Identification of Objectionable and displeasing Contents in Marathi using BERT-CNN

Darshan Mhapaseka[1], Shubham Kubade[2], Suyash Kubade[3], Kanhaiya Ekawde[4] and Sandip Sawant[5]
[1]*Assistant Professor, Dept of Comp. Engg. S.S.P.M.'s College of Engineering*
[2,3,4,5] *Student, Dept of Comp. Engg. S.S.P.M.'s College of Engineering*

Abstract*— Social media platforms enable users to express their opinions, but this freedom often leads to trolling, rumours, objectionable comments, disrespect, and hate speech, potentially inciting communal riots. To address this, the system is developed a tool for detecting hate and displeasing speech in Marathi. The system focuses on accurately classifying Marathi text into four categories: hateful, offensive, profane, and non-hateful. This project use the Muril BERT transformer for pre-processing and encoding the Marathi text dataset and a CNN model to predict the nature of social media posts or speech. Through extensive experiments with various configurations, it has been found that the combination of Muril BERT and the CNN model achieved an accuracy of 80\%. This approach aims to mitigate the negative impact of harmful speech on social media by providing a reliable detection mechanism

Index Terms*— Hate Speech, Marathi Text Classification, BERT, CNN

I. INTRODUCTION

In an era dominated by digital communication, the proliferation of social media platforms has created both opportunities for connectivity and challenges regarding the sharing of disrespectful comments. Objectionable speech can described as any form of message that denigrates, threatens, or insults individuals or community based on age, ethnicity, gender, gender identity, national origin, and disability a significant threat to social cohesion[1]and individual safety. The Marathi-speaking community, encompassing millions of individuals, faces the same risks of encountering hate speech as any other linguistic group. However, the unique linguistic and cultural context of Marathi necessitates tailored solutions for detecting and mitigating hate speech in this language.

The major contribution of this work is hate speech detection system tailored for the Marathi language. This endeavour involves classifying input text into one of four distinct categories: "hate," "offensive," "profane," or "not hate." This initiative works to create a more secure and welcoming online space for all by effectively identifying and addressing harmful content in Marathi, a language of growing significance in the digital landscape.

Hate Speech (HATE): Speech that promotes violence or prejudicial actions against a particular group or individual based on features such as age, gender, or community.
• Offensive Speech (OFFN): Communication that is likely to cause discomfort, anger, or resentment, often breaching societal norms or standards of decency.
• Profane Speech (PRFN): Vulgar or obscene language that is considered socially unacceptable.
• Not (NOT): A text that is free from displeasing, disrespectful content or profane words and appears normal.

Social networks have become fertile grounds for the propagation of hate speech. Individuals exploit the anonymity provided by online platforms to express discriminatory views, amplify prejudices, and incite hatred. The immediacy and vast reach of social media amplify the impact of hate speech, making it a potent tool for spreading toxicity and influencing public opinion.

The need for Marathi hate speech detection arises from the growing digital presence of the Marathi-speaking population and the potential for hate speech to incite violence, discrimination, and harm within this community. Detecting and addressing hate speech is crucial to protect individuals' safety, maintaining social harmony, and upholding legal and ethical standards in online communication. Automated hate speech detection tools can assist in monitoring and moderating online platforms, preventing the spread of hate speech, and ensuring that digital spaces remain inclusive and respectful for all Marathi speakers, promoting a more positive and constructive online environment.
The model is trained and evaluated by data provided by L3Cube[2]. This dataset comprised of 25000 samples annotated with class labels.

II. PRELATED WORK

Detecting and controlling hate speech is a critical issue, primarily addressed in English text analysis. However, recent efforts focus on extending research to regional languages like Marathi, indicating a broader scope in addressing this problem across diverse linguistic contexts.

[2]3Cube-MahaHate released a dataset in two versions: a 4-class version and a 2-class version. The 4-class dataset has 25,000 samples, divided into 21,500 training samples, 2,000 test samples, and 1,500 validation samples. The labels for this dataset are "hate," "offensive," "profane," and "not." The 2-class dataset has 37,500 samples with labels "hate" and "not." They also tested how well different machine learning paradigms such as CNN, LSTM, BiLSTM and BERT.

[3]They introduced the Marathi Offensive Language Dataset (MOLD), which contains around 2,500 labelled tweets categorized as either offensive or not offensive. This is the first dataset created for identifying offensive language in Marathi. Additionally, they assessed the performance of various traditional machine learning and deep learning paradigm (such as LSTM) trained on the MOLD dataset.

[4]They presented a study on disrespectful text in Marathi and multilingual text using TF-IDF and transformer-based BERT variants. For code-mixed languages, they showed that the BERT model outperforms the Term feature extraction method. For Marathi, they demonstrated that pre-programmed BERT models achieve superior performance ,also more effective at understanding the meaning of sentences, providing superior learning representations. As a result, the transfer learning approach using pre-programmed BERT models is more effective in detecting objectionable speech compared to traditional feature extraction methods. Among the three models tested, MuRIL performed the best.

[5]This paper outlines the experimental work of Team Mind Benders in detecting objectionable content in Marathi on Twits data using Regression and decision trees. After analyzing different feature values and model parameters, the team found that Random Forest outperformed Logistic Regression, achieving an accuracy of 0.77platforms.[6]

Anjum , Rahul Katarya[7] They developed a method to classify the specific emotion or tone of a tweet is objectionable using the Profanity Check Technique (PCT).This method integrates a ReLU activation function with a logistic regression classifier.

III.PROPOSED METHOD

This paper proposes a framework utilizing MURIL BERT along with machine learning algorithms including CNN for text classification. The framework consists of three blocks of processing: A) Data collection, B) Data pre-processing, and C) Text classification. In the data collection phase, relevant datasets are gathered. Subsequently, in the data pre-processing phase utilizes transformer based MURIL BERT language model. Finally, Hate and Offensive text classification is performed using the five layer CNN model.[8]

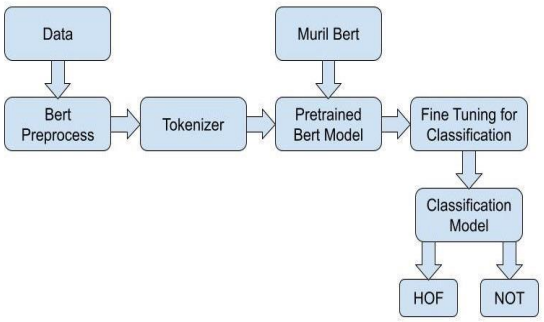

Fig 1- Proposed System

CNN: CNNs in text classification utilize convolution and max-pooling layers. The convolution layer uses filters to extract features from input regions, thereby creating feature vectors. Max pooling selects key features from these vectors, reducing input dimensionality. This process enables effective learning and classification of text data. By focusing on essential information, CNNs achieve accurate classification while efficiently handling text data.

LSTM[9]:Long Short-Term Memory (LSTM) is a kind of neural network specifically built to manage long-term dependencies in sequential data. It employs unique memory cells equipped with logic gates which process the information flow , allowing it to identify and retain facts over lengthy sequences. This structure is particularly adept at tasks such as language modelling, language translation, and time series analysis because it can maintain context over extended periods.

BiLSTM[10]: Bi-LSTM extension of LSTM which process input sequences in both direction , exploring both future and past contexts. It comprises two LSTM layers: one starts reading sequences from front to rear

(forward LSTM), while the other processes sequences in reverse (backward LSTM). Combining outputs from both layers enhances the model's understanding of context in tasks like sequence labelling and sentiment analysis. Bi-LSTM's bidirectional approach enables it to effectively capture complex dependencies in sequential data.

BERT: BERT, or Bidirectional Encoder Representations from Transformers, is a deep learning model designed by Google for natural language processing tasks. Unlike traditional models, BERT processes text in both directions, capturing context from the words on either side of a given word. This ability enhances its performance in tasks such as question answering, sentiment analysis, and language understanding.

MURIL[11] MuRIL, or Multilingual Representations for Indian Languages, is a machine learning model developed by Google to enhance natural language processing for Indian languages. It is trained on a large volume of Native Indian tongue text, including both translated as well as transliterated data, to better handle the complexities of these languages2. MuRIL outperforms other multilingual models on various tasks, making it a valuable tool for Indian language technologies

IV.IMPLEMENTATION

The proposed system Fig.1 uses MURIL BERT model tuned, trained with large collection of multilingual datasets, to effectively capture contextual information and nuanced features from textual data. The design of the proposed system comprised of several key components shown in Fig.2 [12]

A. Input Processing

Tokenizer: The input text is tokenized using the BERT tokenizer, breaking it down into individual tokens.
Positional Embeddings: Each tokens position in input sequence is encoded with embeddings which elaborates sequential nature of input text.
Segment Embeddings: Segment embeddings are utilized to distinguish between different sentences in the case of multiple sentences, aiding the model in understanding sentence boundaries and relationships.

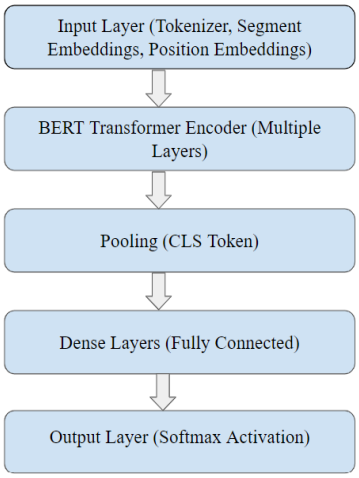

Fig 2- BERT Layers

B. Pre-trained MURIL BERT Model

Transformer Blocks: The pretrained MURIL BERT model comprises different layers of transformer blocks with attention mechanisms. These transformer blocks capture contextual information bidirectionally, allowing the model to describe the meaning of each token related to the entire input sequence.
Bidirectional Representations: MURIL BERT learns bidirectional representations of words, enabling it to capture rich semantic information and contextually relevant features from the input text.

C. Pooling Layer

Pooling Strategy: A pooling layer, employing a pooling strategy such as mean pooling or max pooling, is applied to aggregate information from all tokens into a fixed-size vector. This fixed-size vector serves as a high-level representation of the input text, capturing its salient features.

D. Fully Connected (Dense) Layers

Dense Layers: One or more fully connected (dense) layers are utilized for high-level reasoning and classification. These dense layers process the aggregated information from the pooling layer, extracting higher-order features and patterns relevant to hate and offensive speech detection.

Output Layer: The output layer employs softmax activation for multiclass classification, producing probabilities for each class (e.g., hateful, offensive, profane, or non-hateful) based on the learned features.

E. System Training and Fine-tuning

Training Data: The proposed system is trained on a labelled dataset comprising examples of different classes of Hate speech.
Optimization: The pre-programmed MURIL BERT model refines on the specific objectionable speech detection task using the training data. Fine-tuning adapts the model's parameters to the related context, enhancing its performance and adaptability to the classification task.
Evaluation: The evaluation of the system is calibrated using validation dataset using performance parameters such as accuracy, precision, recall, and F1-score.

F. Integration and Deployment

Once trained and evaluated, the proposed hate and offensive speech detection system can be integrated into various applications and platforms for real-time monitoring and moderation of textual content. Deployment options include web-based APIs, software libraries, or standalone applications, enabling seamless integration with existing systems and workflows.

V. RESULT AND DISCUSSION

Evaluating the performance of hate and displeasing speech detection models in Marathi done based on precision, recall, accuracy, F1-score, and more. Here are key performance evaluation parameters:[13]
Accuracy:
Definition: The proportion of correctly classified instances out of the total instances.
Formula:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Precision:
Definition: The ratio of true positive instances to the sum of true positive and false positive instances. It measures the accuracy of positive predictions.
Formula:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

Recall (Sensitivity):
Definition: The ratio of true positive instances to the sum of true positive and false negative instances. It measures the model's ability to identify all relevant instances.
Formula:

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

• F1-Score:
Definition: The harmonic mean of precision and recall, providing a balance between the two.
Formula:

$$F1 = 2x \text{ Presion x Recall} \quad (4)$$

The proposed system leverages the power of Muril BERT to detect hate and objectionable speech in Marathi.

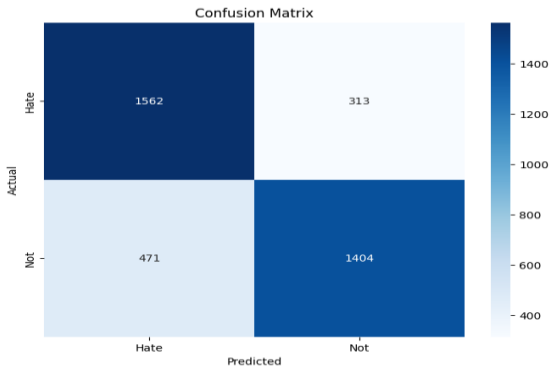

Fig- 3 2- Class Confusion Matrix

In the two-class classification task distinguishing between hate or non-hate speech, the MURIL BERT model achieved an accuracy eq.(1) of 80\% . The confusion matrix Fig.3 obtained from this classification task provides details of the model's performance, allowing for further analysis of precision = 0.82 eq.(2), recall = 0.75 eq.(3), and F1-score = 0.78 eq.(4).

Extending the classification task to four categories - hate, offensive, profane, or non-hate speech, the MURIL BERT model demonstrated robust performance.

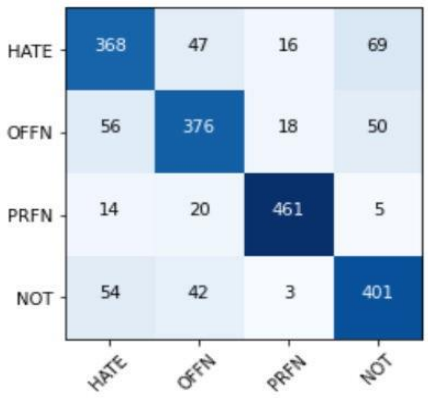

Fig- 4 4-Class Confusion Matrix

The confusion matrix Fig.4 obtained from this classification task provides details of the system's performance. Comprehensive evaluation parameters, such as precision, precision, recall, and F1 score, were calculated to estimate the model's ability to correctly identify the various classes. The results are compared

with the HOF results using linear regression and random forest [5] shown in Table 1.

TABLE 1: RESULTS OF EXPERIMENT

ML Algorithm	Accuracy	Precision	Recall	F-1 score
Logistic Regression [5]	75.955%	85%	39%	54%
Random Forest [5]	77.707%	70%	65%	67%
Deep Learning	80%	82%	75%	78%

The results obtained from the hate and objectionable speech detection for the Marathi language using the MURIL BERT model showcase its effectiveness in accurately classifying text into predefined categories. These findings underscore the effectiveness of using pre-programmed models for addressing objectionable speech identification tasks for multilingual contexts, particularly for languages like Marathi.

A. Key Findings

The research focuses on detecting displeasing speech in Marathi using the MURIL BERT model. Key findings from the study include:

High Accuracy: The model achieved an 80\% accuracy in the two-class classification task (hate vs. non-hate speech).

Effective Multiclass Classification: For the four-class classification (hate, offensive, profane, non-hate), the model demonstrated robust performance, with detailed metrics showing precision of 0.82, recall of 0.75, and an F1-score of 0.78

Model Utilization: The fine-tuned MURIL BERT model, leveraging its pre-training on a large multilingual corpus, effectively captured contextual and nuanced features from the Marathi text

Dataset: The datasets used were substantial, with the two-class dataset containing 30,000 training samples and the four-class dataset containing 21,500 training samples

B. Interpretations

Model Performance: The high accuracy and robust performance metrics indicate that the MURIL BERT model is most acceptable in performing the task of detecting displeasing speech in Marathi. The ability to accurately classify different types of offensive speech suggests that the model can effectively understand and process the linguistic nuances of the language.

Contextual Understanding: The results of evaluation of the model shows the importance of using pre-trained multilingual models like MURIL BERT for natural language processing tasks in less-resourced languages. This model's bidirectional representation and fine-tuning capabilities are crucial in capturing the context and meaning of the text accurately.

Dataset Quality and Size: The comprehensive datasets contributed significantly to the model's performance, ensuring a wide variety of examples for training and validation, which likely helped in achieving the observed accuracy and robustness.

C. Implications

The deployment of such a model can significantly enhance real-time monitoring and moderation of online content, helping to limit the sharing of displeasing contents on social networks.

The success of MURIL BERT in Marathi underscores the potential for similar models to be developed and fine-tuned for other Indian languages, contributing to secure and more wide digital platforms across diverse linguistic contexts.

While the primary goal is to combat hate speech, it is essential to balance this with concerns about censorship and freedom of expression. The model needs to be deployed with mechanisms to ensure that it does not unjustly suppress legitimate speech.

D. Limitations

While the model performs well for Marathi, its performance in other languages or in code-mixed scenarios may vary and requires further research and adaptation.[14]

Although the datasets used are substantial, they may not cover all possible variations of hate and offensive speech. Future work could benefit from even larger and more diverse datasets.

Despite high accuracy, there may be instances where the model misinterprets context, especially in cases involving sarcasm, irony, or cultural references specific to Marathi speakers.

E. Recommendations

The model could be fine-tuned with more diverse datasets, including examples of sarcasm, irony, and other nuanced speech forms, to improve its contextual understanding and accuracy.[15]This work can be extended to include other Indian languages and evaluate the model's performance across different linguistic contexts. This could involve creating and using similar datasets for these languages.

Also the model can leads to frameworks for the ethical deployment of the model, ensuring it respects freedom

of expression while effectively moderating harmful content.

F. Comparative Analysis

Compared to other models and traditional machine learning approaches, the MuRIL BERT model excels in terms of accuracy and effectiveness in the process of detecting hate and objectionable contents in Marathi. Its robust architecture and training on a diverse corpus of Indian languages enable it to outperform other models in this specific task. The uses of transformer-based models and optimized on specific tasks provides a significant advantage over traditional methods such as TF-IDF and LSTM models, which may not capture the same level of contextual detail and nuance.

VI. CONCLUSION

This project built using the MuRIL (Multilingual Representations for Indian Languages) BERT model focused on detecting hate and offensive words in Marathi text, yielding promising results. The model exhibited high accuracy in identifying such language, crucial for fostering a safer online environment. Compared to other models, this work demonstrate superior performance in terms of accuracy and effectiveness. Although the primary goal is to combat hate speech, ethical considerations such as censorship and freedom of expression remain paramount. Future efforts aim to refine the model further, exploring novel approaches and understanding sociocultural factors influencing hate speech in the Marathi-speaking community. Ultimately, this research contributes to creating a more secure and deeper Marathi digital space by effectively identifying hate and offensive language

REFERENCES

[1] Prasanta Mandal, Apurbalal Senapati, and Amitava Nag. Hate-speech detection in news articles: In the context of west bengal assembly election 2021. In Deepak Gupta, Rajat Subhra Goswami, Subhasish Banerjee, M. Tanveer, and Ram Bilas Pachori, editors, Pattern Recognition and Data Analysis with Applications, pages 247–256, Singapore, 2022. Springer Nature Singapore.

[2] Hrushikesh Patil, Abhishek Velankar, and Raviraj Joshi. L3cube- mahahate: A tweet-based Marathi hate speech detection dataset and bert models. In Proceedings of the Third Workshop on Threat, Aggression and Cyberbullying (TRAC 2022), pages 1–9, 2022.

[3] Saurabh Gaikwad, Tharindu Ranasinghe, Marcos Zampieri, and Christo- pher M. Homan. Cross-lingual offensive language identification for low resource languages: The case of marathi, 2021.

[4] Sakshi Kalra, Kushank Maheshwari, Saransh Goel, and Yashvardhan Sharma. Hate speech detection in marathi and code-mixed languages using tf-idf and transformers-based bert-variants. Department of CSIS, BITS Pilani, 333031, Rajasthan, INDIA, 2022,.

[5] Disha Gajbhiye, Swapnil Deshpande, Prerna Ghante, Abhijeet S. Kale, and Deptii D. Chaudhari. Machine learning models for hate speech identification in marathi language. In Fire, 2021.

[6] Pradeep Kumar Roy, Snehaan Bhawal, and Chinnaudayar Navaneethakr- ishnan Subalalitha. Hate speech and offensive language detection in dravidian languages using deep ensemble framework. Computer Speech Language, 75:101386, 2022.

[7] Anjum and Rahul Katarya. Hatedetector: Multilingual technique for the analysis and detection of online hate speech in social networks. Multimedia Tools and Applications, 83(16):48021–48048, May 2024.

[8] Hendri Murfi, Syamsyuriani, Theresia Gowandi, Gianinna Ardaneswari, and Siti Nurrohmah. Bert-based combination of convolutional and recurrent neural network for indonesian sentiment analysis, 2022.

[9] Sepp Hochreiter and Ju¨rgen Schmidhuber. Long short-term memory. Neural Computation, 9(8):1735–1780, 1997.

[10] Abhishek Basu Soham Sarkar Dipankar Das, Anup Kumar Kolya. Computational Intelligence Applications for Text and Sentiment Data Analysis, volume 1. Academic Press, 2023.

[11] Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha P. Talukdar. Muril: Multilingual representations for indian languages. CoRR, abs/2103.10730, 2021.

[12] S. Ravichandiran. Getting Started with Google BERT: Build and train state-of-the-art natural

language processing models using BERT. Packt Publishing, 2021.

[13] Tom M Mitchell. Machine learning, volume 1. McGraw-hill New York, 1997.

[14] Koyel Ghosh, Debarshi Sonowal, Abhilash Basumatary, Bidisha Gogoi, and Apurbalal Senapati. Transformer-based hate speech detection in assamese. In 2023 IEEE Guwahati Subsection Conference (GCON), pages 1–5, 2023.

[15] Deepak Prasad, K.V. Kadambari, Raghav Mukati, and Sunny Singariya. Real-time multi-lingual hate and offensive speech detection in social networks using meta-learning. In TENCON 2023 - 2023 IEEE Region 10 Conference (TENCON), pages 31–35, 2023.