0.568	0.906	0.093	0.906
SVM	0.916	0.303	0.533	0.924	0.075	0.924
Logistic Regression	0.820	0.410	0.604	0.922	0.077	0.922
Conv-BiLSTM	0.873	0.391	0.616	0.943	0.061	0.938
BiLSTM	0.901	0.367	0.585	0.943	0.062	0.937
RoBERTa-Hasoc	0.923	0.230	0.725	0.955	0.044	0.955
BERT (HAM)	0.913	0.272	0.706	0.958	0.041	0.958
XLM-Roberta	0.906	0.322	0.603	0.954	0.045	0.954
BERT (HCMAM)	0.905	0.307	0.659	0.949	0.050	0.949
Hard Ensemble of BERTs (HEB)	**0.935**	**0.214**	**0.733**	**0.959**	**0.040**	**0.959**

Table 10. Results of Targets with different algorithms in boht Original and Augmented Dataset

Model	Original Dataset			Augmented Dataset		
	Acc↑	MMAE↓	F1-Score↑	Acc↑	MMAE↓	F1-Score↑
Naive Bayes	0.560	0.737	0.452	0.635	0.445	0.629
Decision Tree	0.522	0.700	0.446	0.604	0.494	0.595
SVM	0.608	0.292	0.447	0.640	0.434	0.634
Logistic Regression	0.570	0.719	0.466	0.619	0.473	0.609
Conv-BiLSTM	0.540	0.637	0.409	0.601	0.434	0.588
BiLSTM	0.487	0.785	0.433	0.645	0.429	0.637
RoBERTa-Hasoc	0.708	0.392	0.601	0.801	0.259	0.811
BERT (HAM)	0.697	0.424	0.614	0.800	0.218	0.804
XLM-Roberta	0.748	0.357	0.645	0.812	0.208	0.814
BERT (HCMAM)	0.760	0.287	0.652	0.712	0.351	0.695
Hard Ensemble of BERTs (HEB)	**0.765**	**0.278**	**0.726**	**0.868**	**0.151**	**0.866**

6.3 Analysis and Discussion on Performance

In our results, we observed that the machine learning models and non-transformer based models perform low as compared to the transformer-based models. Transformer-based models, particularly RoBERTa and BERT variants, demonstrate substantial potential, outperforming the non-transformer models in both original and augmented data. It is in line with past observations on similar tasks [28, 62]. The high performance by a hard ensemble of BERTs (HEB) shows that the ensemble techniques are effective in classification by leveraging the

strengths of individual BERT models. Furthermore, the augmentation of datasets significantly enhances model performance, highlighting the importance of diverse and representative data in improving generalization and better performance.

6.4 Misclassification and the need for Explainability

It was exciting to observe in Fig. 13 that some of the non-hateful tweets were mistakenly labeled as hateful tweets. Fig. 13 shows some of the misclassification examples by hard ensemble of BERTs. The investigation of these misclassifications is crucial to building strong models. Fig. 13 illustrates that some terms with a negative meaning or words that can confuse the model were mostly misclassified. For example, Fig. 13 (left) shows that the user mentions that some leader of सपा *(Samajwadi Party)* passed hateful comments on राहुल गांधी *(Rahul Gandhi)*. Since there is mention of the party name of the person demeaning *Rahul Gandhi* as well as the party *Rahul Gandhi* is associated with, the model likely flagged the tweet as the organization target instead of the individual target. Similarly, as seen in the Fig. 13 (middle), the tweet is non-hateful but the presence of hateful or negative words such as "भ्रष्ट (Corrupt)" and " चोर (Thief)" are included in the misclassified tweet. This might have affected the model's ability to correctly discriminate between the classes. Similarly, Fig. 13 (right) shows that there is mention of parties (organizations) as well as politician names. This affected the model's ability to distinguish between individual and organization targets. Examining the proper reasons behind misclassification requires more research on the model's explainability. Finding the most important terms and cues that models use for categorization would enable us to dive more into the decision-making process of the model and improve its accuracy. Achieving transparency in model decisions not only helps to build trust but also enables easier refinement of models for effectively combating various forms of hate speech. Techniques such as feature importance analysis [33], attention mechanisms [40], rule extraction [1], counterfactual explanations [59], and human-in-the-loop approaches [42, 66] offer avenues for understanding and improving model interpretability for the analysis of hate speech. Furthermore, there is also a need for more research to develop models capable of accurately distinguishing between various tweeting patterns. Overall, the results highlight the need for more investigation and development in the area of Indian language multi-aspect tweet classification, with the use of sophisticated transformer-based models being one potentially useful tactic.

Fig. 13. Misclassification examples by the Hard Ensemble of BERTs (HEB) model

6.5 Limitations

In this study, we offer a sizable dataset for a thorough examination of Hindi tweets related to a political debate against the establishment. We suggest establishing baselines to evaluate the discourse around hatred and its

targets. The targets of hate speech were also identified. However, our efforts are limited in several ways. First off, our dataset may not always be representative of the forms of hate speech and may not express the same attitude in different situations because it only covers tweets from a certain time period related to the Indian election. Furthermore, the sole tweets in our sample come from a single tweeting network. Second, targets that are more complicated or specialized may not be identified by our target annotation approach as it is limited to generic categories (Individuals, Organizations, and Communities). Moreover, because annotation is a subjective process, different annotators may have differing opinions about a tweet. However, a high inter-annotator agreement shows that our annotations are mostly valid. Thirdly, the characteristics on which the baselines we provide are based are restricted, thus it is possible that alternative features or architectural arrangements will provide higher performance. It is also critical to emphasize that the use of technology to identify targets and analyze speech from several angles may raise ethical concerns, such as potential bias. When developing and implementing such solutions, moral considerations should be made and addressed.

6.6 Broader Impact

The research presented in this paper, particularly focusing on hate speech analysis in low-resource languages such as Hindi through the *CHUNAV* dataset, aligns closely with key principles of the United Nations Sustainable Development Goals (SDGs) [31, 47]. By prioritizing actions to understand and address hate speech in a marginalized linguistic context, this research contributes to SDG 10: Reduced Inequalities, aiming to uplift marginalized communities and reduce discrimination. The public availability of the *CHUNAV* dataset not only fosters further research but also aligns with SDG 9: Industry, Innovation, and Infrastructure, as it supports the development of efficient and sustainable tools for hate speech detection, contributing to technological innovation. Moreover, the prevention of hate speech, as advocated in this research, resonates with SDG 3: Good Health and Well-Being, as hate speech can have severe consequences on individuals' mental health. Detecting and mitigating hate in political events aligns with SDG 16: Peace, Justice, and Strong Institutions, as it contributes to building inclusive societies that are free from discrimination and promote peace. By focusing on hate speech in the distinctive socio-political context of Indian Assembly Elections, this research further supports the achievement of SDG 16 by working towards fostering a more just and inclusive political discourse.

In summary, the research not only addresses the specific challenges of hate speech in low-resource languages but also actively contributes to broader global goals of poverty reduction, inequality reduction, technological innovation, mental health improvement, and the promotion of peace and justice, as outlined in the United Nations Sustainable Development Goals. The *CHUNAV* dataset and its analysis strive to create a more inclusive and respectful online space, particularly within the distinctive realm of Indian Assembly Elections.

7 CONCLUSION

In conclusion, this research makes a significant contribution to the understanding and mitigation of hate speech in the context of Indian Assembly Elections by introducing the *CHUNAV* dataset. Through the meticulous processes of topic modeling, manual annotations for the presence of hate speech, and identification of targets (individuals, organizations, and communities), we have laid the groundwork for comprehensive analysis and benchmarking. Our benchmark evaluation encompasses a diverse set of machine learning, deep learning, and transformer-based algorithms, shedding light on their efficacy in hate speech detection and target identification.

The findings of this study underscore the need for further research in two critical dimensions. Firstly, the call for increased focus on explainability in hate speech detection models becomes evident. Understanding why certain classifications occur is crucial for refining models and ensuring their responsible deployment in real-world applications. Secondly, while our study provides valuable insights into hate speech targeting individuals, organizations, and communities, a more nuanced exploration of hate speech targets is warranted. Unraveling the

intricacies of hate speech towards specific groups within these broad categories will deepen our understanding and inform more targeted interventions.

This work addresses the challenges of hate speech in a socio-politically sensitive context and in a language with limited resources for NLP. The dataset and analysis techniques developed in this study contribute to the broader understanding of hate speech dynamics in Indian social media, offering tools for better monitoring and intervention. This work is crucial for advancing hate speech detection technology and promoting a more respectful online environment in the context of Indian politics.

Furthermore, this study is confined to a single platform, and the digital landscape is diverse. Future research should expand its scope to encompass multiple platforms, considering the unique dynamics each one presents. A cross-platform study could unveil variations in hate speech dynamics and strengthen the generalizability of our findings. In summary, our work opens avenues for continued exploration, encouraging researchers to delve deeper into the intricacies of hate speech, refine detection methodologies, and extend analyses to broader digital contexts. By doing so, we can collectively work towards creating digital spaces that are not only technologically advanced but also socially responsible and culturally sensitive, fostering a more inclusive online environment.

8 ACKNOWLEDGEMENTS

MK is supported by UKRI NERC grant NE/X000192/12.

REFERENCES

[1] Jesse Ables, Nathaniel Childers, William Anderson, Sudip Mittal, Shahram Rahimi, Ioana Banicescu, and Maria Seale. 2024. Eclectic Rule Extraction for Explainability of Deep Neural Network based Intrusion Detection Systems. *arXiv preprint arXiv:2401.10207* (2024).

[2] Surabhi Adhikari, Surendrabikram Thapa, Usman Naseem, Hai Ya Lu, Gnana Bharathy, and Mukesh Prasad. 2023. Explainable hybrid word representations for sentiment analysis of financial news. *Neural Networks* 164 (2023), 115–123.

[3] Ika Alfina, Rio Mulia, Mohamad Ivan Fanany, and Yudo Ekanata. 2017. Hate speech detection in the Indonesian language: A dataset and preliminary study. In *2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*. 233–238. https://doi.org/10.1109/ICACSIS.2017.8355039

[4] Jacob Amedie. 2015. The impact of social media on society. (2015).

[5] Muhammad Umair Arshad, Raza Ali, Mirza Omer Beg, and Waseem Shahzad. 2023. UHated: hate speech detection in Urdu language using transfer learning. *Language Resources and Evaluation* (2023), 1–20.

[6] Femi Emmanuel Ayo, Olusegun Folorunso, Friday Thomas Ibharalu, and Idowu Ademola Osinuga. 2020. Machine learning techniques for hate speech classification of twitter data: State-of-the-art, future challenges and research directions. *Computer Science Review* 38 (2020), 100311.

[7] Aashish Bhandari, Siddhant B Shah, Surendrabikram Thapa, Usman Naseem, and Mehwish Nasim. 2023. CrisisHateMM: Multimodal Analysis of Directed and Undirected Hate Speech in Text-Embedded Images From Russia-Ukraine Conflict. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1993–2002.

[8] Mohit Bhardwaj, Md Shad Akhtar, Asif Ekbal, Amitava Das, and Tanmoy Chakraborty. 2020. Hostility Detection Dataset in Hindi. arXiv:2011.03588 [cs.CL]

[9] Aritz Bilbao-Jayo and Aitor Almeida. 2021. Improving political discourse analysis on twitter with context analysis. *IEEE Access* 9 (2021), 104846–104863.

[10] Michał Bilewicz and Wiktor Soral. 2020. Hate speech epidemic. The dynamic effects of derogatory language on intergroup relations and political radicalization. *Political Psychology* 41 (2020), 3–33.

[11] Carlos Arcila Calderón, Gonzalo de la Vega, and David Blanco Herrero. 2020. Topic modeling and characterization of hate speech against immigrants on Twitter around the emergence of a far-right party in Spain. *Social Sciences* 9, 11 (2020), 188.

[12] Mudit Chaudhary, Chandni Saxena, and Helen Meng. 2021. Countering online hate speech: An nlp perspective. *arXiv preprint arXiv:2109.02941* (2021).

[13] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 8440–8451.

[14] Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. 2011. Political Polarization on Twitter. *Fifth International AAAI Conference on Weblogs and Social Media*.

[15] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning* 20, 3 (1995), 273–297.

[16] Renu Dalal, Manju Khari, John Petearson Anzola, and Vicente García-Díaz. 2021. Proliferation of opportunistic routing: A systematic review. *IEEE Access* 10 (2021), 5855–5883.
[17] Renu Dalal, Manju Khari, and M Hernandez. 2021. Persuasive simulation of optimized protocol for OppNet. *Dynamic Systems and Applications* 30, 5 (2021), 865–900.
[18] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. *Proceedings of the International AAAI Conference on Web and Social Media* 11 (03 2017). https://doi.org/10.1609/icwsm.v11i1.14955
[19] Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate Speech Dataset from a White Supremacy Forum. arXiv:1809.04444 [cs.CL]
[20] Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate speech detection with comment embeddings. In *Proceedings of the 24th international conference on world wide web*. 29–30.
[21] Asif Ekbal and Sriparna Saha. 2013. Simulated annealing based classifier ensemble techniques: Application to part of speech tagging. *Information Fusion* 14, 3 (2013), 288–300.
[22] Atefeh Farzindar, Diana Inkpen, and Graeme Hirst. 2015. *Natural language processing for social media*. Springer.
[23] Vaishali U Gongane, Mousami V Munot, and Alwin D Anuse. 2022. Detection and moderation of detrimental content on social media platforms: current status and future directions. *Social Network Analysis and Mining* 12, 1 (2022), 129.
[24] Vikram Gupta, Sumegh Roychowdhury, Mithun Das, Somnath Banerjee, Punyajoy Saha, Binny Mathew, Animesh Mukherjee, et al. 2022. Multilingual Abusive Comment Detection at Scale for Indic Languages. *Advances in Neural Information Processing Systems* 35 (2022), 26176–26191.
[25] Tim Highfield. 2017. *Social media and everyday politics*. John Wiley & Sons.
[26] Christine Hine. 2013. *The internet*. Oxford University Press, USA.
[27] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
[28] Farhan Ahmad Jafri, Mohammad Aman Siddiqui, Surendrabikram Thapa, Kritesh Rauniyar, Usman Naseem, and Imran Razzak. 2023. Uncovering Political Hate Speech During Indian Election Campaign: A New Low-Resource Dataset and Baselines. arXiv:2306.14764 [cs.CL]
[29] R Tallal Javed, Muhammad Usama, Waleed Iqbal, Junaid Qadir, Gareth Tyson, Ignacio Castro, and Kiran Garimella. 2022. A deep dive into COVID-19-related messages on WhatsApp in Pakistan. *Social Network Analysis and Mining* 12 (2022), 1–16.
[30] Hamed Jelodar, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li, and Liang Zhao. 2019. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications* 78 (2019), 15169–15211.
[31] Pia Katila, Carol J Pierce Colfer, Wil De Jong, Glenn Galloway, Pablo Pacheco, and Georg Winkel. 2019. *Sustainable development goals*. Cambridge University Press.
[32] Yash Khemchandani, Sarvesh Mehtani, Vaidehi Patil, Abhijeet Awasthi, Partha Talukdar, and Sunita Sarawagi. 2021. Exploiting language relatedness for low web-resource language model adaptation: An Indic languages study. *arXiv preprint arXiv:2106.03958* (2021).
[33] Kwang Hyeon Kim, Woo-Jin Choi, and Moon-Jun Sohn. 2022. Feature Importance Analysis for Postural Deformity Detection System Using Explainable Predictive Modeling Technique. *Applied Sciences* 12, 2 (2022), 925.
[34] Krishanu Maity, Gokulapriyan Balaji, and Sriparna Saha. 2023. Towards Analyzing the Efficacy of Multi-task Learning in Hate Speech Detection. In *International Conference on Neural Information Processing*. Springer, 317–328.
[35] Krishanu Maity, Shaubhik Bhattacharya, Salisa Phosit, Sawarod Kongsamlit, Sriparna Saha, and Kitsuchart Pasupa. 2023. Ex-ThaiHate: A Generative Multi-task Framework for Sentiment and Emotion Aware Hate Speech Detection with Explanation in Thai. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 139–156.
[36] Krishanu Maity, Shaubhik Bhattacharya, Sriparna Saha, and Manjeevan Seera. 2023. A deep learning framework for the detection of Malay hate speech. *IEEE Access* (2023).
[37] Krishanu Maity, Nilabja Ghosh, Raghav Jain, Sriparna Saha, and Pushpak Bhattacharyya. 2019. StereoHate: Towards identifying Stereotypical Bias and Target group in Hate Speech Detection. *Natural Language Engineering* 1 (2019), 00.
[38] Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2022. HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. arXiv:2012.10289 [cs.CL]
[39] David Mimno, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the 2011 conference on empirical methods in natural language processing*. 262–272.
[40] Akash Kumar Mohankumar, Preksha Nema, Sharan Narasimhan, Mitesh M Khapra, Balaji Vasan Srinivasan, and Balaraman Ravindran. 2020. Towards Transparent and Explainable Attention Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 4206–4216.
[41] Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. 2022. ETHOS: a multi-label hate speech detection dataset. *Complex & Intelligent Systems* 8, 6 (2022), 4663–4678.
[42] Eduardo Mosqueira-Rey, Elena Hernández-Pereira, David Alonso-Ríos, José Bobes-Bascarán, and Ángel Fernández-Leal. 2023. Human-in-the-loop machine learning: A state of the art. *Artificial Intelligence Review* 56, 4 (2023), 3005–3054.

[43] Zewdie Mossie and Jenq-Haur Wang. 2018. Social network hate speech detection for Amharic language. *Computer Science & Information Technology* (2018), 41–55.

[44] Natarajan Narasimhamurthy. 2014. Use and Rise of Social Media as Election Campaign Medium in India. https://api.semanticscholar.org/CorpusID:24787289

[45] Edi Surya Negara, Dendi Triadi, and Ria Andryani. 2019. Topic Modelling Twitter Data with Latent Dirichlet Allocation Method. In *2019 International Conference on Electrical Engineering and Computer Science (ICECOS)*. 386–390. https://doi.org/10.1109/ICECOS47637.2019.8984523

[46] Taberez Ahmed Neyazi. 2020. Digital propaganda, political bots and polarized politics in India. *Asian Journal of Communication* 30, 1 (2020), 39–57.

[47] Derek Osborn, Amy Cutter, and Farooq Ullah. 2015. Universal sustainable development goals. *Understanding the transformational challenge for developed countries* 2, 1 (2015), 1–25.

[48] Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and Multi-Aspect Hate Speech Analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 4675–4684. https://doi.org/10.18653/v1/D19-1474

[49] Anil Singh Parihar, Surendrabikram Thapa, and Sushruti Mishra. 2021. Hate speech detection using natural language processing: Applications and challenges. In *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)*. IEEE, 1302–1308.

[50] Rahul, Vasu Gupta, Vibhu Sehra, and Yashaswi Raj Vardhan. 2021. Hindi-English Code Mixed Hate Speech Detection using Character Level Embeddings. In *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*. 1112–1118. https://doi.org/10.1109/ICCMC51019.2021.9418261

[51] Kritesh Rauniyar, Sweta Poudel, Shuvam Shiwakoti, Surendrabikram Thapa, Junaid Rashid, Jungeun Kim, Muhammad Imran, and Usman Naseem. 2023. Multi-Aspect Annotation and Analysis of Nepali Tweets on Anti-Establishment Election Discourse. *IEEE Access* (2023).

[52] Biswarup Ray and Avishek Garain. 2020. JU at HASOC 2020: Deep Learning with RoBERTa and Random Forest for Hate Speech and Offensive Content Identification in Indo-European Languages.. In *FIRE (Working Notes)*. 168–174.

[53] Kelly Reynolds, April Kontostathis, and Lynne Edwards. 2011. Using Machine Learning to Detect Cyberbullying. In *2011 10th International Conference on Machine Learning and Applications and Workshops*, Vol. 2. 241–244. https://doi.org/10.1109/ICMLA.2011.152

[54] Irina Rish et al. 2001. An empirical study of the naive Bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, Vol. 3. 41–46.

[55] Lior Rokach and Oded Maimon. 2005. Decision trees. *Data mining and knowledge discovery handbook* (2005), 165–192.

[56] Niloofar Safi Samghabadi, Parth Patwa, Srinivas PYKL, Prerana Mukherjee, Amitava Das, and Thamar Solorio. 2020. Aggression and Misogyny Detection using BERT: A Multi-Task Approach. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*. European Language Resources Association (ELRA), Marseille, France, 126–131. https://aclanthology.org/2020.trac-1.20

[57] Sriparna Saha and Asif Ekbal. 2013. Combining multiple classifiers using vote based classifier ensemble technique for named entity recognition. *Data & Knowledge Engineering* 85 (2013), 15–39.

[58] Chiranjibi Sitaula, Anish Basnet, Ashish Mainali, Tej Bahadur Shahi, et al. 2021. Deep learning-based methods for sentiment analysis on Nepali COVID-19-related tweets. *Computational Intelligence and Neuroscience* 2021 (2021).

[59] Ilia Stepin, Jose M Alonso, Alejandro Catala, and Martín Pereira-Fariña. 2021. A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. *IEEE Access* 9 (2021), 11974–12001.

[60] Jill C. Stoltzfus. 2011. Logistic Regression: A Brief Primer. *Academic Emergency Medicine* 18, 10 (2011), 1099–1104. https://doi.org/10.1111/j.1553-2712.2011.01185.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1553-2712.2011.01185.x

[61] Surendrabikram Thapa, Farhan Ahmad Jafri, Ali Hürriyetoğlu, Francielle Vargas, Roy Ka-Wei Lee, and Usman Naseem. 2023. Multimodal Hate Speech Event Detection - Shared Task 4, CASE 2023. In *Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*.

[62] Surendrabikram Thapa, Rauniyar Kritesh, Shiwakoti Shuvam, Sweta Poudel, Usman Naseem, and Mehwish Nasim. 2023. NEHATE: Large-Scale Annotated Data Shedding Light on Hate Speech in Nepali Local Election Discourse. In *26th European Conference on Artificial Intelligence*.

[63] Surendrabikram Thapa, Aditya Shah, Farhan Ahmad Jafri, Usman Naseem, and Imran Razzak. 2022. A multi-modal dataset for hate speech detection on social media: Case-study of russia-ukraine conflict. In *CASE 2022-5th Workshop on Challenges and Applications of Automated Extraction of Socio-Political Events from Text, Proceedings of the Workshop*. Association for Computational Linguistics.

[64] Zhou Tong and Haiyi Zhang. 2016. A text mining research based on LDA topic modelling. In *International conference on computer science, engineering and information technology*. 201–210.

[65] Javier Torregrosa, Sergio D'Antonio-Maceiras, Guillermo Villar-Rodríguez, Amir Hussain, Erik Cambria, and David Camacho. 2023. A mixed approach for aggressive political discourse analysis on Twitter. *Cognitive computation* 15, 2 (2023), 440–465.

[66] Konstantinos Tsiakas and Dave Murray-Rust. 2022. Using human-in-the-loop and explainable AI to envisage new future work practices. In *Proceedings of the 15th International Conference on PErvasive Technologies Related to Assistive Environments*. 588–594.

[67] Abhishek Velankar, Hrushikesh Patil, Amol Gore, Shubham Salunke, and Raviraj Joshi. 2021. Hate and Offensive Speech Detection in Hindi and Marathi. arXiv:2110.12200 [cs.CL]

[68] Zeerak Waseem and Dirk Hovy. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*. Association for Computational Linguistics, San Diego, California, 88–93. https://doi.org/10.18653/v1/N16-2013

[69] Yakobus Keenan Wiciaputra, Julio Christian Young, and Andre Rusli. 2021. Bilingual Text Classification in English and Indonesian via Transfer Learning using XLM-RoBERTa. *International Journal of Advances in Soft Computing & Its Applications* 13, 3 (2021).

[70] Wenjie Yin and Arkaitz Zubiaga. 2021. Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Computer Science* 7 (2021), e598.

[71] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the Type and Target of Offensive Posts in Social Media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 1415–1420. https://doi.org/10.18653/v1/N19-1144

Received 19 November 2023; revised 9 March 2024; accepted 9 May 2024