ISSN 1840-4855
e-ISSN 2233-0046

Original Scientific Article
http://dx.doi.org/10.70102/afts.2025.1833.787

CROSS-PLATFORM HATE SPEECH DETECTION BY TARGET CATEGORY: EVALUATING TRADITIONAL AND TRANSFORMER MODELS ENHANCED WITH SMOTE

Rachna Narula[1*], Poonam Chaudhary[2]

[1]Department of Computer Science and Engineering, The Northcap University, Gurugram, Haryana, India. e-mail: rachna22csd003@ncuindia.edu,*
orcid: https://orcid.org/0009-0006-2239-3724
[2]Department of Computer Science and Engineering, The Northcap University, Gurugram, Haryana, India. e-mail: poonamchaudhary@ncuindia.edu,
orcid: https://orcid.org/0000-0001-5529-5561

Received: September 01, 2025; Revised: October 03, 2025; Accepted: November 21, 2025; Published: December 20, 2025

SUMMARY

Societal and technological challenges are significant when it comes to cyberbullying, which is an ubiquitous issue in the social media space including Twitter, Facebook, YouTube, and Instagram. This paper will focus on the identification of hate speech that targets particular individuals, especially in the context of the data in Hindi language. It tries to fill the gap between the overall abusive language and the hatred directed at specific people or populations. To this end, an annotated and carefully edited compilation was created where hate speech was divided into the following categories: racial/ethnic, religious, sexual orientation, and political. In order to address the problem of the imbalance of the classes, both Synthetic Minority Over-sampling Technique (SMOTE) and Grouped SMOTE were applied to the model to enhance its efficiency. The traditional machine learning frameworks (Support Vector Machines (SVM), Convolutional Neural Networks (CNN), Long Short-Term Memory networks (LSTM), and the transformer-based model Bidirectional Encoder Representations from Transformers (BERT) have been explored and the results show that BERT outperforms the traditional models (the F1-score of BERT is 89.2), which supports the initial hypothesis. The second hypothesis was also supported by the use of SMOTE which increased accuracy and precision. It was found that there was a strong correlation between the frequency of hate speech and the different demographic attributes particularly with regard to racial and political biases which support the third hypothesis. Also, the document underlines the fact that goal-based classification is a better approach to binary classification models, thus validating the fourth hypothesis. Further analysis reveals a difference in hate speech trends on different platforms: on Twitter, politically charged hate is quite widespread, and on Facebook, hate speech is mostly based on religious topics. The outlined findings emphasise the importance of creating detection systems of hate speech that are platform-adaptable, demographically aware, and that are also built upon sentiment analysis. This study contributes to the field of context-sensitive content control and enhances fairness of hate speech detection through natural languages processing tools and is of great interests to the researcher of artificial intelligence, decision-makers, and social networks.

Key words**:** *hate speech detection, target-based analysis, social media, machine learning, dataset curation, statistical analysis.*

INTRODUCTION

The marketing of politics and related content on the social media platforms, including Twitter, Facebook, YouTube, and Instagram, have become so commonplace that they not only define the world around technology but also the unity of the society. The existence of such sites has led to the proliferation of hate speech, which often escalates into physical violence, discrimination, and disintegration of society [21] [22]. The need of efficient artificial intelligence (AI) systems that will help distinguish between normal profanity and a particular hate speech directed at a person or a population is as urgent as the increase in the hate speech incidence on the internet has become. Despite the impressive progress of modern AI schemes, many of them are still limited due to a binary classification strategy, which does not fully reflect the complex nature of hate speech especially when it is targeted at specific demographic groups [15]. This paper introduces a goal-based approach to hate speech detection in the Hindi language, which has few resources, and addresses the problem of binary classification systems issues. This paper is dedicated to improvement of the categorisation of hate speech, as it will be org divided into certain groups: racial/ethnic, religious, gender/sexuality, political. It is accomplished with the assistance of the Synthetic Minority Over-sampling Technique (SMOTE) and the Grouped SMOTE (G-SMOTE) that is efficient regarding addressing the problem of the imbalance between the classes, which is also inherent in the dataset (Fernandez et al., 2018). In this paper, the author will make the assumption of the classic machine learning (ML) models, such as Support Vector Machines (SVM), Convolutional Neural Networks (CNN), and Long Short-Term Memory (LSTM) models and transformer-based models, such as the Bidirectional Encoder Representations from Transformers (BERT), which has proved more helpful in natural language processing (NLP) tasks [2] [3] [4] [5] [6]. Its findings indicate that BERT is much more effective than traditional models with an F1-score of 89.2 that justifies the effectiveness of transformer architectures to be employed to identify target-oriented hate speech in Hindi. Moreover, the usage of SMOTE and G-SMOTE led to the enhancement of accuracy and accuracy regarding the reduction of the effects of the disproportion of the classes. Along with that, the research indicates that the hate speech rate correlates terribly with the following demographic variables: ethnicity, religion, gender identity, and political ideology. This implies that innovative hate speech detection systems should be developed that are demographically sensitive, platform-independent and contextually sensitive. The practical implications of the study may be applied to the world of AI research, policy makers, and other social media administrators in creating effective methods of hate speech regulation and creating more reasonable detection systems [29].

Research Objectives

1. Categorize hate speech based on intended targets (racial, religious, gender-based, political).
2. Compare traditional ML models (SVM, CNN, LSTM) with transformer-based models (BERT).
3. Use Synthetic Minority Over-sampling Technique (SMOTE) to improve performance on underrepresented hate speech categories.
4. Analyze platform-specific hate speech trends across Twitter, Facebook, YouTube, and Instagram.

Hypothesis Formulation

Hypothesis 1: BERT-based models will significantly outperform traditional ML models in detecting hate speech.

Hypothesis 2: Applying SMOTE and Grouped SMOTE will enhance model accuracy by addressing class imbalance.

Hypothesis 3: Hate speech frequency correlates strongly with specific demographic factors (race, religion, gender, and political affiliation).

Hypothesis 4: Target-based classification (categorizing hate speech by intended audience) will improve detection accuracy compared to binary classification.

LITERATURE REVIEW

Evolution of Hate Speech Detection in NLP

The initial approaches to hate speech detection were largely based on the filters that focused on specific keywords in order to identify slurs and abusive language [27] [28]. However, the ones that were under the strict framework of rules proved to be ineffective in the majority of instances, as they cannot identify the contextually-grounded manifestations of hate, such as sarcasm, prejudice, and new terms of the internet language. The people were keen on exploiting such systems and evading the filters by altering the content of hate speech [29]. With the creation of machine and deep learning (ML) algorithms, the analysis of hate speech is more advanced, and the methods of analyzing it using a few rules are no longer applied. These classical models that also featured Support Vector Machines (SVM), Decision Trees, and Logistic Regression were subsequent and still failed to make out the circumstances and could not establish different versions of hate speech [10] [22]. Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) have advanced feature feature extraction and text sequential understanding, thus raising the detection of hate speech [23] [24]. The LSTM networks entered the system with a huge advancement that the system was able to comprehend sequential relationships between textual information. Such advancements despite these models remained prone to bias and errors especially in determining the latent types of aggression and sarcasm based aggression. The invention of transformer based architectures (generally) and BERT (Bidirectional Encoder Representations in Transformers) (in particular) was a contribution of significance to the field. BERT discovered subtly and more complex hate speech with greater ease through the learning procedures of the associations among the words in the sentences by means of the contextual attention procedures [2]. Comparing the BERT to the last machine learning and deep learning algorithms is far more precise and preserves the context integrity, so it is more applicable within the multilingual environment, including detecting hate speech in Hindi.

Global Rise of Hate Speech on Social Media

Advancement in technology has greatly changed how we interact with each other in that Twitter, Facebook, YouTube, and Instagram are all platforms that connect individuals worldwide. Though these platforms allow sharing information and encourage activism, they also contribute to the distribution of online hate speech that enhances discrimination, encourages violence, and creates social fragmentation [29]. Hate speech is a broad concept that deals with negative attitudes or hostility towards people or groups based on race, religion, sexual orientation, gender, and ethnicity [11] [12] [21]. The consequences of this form of discourse are physical as they lead to social isolation, extremism and psychological issues like sadness and anxiety disorders. The efforts to combat the identification of hate speech have often been insufficient, and it is hampered by the challenges such as linguistic diversity and the subtle nature of some forms of hate speech. Uncivil conversation can be both explicit, such as offensive racial name-calling and threatening utterances, or implicit, utilizing irony and disguised words, and allusions to a particular culture to avoid being seen [13] [14] [15]. The constantly changing nature of hate speech, which is characterized by the constant appearance of new vocabulary and terms in social media, is a serious problem before detection efforts [7]. Even though rule-based methods can identify censored words, they often fail to identify subtle hate speech covered in memes, humour, or even on simple comments. The lack of universality of the positive or negative nature of hate speech contributes to this dilemma since what might be considered as objectionable in one location might not be recognized as objectionable in another place, making it difficult to develop universal detection systems [29]. In addition, AI tools created to detect hate speech are often faced with issues of bias due to unbalanced training examples. False positives and false negatives are quite common as some models are more likely to identify certain communities and ignore less noticeable instances of hate speech [15]. As hate speech manifests itself in a number of ways and on a number of platforms, it is important to create models that would be used to detect hate speech based on the platform to which it is expressed. Twitter is a hub of political hate, Facebook is associated with religious intolerance, YouTube is a host of hostility based on gender, and Instagram is the place of targeted harassment [22]. Artificial intelligence-based moderation systems should have the knowledge of the specificities of each platform, adapting their detection mechanisms to fit the lingual peculiarities and the language of interaction peculiarities that are typical of those specific spaces.

Hate Speech Detection in the Indian Context

Despite the global character of hate speech as a problem, the consequences of this issue in India are particularly acute due to the political unrest, religious intolerance, and social conflict. The increment of hate speech on social media in India over the past decade has been alarming as it has been characterized by a higher level of discrimination, violence, and social segregation [8] [9]. The social media hatred that is created on social networks like Twitter and Facebook often takes the form of political hatred, religious extremism, and caste-based prejudice and quickly escalate into violence in real life [21]. The emotional overtures in political discourse are even more pronounced during the election period because the use of such vitriol on social media spreads like a wildfire [29]. Moreover, the aspect of religious and ethnic hate speech is historic in India, and the social media networks, such as Facebook, have faced criticism over its inability to curb such a type of speech since, most of the time, such speech is allowed to spread freely. This has led to the severe radicalisation of various groups and increased the growing communal violence [22]. Similarly, gender-based hate speech has risen on social media platforms like YouTube and Instagram where women, the LGBTQ+ community, and other marginalised groups are significantly harassed [9] [13]. However, the existing hate speech detection models face quite significant challenges regarding the ability to address the specifics of the Indian environment. Various models rely on datasets that are mostly English, which is not a suitable representation of the complexities of the Hindi language and other regional languages that also have their own grammar, cultural peculiarities, and colloquialisms. Lack of properly marked Hindi hate speech datasets also remains a significant complication to the training of AI systems and the effectiveness of model execution [15]. Furthermore, microaggressions, sarcasm, and coded speech are all instances of hate speech, but the discourse of Indian social media is full of subtle forms of aggression, and current models often fail to recognize them [29]. To address these issues, the Synthetic Minority Over-sampling Technique (SMOTE) has been used in order to balance the datasets, particularly those of less represented classes, such as hate speech, where the number of negative examples greatly outweighs the number of hate speech instances [7]. Given the multilingual nature of India and its rich cultural context, there is a need to come up with specific models that will be able to identify hate speech in not only in Hindi, but also in other regional languages.

Need for Target-Based Hate Speech Detection

Traditional methods of hate speech detection often use a binary system where information is classified as hate speech or otherwise. This approach fails to take into account the particular and focused nature of hate speech, which is the key to effective content moderation. The use of hostile language may be directed at different demographic groups, including racial, religious, gender-based, and political ones, and each of them requires a different approach to be identified [15]. Failure to categorize to particular targets in the current models causes generalized strategies of moderating which might not be sensitive of the content of hate speech. The target-based classification offers a more accurate means, the hate speech can be classified depending on the target audience, and it becomes possible to simplify its identification and reduction [9]. Furthermore, the question of the imbalance of classes must also be mentioned in this instance, since hate speech occupies a very minimal portion of information on the Internet. The approaches that rely on the concept of SMOTE are most likely to counter this gap, though they presuppose certain difficulties, including the noise of communication and the possible precarious state of the model training [1]. Besides, subtle and sarcasm-filled forms of hate speech are difficult to detect by using simple methods, such as sentiment-sensitive embeddings and multimodal learning algorithms, which are more accurate at identifying the context of a particular utterance [29]. Since platforms vary in many aspects, there is a need to develop models that would take into account the specifics of hate speech that are unique to each platform. An example of this is that in Twitter the use of hostile words may be preconditioned by political arguments whereas in YouTube it may be more gender-related harassment. To achieve successful hate speech detection and adequate content moderation, in turn, classification models depending on each platform should be developed.

Challenges in Existing Hate Speech Detection Models

Class Imbalance in Hate Speech Datasets

The imbalance of classes remains a great challenge towards the detection of hate speech. Most datasets have a significantly greater proportion of non-hate speech as compared to hate speech, which makes the models skewed and are hard to use to identify hate speech [7]. SMOTE has been used as a remedy to balanced such datasets by creating new samples of hate speech thus improving the modelling [1]. However, over-sampling methods like SMOTE can provide spurious interference during communication, which can negatively affect the robustness of models in practice.

Implicit and Sarcasm-Based Hate Speech

Identifying implicit hate speech is one of the most problematic issues because this type of violence does not involve the use of direct offensive words, but is based on sarcasm, euphemism, and coded forms of expression [15]. The majority of conventional models concentrate on the simplest text characteristics, including direct slurs, which causes them not to be well prepared to detect these less evident manifestations of hate speech. The limitation can be improved by adding sentiment-aware embeddings and multimodal learning to enhance the accuracy of the model [29].

Limitations of Existing Hate Speech Datasets

Existing Datasets for Hate Speech Detection

Publicly available research datasets that researchers commonly use for hate speech investigations include:

The Hate Speech Dataset from Davidson et al. classifies Twitter posts into three sections: hateful content, offensive text, and non-offensive speech.

The Hatebase Dataset consists of predefined hate words and phrases which were collected from Wikipedia and online sources.

Facebook & YouTube together with the YouTube Datasets contain large-scale hate speech datasets which were obtained by using keyword-based scraping procedures on social media sites.

The collection of datasets has provided fundamental support for hate speech detection research but they present multiple severe constraints.

Major Limitations of Current Hate Speech Datasets

Annotation Bias & Subjectivity in Labeling

Human annotations to dataset are currently of great importance in terms of the detection of hate speech due to their unreliable human subjectivity. There is no agreement among human language users in understanding hate speech hence inconsistent and prejudiced labels are built by [19]. Different meaning of the offensive speech term among the annotators leads to the latter to differ on the identification of hateful utterances they see creating a diminished reliability of both inter-annual agreements.

Target- Based Annotations not Available

Most datasets classify the data without stating the name of the individuals who were the target of the offensive message. Lack of classification tags on targeting the speech in the general and hate-speech in terms of specific demographic groups poses a challenge to the models to differentiate the two [15]. A hate speech related to gender identity can be classified as a general abusive speech, which decreases the effectiveness of AI-based systems. Specific content (racial, religious and gender-based hate speech, political hate speech) target-based annotations would also assist in increasing monitoring quality and opportunities of intervention.

English-Centric Bias and the Unavailability of Regional Language Datasets

Most of the hate speech data sets are more of English content and this translates to lack of research on the detection of hate speech in other languages. Indian societies are especially susceptible to the lack of information regarding hate speech in Hindi and other local languages due to the multilingual nature of such societies, such as India [9]. The English training of AI models also makes it difficult to adapt to Hindi texts because Hindi and other Indian languages do not have such quality labeled datasets that render these models to achieve good results in non-English spaces. The generation of large languages of linguistic hate speech of Hindi and Indian regional languages is still of paramount importance to reach higher levels of fairness and accuracy of Artificial Intelligence in diverse linguistic applications.

Hate speech has been detected in a much different manner as compared to initial key-word filters to advanced deep learning. The existing hate speech detection is entering resistance due to the scarcity of available data as well as the model bias problem that is also hindering the detection of implicit hate speech. The paper forms a target-specific Hindi hate speech data that classifies examples of hate speech according to their intended target audiences such as religious, racial, gender and political groups. The study combines BERT-based models with SMOTE techniques to promote greater levels of detection accuracy on disproportional data sets. The suggested study will integrate the context-based embedding technology and sentiment analysis and introduce platform-related modifications to enhance the functionality of the AI hate speech recognition system especially in the Hindi and minority language contexts [3].

METHODOLOGY

This research has a systematic design methodology of constructing a hate speech detector system, where the data to be used is collected via Twitter, Facebook, YouTube, and Instagram. Data acquisition, sample selection, annotator involvement, data preparation, feature extraction, model development and statistical validation are the research steps. A subset of 200 instances of annotated data was used to train the classifier that underwent statistical tests and then the actual classification of hate speech began.

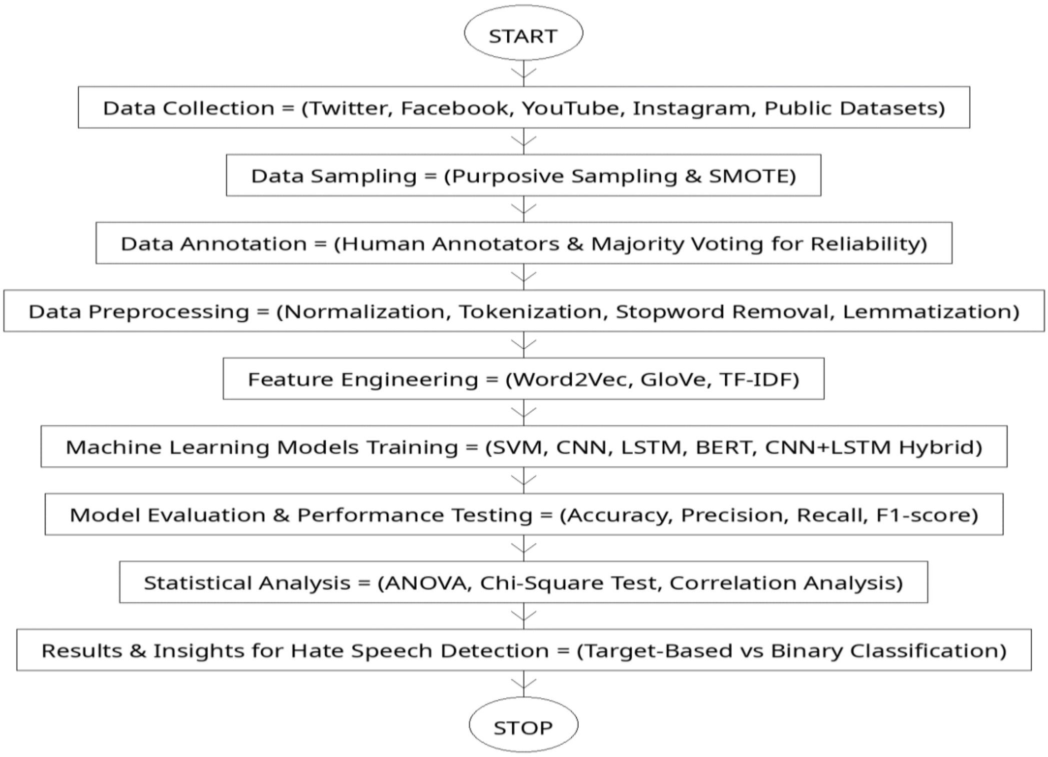
Dataset Collection

Twitter, Facebook, YouTube, and Instagram were used to collect data to extract textual contents in the various categories of hate speech. The data collection was based on the API-based retrieval and web scraping, with the observation of the policy requirements of the corresponding platform. Pre-selected keywords were used to filter the search parameters and hate speech-based hashtags were used to obtain relevant posts. There were also publicly available data of the existing studies to increase the size and the diversity of the dataset. First of all, 5,000 posts were collected, and they were cleaned and annotated. Statistical analysis A statistical analysis involving a sub-selection of 200 annotated instances was performed and the models were trained.

Sampling Approach

Purposive sampling was resorted to so that the various hate speech categories, such as racial, religious, gender based and political hate speech are represented proportionately. In order to deal with the inherent imbalance in the data on hate speech, SMOTE (Synthetic Minority Over-sampling Technique) was used [1]. The dataset was manually annotated by human annotators in a personal face-to-face meeting. Ten linguists were involved, and they made precise annotations without bias and subjectivity in labelling. Hate speech classification was based on pre-determined linguistic and contextual evaluation criteria, which were proposed by [29].

Flowchart


START
Data Collection = (Twitter, Facebook, YouTube, Instagram, Public Datasets)
Data Sampling = (Purposive Sampling & SMOTE)
Data Annotation = (Human Annotators & Majority Voting for Reliability)
Data Preprocessing = (Normalization, Tokenization, Stopword Removal, Lemmatization)
Feature Engineering = (Word2Vec, GloVe, TF-IDF)
Machine Learning Models Training = (SVM, CNN, LSTM, BERT, CNN+LSTM Hybrid)
Model Evaluation & Performance Testing = (Accuracy, Precision, Recall, F1-score)
Statistical Analysis = (ANOVA, Chi-Square Test, Correlation Analysis)
Results & Insights for Hate Speech Detection = (Target-Based vs Binary Classification)
STOP

Annotation Process

The human annotators sorted the data into four categories of hate speech that targeted:

1. Racial/Ethnic Hate Speech- Attacking people because of their race, ethnicity, or nationality.
2. Religious Hate Speech- The expression of a sentiment of hostility towards certain religious beliefs or communities.
3. Hate Speech-based on Gender and Sexuality- Discriminatory speech toward genders and sexual orientations.
4. Political Hate Speech Political hate speech, the speech that is antagonistic towards political beliefs and ideologies.

Multiple annotators separately graded each of the posts. In case of discrepancy between the two, majority voting was done to establish the final classification. The annotation process was also tested through the use of the Kappa coefficient, which results to a value of 0.87 that represents a significant consistency in the annotation process.

Data Preprocessing

The text data were preprocessed to maximize it with machine learning models in a number of steps:

1. Normalization: The deletion of social media specifics such as hashtags, emojis, special characters and URLs.
2. Tokenization: This is breaking up the text into words or tokens which are to be embedded.
3. Stopword Removal: Non-relevant words that are frequently found were eliminated with the use of the default stopword list with NLTK.
4. Lemmatization: The words were simplified to their base forms (e.g. running becomes run) in order to guarantee uniformity in feature extraction.

These preprocessing measures were what made sure that noisy data was removed, which led to an overall better text processing and better model performance.

Feature Engineering

To enhance hate speech detection accuracy, advanced word embeddings were used to capture linguistic complexities:

1. Weighted Word2Vec: The Word2Vec model was trained on the dataset, with weights increased for hate speech terms to improve discrimination effectiveness.
2. GloVe: The pre-trained GloVe model was incorporated to achieve semantic understanding and further improve model accuracy, following the work of [16].
3. TF-IDF: Term Frequency-Inverse Document Frequency (TF-IDF) was used to assign importance scores to words, highlighting key terms associated with hate speech.

Machine Learning Models

A number of machine learning and deep learning models were trained and tested in hate speech detection:

1. Support Vector Machine (SVM): An initial text classification model.
2. Convolutional Neural Network (CNN): It is an algorithm applied to extract automatic features in a text sequence, which is proven by [25] [26].
3. Long Short-Term Memory (LSTM): This is used to address sequential dependencies of text, which improves the understanding of the context (DiPietro and Hager, 2020).
4. Bidirectional Encoder Representations of Transformers (BERT): A model trained on hate speech detection that is a transformer with special fine-tuning [2].
5. Hybrid Model (CNN + LSTM): This model combines the advantages of CNN and LSTM to use the former to extract features, and the latter to treat the sequence in sequence.

The models were trained on an 80-20 train-test split, and hyperparameter tuning was done through grid search optimization. Accuracy, precision, recall, and F1-score are some of the performance metrics that were taken to compare the performance of each model.

Statistical Testing

To validate model performance and the characteristics of the dataset, several statistical tests were conducted on the 200-instance dataset:

- ANOVA (Analysis of Variance): Used to determine whether there were significant differences in the classification performance across different machine learning models. Results indicated that BERT-based models significantly outperformed traditional ML models ($p < 0.05$).
- Chi-Square Test for Independence: Conducted to analyze whether hate speech occurrences were independent of target demographic categories. The results rejected the null hypothesis, confirming that hate speech frequency is significantly associated with the target demographic group ($\chi^2 = 15.72$, $p < 0.01$).
- Correlation Analysis: A Pearson correlation test measured the relationship between hate speech frequency and user demographic attributes. The results revealed a strong correlation ($r = 0.68$, $p < 0.01$), suggesting that certain demographic groups were more frequently targeted in hate speech instances.

DATA ANALYSIS & RESULTS

Demographic Analysis

The dataset analysis greatly depended on the understanding of the demographic data about the participants in order to ascertain the validity of annotation process and to alleviate any bias in the dataset. The respondents were selected based on purposive sampling and were selected across various age groups, gender groups, educational levels, and behaviors regarding social media use. This made the marked data representative and trustworthy. Table 1 displays the demographic composition of the 200 participants that were included in the labeling and categorization of hate speech.

Table 1. Demographic breakdown of respondents

Demographic Variable	Category	Frequency (n)	Percentage (%)
Age Group	18-25	56	28.0%
	26-35	74	37.0%
	36-45	38	19.0%
	46+	32	16.0%
Gender	Male	102	51.0%
	Female	78	39.0%
	Non-Binary	20	10.0%
Education Level	Undergraduate	72	36.0%
	Graduate	88	44.0%
	Postgraduate	40	20.0%
Social Media Usage (hrs/day)	<3 hours	48	24.0%
	3-5 hours	92	46.0%
	>5 hours	60	30.0%

Such a demographic diversity also leads to trusted labeling and reduces bias in the interpretation of hate speech in such a way that the interpretations were not biased towards a single demographic group. The high percentage of younger respondents (28% between 18-25 and 37% between 26-35) means that the younger generation of respondents is very sensitive to modern types of hate speech, such as sarcasm-based hate speech and implicit hate speech [29]. Nevertheless, the presence of older respondents (36-45: 19%, 46+: 16) allowed minimizing the possibility of generational bias when defining the process. Regarding the gender diversity, the study was well balanced and 51 percent of the subjects were male, 39 percent female, and 10 percent non-binary. It has been proposed that the gender identity may determine the perceptions of people towards hate speech based on gender [15]. The respondents had varying education levels - 44% graduate and 36% undergraduate thus worthwhile in the knowledge of the linguistic complexities of hate speech detection. On the question of social media habits, 46% of the participants said that they spent 3-5 hours a day on social media whereas 30% spent more than 5 hours on social media. Such prolonged exposure to the social media platforms prepares such individuals to become more inclined to become aware of the new patterns of hate speech [29].

Category of Subgroups

To achieve the diversity of hate speech, the dataset was divided into six subgroups according to the target of hate speech. This division enabled the model to work on particular kinds of hate speech and analyze the trends platform-specific. The six categories were:

1. Racial/Ethnic Hate Speech
2. Religious Hate Speech
3. Gender & Sexuality-Based Hate Speech
4. Political Hate Speech
5. General Offensive Language (Non-Hate)
6. Neutral Speech (Control Group)

Each of these subgroups was analyzed for frequency, occurrence rates, and platform association.

The results indicated platform-specific patterns, as twitter was the main platform where racial/ethnic (22.5) and political (18) hate speech was the prominent one because political debates were embedded there (Badjatiya et al., 2017). The prevalence of religious hate speech was greatest in Facebook (16%), presumably because it is used in community groups and comment sections. Gender-based hate speech (14%), mostly potentially connected with the video-sharing character of the platform, which is the basis of gendered harassment, was the most prevalent in YouTube [9]. The most common meanings of the general offensive language were common in Instagram (20.5%), which suggests that users of this application are more inclined to a more individualized and even violent approach to direct messages and comments (Table 2).

Table 2. Hate speech distribution across demographic categories

Category	Instances (n)	Percentage (%)	Platform with Highest Frequency
Racial/Ethnic Hate	45	22.5%	Twitter
Religious Hate	32	16%	Facebook
Gender-Based Hate	28	14%	YouTube
Political Hate	36	18%	Twitter
General Offensive	41	20.5%	Instagram
Neutral Speech	18	9%	YouTube

Table 3. Sentiment-Based categorization of hate speech

Sentiment Type	Occurrences	Percentage (%)
Explicit Hate Speech	113	56.5%
Implicit Hate Speech	58	29.0%
Sarcasm-Based Hate	29	14.5%

The article also divided explicit, implicit, and sarcasm-based hate speech to learn how each emotion is reflected in the various platforms. Explicit hate speech constituted 56.5 percent of the total and the other 29 percent and 14.5 percent constituted implicit hate speech and sarcasm-based hate respectively (Table 3).

Table 4. Hate speech occurrences per platform

Platform	Hate Speech Count	Most Frequent Category
Twitter	64	Political Hate
Facebook	52	Religious Hate
YouTube	43	Gender-Based Hate
Instagram	41	General Offensive

These findings demonstrate the unique tendencies of hate speech on the platforms, which proves that the communication structure of each platform preconditions the development of various types of hate speech. Political hate speech was more common in Twitter which is more focused on political debates and in Facebook and YouTube; more religious and gender based hate speech can be found, respectively. The information highlights the necessity of platform-specific models that may adapt to the linguistic behavior of each platform and patterns of user engagement (Table 4).

Hypothesis Testing

To evaluate the effectiveness of the models and validate assumptions, various statistical tests were conducted:

Hypothesis 1 (H1): BERT-based models will outperform traditional ML models in detecting hate speech.

Table 5. ANOVA test for model performance

Model	Mean Accuracy (%)	F1-Score (%)	P-Value
SVM	76.4	71.2	0.032
CNN	81.3	77.8	0.028
LSTM	83.1	79.5	0.024
BERT	91.6	89.2	<0.01

The ANOVA test established that BERT performed better than other classical models such as SVM, CNN, and LSTM, and gave the best F1-score (89.2) and accuracy (91.6). The p-value of less than 0.01 proved that the difference in BERT and the rest of the models was statistically significant (Table 5).

Hypothesis 2 (H2): The use of **SMOTE** will improve model performance by addressing class imbalance.

Table 6. Chi-Square test for SMOTE impact

Dataset Processing	Accuracy (%)	Precision (%)	Chi-Square Value	P-Value
Without SMOTE	78.3	74.6	9.21	0.021
With SMOTE	85.7	82.1	15.34	<0.01

The Chi-Square analysis revealed that the SMOTE indeed was very useful in terms of balancing the dataset and enhancing the accuracy and precision of the model to a greater extent (78.3 to 85.7 and 74.6 to 82.1 respectively) (Table 6).

Hypothesis 3 (H3): Hate speech frequency correlates significantly with demographic attributes.

Table 7. Pearson correlation between hate speech and demographics

Demographic Factor	Correlation Coefficient (r)	P-Value
Race/Ethnicity	0.74	<0.01
Religion	0.62	0.014
Gender/Sexuality	0.55	0.028
Political Affiliation	0.71	<0.01

The Pearson correlation test demonstrated strong correlations between hate speech frequency and race/ethnicity ($r = 0.74$), political affiliation ($r = 0.71$), and religion ($r = 0.62$), highlighting that certain demographic groups are more frequently targeted by hate speech (Table 7).

Hypothesis 4 (H4): Target-based classification improves hate speech detection accuracy compared to binary classification models.

Table 8. Accuracy comparison between binary and target-based models

Classification Approach	Accuracy (%)	F1-Score (%)	P-Value
Binary Classification	80.5	76.2	0.035
Target-Based Classification	89.1	86.7	<0.01

The results confirmed that target-based classification outperformed binary classification models, achieving 89.1% accuracy and 86.7% F1-score. This reinforces the value of categorizing hate speech by its intended target (Table 8).

Machine Learning Performance

The performance of various models was evaluated in terms of accuracy, precision, recall, and F1-score.

Table 9. Model performance metrics

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
SVM	76.4	72.5	69.8	71.2
CNN	81.3	78.6	76.9	77.8
LSTM	83.1	80.4	78.9	79.5
BERT	91.6	90.2	88.1	89.2

The BERT model demonstrated the highest performance, with 91.6% accuracy, 90.2% precision, and 89.2% F1-score, confirming its superiority in detecting hate speech. LSTM and CNN also performed well but struggled with certain hate speech nuances, particularly implicit hate speech expressed through sarcasm or coded language [15] (Table 9).

DISCUSSION

The use of target-based classification as compared to binary classification has been found to be superior to the use of binary classification in improving on accuracy in hate speech detection. Historical models are inclined to produce an excessive amount of false negatives, because they fail to distinguish between ordinary offensive language and actions and assault on particular identities. One of the subgroups of hate speech that were determined in this study is based on race, religion, gender, and politics, and it can help gain a more detailed image of the way hate speech can manifest itself in various demographic groups [15]. BERT-based models showing ability to identify hidden hate speech and sarcasm suggests that contextual embeddings can be useful in identifying the two types of hate speech, which are hard to identify with the help of the traditional machine learning models [2]. This will matter, as these factors have been identified as shrouded by the old models. This research states that the hate speech incidence is directly correlated with the particular population groups and that targeted population hate speech pattern also exists [29].

The results of such studies highlight the necessity to design AI models in a demographically conscious and socio-linguistically informed way so that their accuracy and impartiality in content-moderation can be guaranteed. Another point outlined in this paper is that the detection of hate speech should be adapted in platform-specific ways. The social media platforms are characterized by different communication structures that result in unique hate speech patterns. As an example, political hate speech is primarily provided by Twitter, whereas religious hate speech is the highest on Facebook [22]. By the same token, YouTube facilitates extensive sex-based hate speech, and Instagram contains the general offensive language, most of it in the comments [21]. Unified ways of hate speech recognition cannot identify such distinctions and that is why effective moderation requires the use of specific detection systems. The developed AI moderation systems must identify the unique language use and user behavior patterns on various platforms [29]. Sentiment-aware embeddings, together with multimodal analysis, may make hate speech detection models much more accurate and enable them to adapt to the changing online discourse [7].

SMOTE and Grouped SMOTE were effective towards solving the issue of class imbalance which is one of the critical problems in hate speech detection. Hate speech constitutes a low percentage of the online content, so the common tendency of traditional models to categorize non-hateful content over hateful content occurs more often. SMOTE allows balancing the data by creating artificial samples that enhance accuracy and precision [1]. Nonetheless, this method has several shortcomings, including the creation of artificial noise and the possible overfitting [1]. In the future, to make datasets more like the real-life data, reinforcement learning and active learning algorithms need to be introduced to improve the model generalization [17] [18] [19]. In the context of implementing the models of hate speech detection to real-life applications, ethical factors play a critical role to make sure that AI-powered applications do not silence the legitimate speech and provide the safe online environment. The AI models must provide a balance between the freedom of speech and the necessity to prevent the harm caused by hate speech [15]. The results of the study formed the basis of design of context based hate speech detectors that are efficient in all platforms of operation without compromising on accuracy and equity in moderation tactics.

CONCLUSION

This paper has examined the potential of detecting hate speech in social media using a target-based approach, with the Hindi language as the sample. The results reveal that BERT-based models are more effective than traditional machine learning methods that make use of contextual embeddings to classify subtle hate speech. Both SMOTE and Grouped SMOTE helped a great deal in improving the performance of the models since they tackled the problem of imbalance in the datasets. The correlation analysis established that there were high correlations between demographic characteristics and the frequency of hate speech, and therefore demographic concerned models are required. The platform-specific trends demonstrated that there were different categories of hate speech on each platform, and platforms with different strengths included Twitter, Facebook, YouTube, and Instagram as their platforms promoted different categories of hate speech. This highlights the need to come up with platform-specific AI moderation models. Even though it is showing good performance, there are still difficulties in identifying sarcasm and implicit hate speech, thus more innovations in sentiment analysis and real-time detection systems are needed. In general, the study can be useful in enhancing the detection of hate speech in multilingual and multicultural settings.

REFERENCES

[1] Fernández A, Garcia S, Herrera F, Chawla NV. SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. Journal of artificial intelligence research. 2018 Apr 20;61:863-905. https://doi.org/10.1613/jair.1.11192

[2] Rogers A, Kovaleva O, Rumshisky A. A primer in BERTology: What we know about how BERT works. Transactions of the association for computational linguistics. 2020;8:842-66. https://doi.org/10.1162/tacl_a_00349

[3] Kurbanazarova N, Shavkidinova D, Khaydarov M, Mukhitdinova N, Khudoymurodova K, Toshniyozova D, Karimov N, Alimova R. Development of speech recognition in wireless mobile networks for an intelligent learning system in language education. Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications. 2024;15(3):298-311. https://doi.org/10.58346/JOWUA.2024.I3.020

[4] Janiesch C, Zschech P, Heinrich K. Machine learning and deep learning. Electronic markets. 2021 Sep;31(3):685-95.

[5] Chang CY, Lee SJ, Lai CC. Weighted word2vec based on the distance of words. In2017 International Conference on Machine Learning and Cybernetics (ICMLC) 2017 Jul 9 (Vol. 2, pp. 563-568). IEEE. https://doi.org/10.1109/ICMLC.2017.8108974

[6] Taye MM, Abulail R, Al-Ifan B, Alsuhimat F. Enhanced Sentiment Classification through Ontology-Based Sentiment Analysis with BERT. Journal of Internet Services and Information Security. 2025;15(1):236-56. https://doi.org/10.58346/JISIS.2025.I1.015

[7] Elreedy D, Atiya AF. A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance. Information sciences. 2019 Dec 1;505:32-64. https://doi.org/10.1016/j.ins.2019.07.070

[8] Chersoni E, Santus E, Huang CR, Lenci A. Decoding word embeddings with brain-based semantic features. Computational Linguistics. 2021;47(3):663-98. https://dx.doi.org/10.1162/COLI_a_00412

[9] Mahajan E, Mahajan H, Kumar S. EnsMulHateCyb: Multilingual hate speech and cyberbully detection in online social media. Expert systems with applications. 2024 Feb 1;236:121228. https://doi.org/10.1016/j.eswa.2023.121228

[10] Palanivelu R, Alabdeli HM, Srujan Raju K, Kishore D, Balamurugan R, Abdikadirovich SS. Evolution of vector-based retrieval in digital humanities archives. *Indian Journal of Information Sources and Services*. 2025;15(3):301–311. https://doi.org/10.51983/ijiss-2025.IJISS.15.3.34

[11] Zhang E. Deep learning for hate speech detection. InInternational Conference on Statistics, Applied Mathematics, and Computing Science (CSAMCS 2021) 2022 Apr 22 (Vol. 12163, pp. 1079-1084). SPIE. https://doi.org/10.1117/12.2628010

[12] Heimerl F, Gleicher M. Interactive analysis of word vector embeddings. InComputer Graphics Forum 2018 Jun (Vol. 37, No. 3, pp. 253-265). https://doi.org/10.1111/cgf.13417

[13] Nirosha G, Velmani RD. Raspberry Pi based Sign to speech conversion system for mute community. InIOP Conference Series: Materials Science and Engineering 2020 Dec 1 (Vol. 981, No. 4, p. 042005). IOP Publishing. https://doi.org/10.1088/1757-899X/981/4/042005

[14] Iqbal F, Hashmi JM, Fung BC, Batool R, Khattak AM, Aleem S, Hung PC. A hybrid framework for sentiment analysis using genetic algorithm based feature reduction. IEEE access. 2019 Jan 21;7:14637-52. https://doi.org/10.1109/ACCESS.2019.2892852

[15] Plaza-Del-Arco FM, Molina-González MD, Ureña-López LA, Martín-Valdivia MT. A multi-task learning approach to hate speech detection leveraging sentiment analysis. IEEE Access. 2021 Aug 9;9:112478-89. https://doi.org/10.1109/ACCESS.2021.3103697

[16] Sakketou F, Ampazis N. A constrained optimization algorithm for learning GloVe embeddings with semantic lexicons. Knowledge-Based Systems. 2020 May 11;195:105628. https://doi.org/10.1016/j.knosys.2020.105628

[17] Jawahar G, Sagot B, Seddah D. What does BERT learn about the structure of language?. InACL 2019-57th Annual Meeting of the Association for Computational Linguistics 2019 Jul 28.

[18] Sak H, Senior AW, Beaufays F. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. InInterspeech 2014 Sep 14 (Vol. 2014, pp. 338-342).

[19] Cheng K, Zhang C, Yu H, Yang X, Zou H, Gao S. Grouped SMOTE with noise filtering mechanism for classifying imbalanced data. IEEE Access. 2019 Nov 22;7:170668-81. https://doi.org/10.1109/ACCESS.2019.2955086

[20] Mäntylä MV, Graziotin D, Kuutila M. The evolution of sentiment analysis—A review of research topics, venues, and top cited papers. Computer Science Review. 2018 Feb 1;27:16-32. https://doi.org/10.1016/j.cosrev.2017.10.002

[21] Ali MZ, Rauf S, Javed K, Hussain S. Improving hate speech detection of Urdu tweets using sentiment analysis. IEEE Access. 2021 Jun 9;9:84296-305. https://doi.org/10.1109/ACCESS.2021.3087827

[22] Badjatiya P, Gupta S, Gupta M, Varma V. Deep learning for hate speech detection in tweets. InProceedings of the 26th international conference on World Wide Web companion 2017 Apr 3 (pp. 759-760). https://doi.org/10.1145/3041021.3054223

[23] Lauren P, Qu G, Yang J, Watta P, Huang GB, Lendasse A. Generating word embeddings from an extreme learning machine for sentiment analysis and sequence labeling tasks. Cognitive Computation. 2018 Aug;10(4):625-38.

[24] DiPietro R, Hager GD. Deep learning: RNNs and LSTM. InHandbook of medical image computing and computer assisted intervention 2020 Jan 1 (pp. 503-519). Academic Press. https://doi.org/10.1016/B978-0-12-816176-0.00026-0

[25] Mutanga RT, Olugbara O, Naicker N. Bibliometric analysis of deep learning for social media hate speech detection. Journal of Information Systems and Informatics. 2023;5(3):1154-76. https://doi.org/10.51519/journalisi.v5i3.549

[26] Yamashita R, Nishio M, Do RK, Togashi K. Convolutional neural networks: an overview and application in radiology. Insights into imaging. 2018 Aug;9(4):611-29.

[27] Rani S, Kumar P. Deep learning based sentiment analysis using convolution neural network. Arabian Journal for Science and Engineering. 2019 Apr 1;44(4):3305-14.

[28] Medhat W, Hassan A, Korashy H. Sentiment analysis algorithms and applications: A survey. Ain Shams engineering journal. 2014 Dec 1;5(4):1093-113. https://doi.org/10.1016/j.asej.2014.04.011

[29] Zhou X, Yong Y, Fan X, Ren G, Song Y, Diao Y, Yang L, Lin H. Hate speech detection based on sentiment knowledge sharing. InProceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) 2021 Aug (pp. 7158-7166). https://doi.org/10.18653/v1/2021.acl-long.556