Decoding Hate: KNN, LSTM and BERT in the Pursuit of Safer Online Spaces in Hindi

Mansi Somani*, Radhika Mamidi*
IIIT Hyderabad, India

{somani.girishbhai}@students.iiit.ac.in
{radhika.mamidi}@iiit.ac.in

***Abstract*—Detecting and mitigating abusive language online is a crucial task, particularly in multilingual contexts like India, where harmful content can exacerbate societal divides. This paper presents a comprehensive study on classifying abusive text in Hindi, leveraging various machine learning techniques. We evaluate the performance of three approaches: (1) TF-IDF feature extraction with a KNN classifier, (2) a Long Short-Term Memory (LSTM) neural network, and (3) a BERT-based model. Our experiments are conducted on a dataset of approximately 20,000 Hindi comments, labeled as abusive or non-abusive. Preprocessing steps include stopword removal, punctuation handling, emoji translation, and digit removal. For the TF-IDF+KNN approach, we achieve moderate performance with an overall accuracy of 64%. The LSTM model, with carefully tuned hyperparameters, yields an accuracy of 79.5% and a macro F1-score of 79%. However, the BERT-based model outperforms the other techniques, attaining an impressive validation accuracy of around 90.4% after 5 epochs of training. Our findings demonstrate the effectiveness of transformer-based language models like BERT in capturing semantic nuances for abusive language detection. We discuss the strengths and limitations of each approach, providing insights into their real-world applicability and potential for generalization to other languages.**

I. Introduction

The rise of social media and online platforms has changed how we talk to each other and find information. But along with these changes, there's been an increase in harmful and hurtful content online. This poses a big challenge to keeping the internet a safe and welcoming place for everyone. Abusive language, which includes hurtful words and insults, can deeply affect people's feelings and can even make existing social problems worse, especially in places with many different languages and cultures. India is a good example of this diversity, with its many languages and ways of speaking. Here, it's really important to find and stop hurtful content online to protect people's feelings and keep online conversations respectful. The use of hurtful language online, targeting people based on things like their gender, caste, religion, or ethnicity, can reinforce negative stereotypes and make it harder for communities to come together.

This study delves into the intricate task of classifying abusive text in Hindi, one of the most widely spoken languages in India and a pivotal medium of online discourse. By leveraging state-of-the-art machine learning techniques and natural language processing methodologies, we aim to develop robust models capable of accurately identifying and flagging abusive content. Our approach encompasses a comprehensive evaluation of three distinct strategies, each offering unique advantages and insights.

Firstly, we explore a traditional feature engineering method that employs Term Frequency-Inverse Document Frequency (TF-IDF) vectorization to extract meaningful features from the text data, coupled with a K-Nearest Neighbors (KNN) classifier for prediction. This approach leverages the statistical properties of the text to capture discriminative patterns, providing a baseline for comparison with more sophisticated techniques.

Secondly, we harness the power of deep learning by implementing a Long Short-Term Memory (LSTM) neural network architecture. LSTMs, with their ability to capture long-range dependencies in sequential data, have proven effective in various natural language processing tasks. By carefully tuning the hyperparameters and leveraging the representational capacity of these models, we aim to uncover intricate patterns and semantic nuances that can aid in the accurate classification of abusive content.

Thirdly, we embrace the cutting-edge advancements in transformer-based language models by employing Bidirectional Encoder Representations from Transformers (BERT). BERT has revolutionized the field of natural language processing, demonstrating remarkable performance in a wide range of tasks, including text classification. By leveraging its ability to capture contextual information and the underlying semantics of language, we seek to push the boundaries of abusive language detection.

Through rigorous experimentation on a curated dataset of Hindi comments labeled as abusive or non-abusive, we meticulously evaluate the performance of each approach, quantifying their respective strengths and limitations. Our study not only contributes to the development of effective abusive language detection models but also sheds light on the broader pursuit of fostering safer and more inclusive online spaces, particularly in the linguistically diverse Indian context. By conducting a comprehensive analysis and comparison of these techniques, we aim to uncover the most effective strategies for addressing the challenges posed by abusive content online. Ultimately, our research endeavors to provide insights and recommendations that can inform the development of robust content moderation systems, empowering platforms and communities to proactively identify and mitigate the spread of harmful language, thereby cultivating an environment of respect.

II. Related Work

The detection and mitigation of abusive language online has garnered significant attention from the research community, given the detrimental effects of such content on individuals and society. Numerous studies have explored various techniques and approaches to tackle this challenging problem across different languages and contexts.

Early work in this domain focused on the use of traditional machine learning methods, such as Support Vector Machines (SVMs), Logistic Regression, and Naive Bayes classifiers, in conjunction with feature engineering techniques like bag-of-words and TF-IDF vectorization [1, 2]. While these methods achieved reasonable performance, they often struggled to capture the contextual and semantic nuances inherent in abusive language.

With the advent of deep learning, researchers began exploring the use of neural network architectures for abusive language detection. Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, have been widely employed due to their ability to model sequential data and capture long-range dependencies in text [3, 4]. LSTMs have demonstrated promising results in various languages, including English, German, and Arabic.

More recent studies have leveraged the power of transformer-based language models, such as Bidirectional Encoder Representations from Transformers (BERT) [5] and its variants (e.g., RoBERTa, XLNet). These models, pre-trained on large corpora, have shown remarkable performance in capturing contextual information and understanding the underlying semantics of language, making them well-suited for tasks like abusive language detection [6, 7].

In the Indian context, several researchers have focused on addressing the challenges posed by the linguistic diversity and code-switching prevalent in online communication. Mathur et al. [8] proposed a CNN-based approach for detecting gendered abuse in Hindi, Tamil, and Indian English social media content. Chakravarthi et al. [9] investigated the use of character-level and word-level CNN models for abusive language detection in Hinglish (a combination of Hindi and English). Despite these advancements, the detection of abusive language in low-resource languages like Hindi remains a challenging task due to the scarcity of labeled data and the complexities arising from regional variations, linguistic nuances, and cultural contexts. Several studies have explored techniques such as transfer learning, data augmentation, and semi-supervised learning to mitigate the data scarcity issue [10, 11].

Our study builds upon this existing body of work by conducting a comprehensive evaluation of three distinct approaches – traditional feature engineering with KNN, LSTM neural networks, and BERT-based models – for the task of abusive text detection in Hindi. By comparing and analyzing the performance of these techniques on a curated dataset, we aim to provide insights into their respective strengths and limitations, contributing to the ongoing research efforts in this crucial domain.

III. Dataset

For our study on abusive text detection in Hindi, we leveraged a dataset containing a collection of comments in the Hindi language. These comments were manually labeled as either "abusive" (0) or "non-abusive" (1) based on their content. The dataset comprises approximately 20,000 comments, providing a substantial corpus for training and evaluating our models. The comments were sourced from various online platforms, ensuring a diverse representation of language usage and contexts. The dataset can be downloaded from given **Link**.

The final dataset was stratified to maintain a balanced distribution of abusive and non-abusive comments, enabling effective model training and performance evaluation. Additionally, the dataset was carefully cleaned and preprocessed to handle any potential noise or inconsistencies, such as removing irrelevant content, handling encoding issues, and normalizing text representations.

It is important to note that the dataset used in this study focuses specifically on the Hindi language. While this allows for a comprehensive analysis of abusive language detection in this widely spoken language, it may limit the generalizability of the findings to other languages or multi-lingual contexts. Future work could explore the extension of this research to other languages or the development of language-agnostic models capable of handling code-switching and multilingual data.

By leveraging this carefully curated and labeled dataset, our study aims to provide valuable insights into the effectiveness of various machine learning techniques for detecting abusive content in Hindi social media and online platforms. The dataset's diversity and size ensure robust model training and evaluation, contributing to the ongoing efforts in fostering safer and more inclusive online spaces.

IV. Task Description

The primary objective of this study is to develop effective models for detecting abusive language in Hindi text. We approach this task through three distinct methods, each leveraging different machine learning techniques and architectures. The tasks are defined as follows: Task 1: Abusive Text Classification using TF-IDF and K-Nearest Neighbors (KNN) Task 2: Abusive Text Classification using Long Short-Term Memory (LSTM) Networks Task 3: Abusive Text Classification using Bidirectional Encoder Representations from Transformers (BERT)

A. *Task 1: TF-IDF and K-Nearest Neighbors (KNN)*

1. Approach:

In this task, we employ a traditional feature engineering technique, Term Frequency-Inverse Document Frequency (TF-IDF), to extract meaningful features from the text data. TF-IDF captures the relevance of words in a document by considering their frequency and inverse document frequency.

The preprocessed text data is converted into a matrix of TF-IDF features, which serves as input to a K-Nearest Neighbors (KNN) classifier. KNN is a simple yet effective algorithm that classifies new instances based on their similarity to the nearest neighbors in the training data.

2. Data Preprocessing:

– Created a set of stop words in the Hindi language from various resources.

– Removed stop words from the dataset sentences. Removed punctuation marks from sentences.

– Converted emojis to their text equivalent representations using the emot library.

– Removed digits from the text.

3. Equations:

The TF-IDF value for a word 'w' in a document 'd' is calculated as:

TF-IDF(w, d) = TF(w, d) × IDF(w)

where,

TF(w, d) = (Number of times word 'w' appears in document 'd') / (Total number of words in document 'd')

IDF(w) = log((Total number of documents) / (Number of documents containing word 'w'))

The K-Nearest Neighbors (KNN) algorithm is a non-parametric method that classifies a new instance based on its similarity to the 'k' nearest neighbors in the training data. The similarity is typically measured using distance metrics such as Euclidean distance.

The formula for Euclidean distance between two points P1(x1,y1) and P2(x2,y2) in a two-dimensional space is:

$$\text{Euclidean Distance} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

4. Results:

The TF-IDF + KNN model achieved moderate performance in classifying text comments as abusive or non-abusive. The model's performance metrics are as follows:

TABLE I
PERFORMANCE METRICS

Metric	Class 0 (Abusive)	Class 1 (Non-Abusive)
Precision	0.61	0.74
Recall	0.87	0.40
F1-Score	0.72	0.52
Overall Accuracy	0.64	0.64

B. Task 2: Long Short-Term Memory (LSTM) Networks

1. Approach: In this task, we leveraged the power of deep learning by implementing a Long Short-Term Memory (LSTM) neural network architecture. LSTMs are a type of recurrent neural network capable of capturing long-range dependencies in sequential data, making them well-suited for natural language processing tasks.

2. Data Preprocessing:

– Created a set of stop words in the Hindi language from various resources.

– Removed stop words from the dataset sentences. Removed punctuation marks from sentences.

– Converted emojis to their text equivalent representations using the emot library.

– Removed digits from the text.

TABLE II
HYPERPARAMETERS

Parameter	Values
Vocabulary Size	29941
Embedding Dimensions	300
Hidden Dimension	600
Number of LSTM layers	2
Epochs	10
Learning Rate	0.001
in_features	600
out_features	1
Dropout (p)	0.3

TABLE III
MODEL ARCHITECTURE

Embedding Layer	Embedding(vocab_size,emb_dim)
LSTM Layer	LSTM(ip_sz,hid_sz,layers,batch_first,dropout)
Fully connected Layer	Linear(in_features,out_features)
Dropout Layer	Dropout(p)
Sigmoid Activation Function	Sigmoid()

3. Preparation:

– Vectorized the dataset using word_tokenizer().

– Padded all sentences to a maximum length.

– Split the dataset into an 80:20 ratio for training and validation.

4. Eqautions:

The key equations involved in an LSTM cell are:

$$\text{Forget Gate: } f_t = \sigma(W_f \times [h_{t-1}, x_t] + b_f)$$
$$\text{Input Gate: } i_t = \sigma(W_i \times [h_{t-1}, x_t] + b_i)$$
$$\text{Cell State: } C_t = f_t \cdot C_{t-1} + i_t \cdot \tanh(W_C \times [h_{t-1}, x_t] + b_C)$$
$$\text{Output Gate: } o_t = \sigma(W_o \times [h_{t-1}, x_t] + b_o)$$
$$\text{State: } h_t = o_t \cdot \tanh(C_t)$$

Where,

σ is the sigmoid activation function

$\tanh$ is the hyperbolic tangent activation function

W and b are the weight matrices and bias vectors

x_t is the input at time step t

h_t is the hidden state at time step t

C_t is the cell state at time step t

5. Results:

The LSTM model achieved a relatively high accuracy and F1-score, indicating its effectiveness in classifying text comments as abusive or non-abusive. The model's performance suggests that it has learned meaningful patterns in the text data, allowing it to make accurate predictions.

TABLE IV
RESULTS

Accuracy	79.504%
Macro F1-score	78.984%

C. Task 3: Bidirectional Encoder Representations from Transformers (BERT)

1. Approach:

In this task, we embraced the cutting-edge advancements in transformer-based language models by employing Bidirectional Encoder Representations from Transformers (BERT). BERT has revolutionized the field of natural language processing, demonstrating remarkable performance in various tasks, including text classification.

2. Data Preprocessing:

– Created a set of stop words in the Hindi language from various resources.

– Removed stop words from the dataset sentences. Removed punctuation marks from sentences.

– Converted emojis to their text equivalent representations using the emot library.

– Removed digits from the text.

3. Model Architecture:

a) BertClassifier:

– Defines a custom classifier model using BERT as the base model.

– Takes a dropout rate as an input parameter.

– Initializes the BERT model, a dropout layer, a linear layer for classification, and a ReLU activation function.

– In the forward method, it passes the input through the BERT model and applies dropout, linear transformation, and ReLU activation to obtain the final output.

b) Training Function:

– Trains the BertClassifier model on the provided training data and evaluates it on the validation data.

– Takes the model, training data, validation data, learning rate, batch size, and number of epochs as input parameters.

– Prepares the data loaders for training and validation datasets.

– Iterates over each epoch, calculating the loss and accuracy for each batch of training and validation data.

– Prints the average training loss, training accuracy, validation loss, and validation accuracy for each epoch.

TABLE V
HYPERPARAMETERS

Parameter	Values
Learning Rate	1e-6
Batch Size	20
Epochs	5

4. Equations:

BERT is a transformer-based language model that uses an attention mechanism to capture long-range dependencies in text. The core components of BERT are the encoder and decoder, which rely on the multi-head attention mechanism. The key equations involved in the attention mechanism are:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \cdot V$$

$$\text{Multi-Head Attention}(Q, K, V) = \text{Concat}(head_1, \ldots, head_h) \cdot W^O$$

where:

$Q, K,$ and V are the query, key, and value matrices, respectively

d_k is the dimension of the key vectors

h is the number of attention heads

5. Results:

TABLE VI
ITERATION WISE RESULTS

Epoch	Train Loss	Train Accuracy	Val Loss	Val Accuracy
1	0.547	0.743	0.362	0.824
2	0.332	0.856	0.297	0.868
3	0.268	0.890	0.262	0.887
4	0.231	0.907	0.243	0.898
5	0.207	0.919	0.233	0.904

The BERT-based model outperformed the other techniques, attaining an impressive validation accuracy of around 90.4% after 5 epochs of training. The model achieved an increasing accuracy and decreasing loss over the epochs, indicating that it was learning and improving its predictions. The high validation accuracy demonstrates the model's ability to generalize well to unseen data.

CODE LINK: Github Link

V. Conclusion

This study presents a comprehensive evaluation of three distinct approaches for detecting abusive language in Hindi text, highlighting the remarkable potential of advanced language models and deep learning techniques in tackling this intricate challenge. By leveraging a carefully curated dataset of approximately 20,000 Hindi comments, we conducted rigorous experiments to assess the strengths and limitations of each approach, yielding insights that could shape the future of online content moderation.

The traditional feature engineering approach with TF-IDF and KNN, while providing a baseline performance, struggled to capture the contextual and semantic nuances inherent in abusive language, achieving a moderate overall accuracy of 64%. This underscores the limitations of relying solely on statistical features and highlights the need for more sophisticated methods to handle the complexities of natural language.

The LSTM model, harnessing the power of deep learning and its ability to model sequential data, demonstrated a significant improvement with an accuracy of 79.5% and a macro F1-score of 79%. This performance attests to the effectiveness of neural networks in learning meaningful patterns and representations from text data, paving the way for more robust and nuanced abusive language detection models.

However, the true potential of cutting-edge natural language processing techniques was unveiled by the BERT-based model, which outperformed the other approaches with an impressive validation accuracy of around 90.4% after just 5 epochs of training. This remarkable achievement underscores the transformative impact of transformer-based language models like BERT, which have revolutionized the field by capturing contextual information and the underlying semantics of language with unprecedented accuracy.

The stark contrast in performance between the traditional and deep learning-based approaches highlights the rapid advancements in the field of natural language processing and the crucial role these technologies can play in fostering safer and more inclusive online spaces. By effectively identifying and mitigating the spread of harmful content, such models can contribute to promoting constructive discourse, empowering communities, and cultivating an environment of respect and social cohesion.

Importantly, this study's findings extend beyond the realm of abusive language detection, serving as a testament to the broader potential of deep learning and transformer models in addressing complex natural language processing tasks. As we continue to navigate the digital landscape, these cutting-edge techniques will undoubtedly play a pivotal role in unlocking new frontiers of understanding, enabling more effective communication, and enhancing our ability to harness the power of language in a responsible and ethical manner.

VI. Future Work

While this study provides valuable insights into the detection of abusive language in Hindi, there are several avenues for future exploration and research: **1) Multi-lingual and Code-Switching Scenarios**: Online communication in India often involves code-switching between multiple languages, including Hindi, English, and regional dialects. Extending the research to handle multi-lingual and code-switched data could enhance the applicability and robustness of the models in real-world scenarios.

2) Transfer Learning and Domain Adaptation: Investigating transfer learning techniques and domain adaptation strategies can potentially improve the performance of the models on specific domains or platforms, accounting for variations in language usage and contexts.

3) Explainable AI and Interpretability: While deep learning models like BERT achieve high accuracy, they often lack interpretability. Exploring methods for interpreting and explaining the model's predictions could provide valuable insights into the decision-making process and facilitate better understanding and trust in the system.

4) Incorporating Contextual and Multimodal Information: Abusive language can be influenced by various contextual factors, such as user profiles, conversational history, and multimedia content (e.g., images, videos). Integrating these additional modalities into the models could enhance their ability to accurately identify and mitigate abusive content.

5) Deployment and Real-World Impact Assessment: Collaborating with online platforms and social media companies to deploy and evaluate the effectiveness of the developed models in real-world settings would be crucial. Conducting user studies and assessing the broader societal impact of such systems could inform further improvements and ethical considerations.

While the findings of this study have made significant strides in advancing abusive language detection for Hindi, several avenues for future research remain unexplored. One promising direction lies in the exploration of few-shot and zero-shot learning techniques, which could alleviate the reliance on large labeled datasets and enable the adaptation of models to new domains or languages with minimal data requirements.

By addressing these future research directions, the field of abusive language detection can continue to advance, contributing to the development of more robust, inclusive, and ethical systems for promoting online safety and responsible digital citizenship.

REFERENCES

1 Warner, W. and Hirschberg, J., 2012. Detecting hate speech on the world wide web. In Proceedings of the Second Workshop on Language in Social Media (pp. 19-26).

2 Waseem, Z. and Hovy, D., 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In Proceedings of the NAACL student research workshop (pp. 88-93).

3 Pavlopoulos, J., Malakasiotis, P. and Androutsopoulos, I., 2017. Deep learning for user comment moderation. arXiv preprint arXiv:1705.09993.

4 Badjatiya, P., Gupta, S., Gupta, M. and Varma, V., 2017. Deep learning for hate speech detection in tweets. arXiv preprint arXiv:1706.00188.

5 Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

6 Caselli, T., Basile, V., Mitrović, J., Granitzer, M. and Fell, M., 2021. Hatebert: Retraining bert for abusive language detection in english. arXiv preprint arXiv:2010.12472.

7 Liu, P., Li, W. and Zou, L., 2019. NULI at SemEval-2019 Task 6: Transfer BERT for Offensive Language Detection. arXiv preprint arXiv:1910.09698.

8 Mathur, P., Bhatia, A., Roy, P. and Choudhary, M., 2022. Breaking the Silence: Detecting and Mitigating Gendered Abuse in Hindi, Tamil and Indian English Online Spaces. arXiv preprint arXiv:2201.07749.

9 Chakravarthi, B.R., Muralidaran, V., Priyadharshini, R. and McCrae, J.P., 2020. Corpus creation for sentiment analysis in code-mixed Tamil-English text. arXiv preprint arXiv:2005.00317.

10 Arango, A., Pérez, J. and Nascimiento, B., 2020. Hate speech detection for languages in the latin-american context. arXiv preprint arXiv:2003.07297.

11 Boukkouri, H.E., Aueb, M.E., Abdulrahman, A.A., Laachache, A. and Saad, M., 2021. HASTACK @ DravidianLangTech-EACL2021: Offensive Language Identification in Dravidian Languages using Transformers. arXiv preprint arXiv:2102.11516.

12 Mathur, P., Bhatia, A., Roy, P., & Choudhary, M. (2022). Breaking the Silence: Detecting and Mitigating Gendered Abuse in Hindi, Tamil and Indian English Online Spaces. arXiv preprint arXiv:2201.07749.

13 Chakravarthi, B. R., Muralidaran, V., Priyadharshini, R., & McCrae, J. P. (2020). Corpus creation for sentiment analysis in code-mixed Tamil-English text. arXiv preprint arXiv:2005.00317.

14 Chhablani, G., Jain, S., Pandey, A., & Gupta, M. (2021). One-step and Two-step Classification for Abusive Language Detection on Twitter. arXiv preprint arXiv:2109.09745.