Curated Hinglish Dataset for Deep Learning-Based Misogyny Detection

Deepti Negi[1,*], Himani Maheshwari[1], Chandrakala Arya[1], Umesh Chandra[2] and Gaurav Shukla[2]

[1]*School of Computing, Graphic Era Hill University, Dehradun, 248001, India*
[2]*Department of Statistics & Computer Science, Banda University of Agriculture & Technology, Banda, 210001, India*
E-mail: dnegi@gehu.ac.in
[*]*Corresponding Author*

Received 11 September 2025; Accepted 25 January 2026

Abstract

Social networking sites serves as influential medium for sharing information and communication; however, their mostly unregulated and open frameworks have also turned them into fertile ground for the dissemination of offensive content. The simplicity of sharing content, coupled with user anonymity and vast reach, facilitates the swift circulation of offensive, abusive, and discriminatory remarks. Engagement-driven algorithms may unintentionally promote such harmful content, increasing its visibility and impact. Consequently, offensive content on platforms like Twitter, YouTube, Facebook, and Reddit frequently gains traction, fuelling online hostility, social division, and tangible real-world effects. Offensive content about women is a prevailing subject on social media platforms. Instances of misogyny are disproportionately represented on social media platforms and misogyny is a substantial societal concern which needs to be addressed.

While exhaustive research work has been done for offensive language detection in monolingual settings, the domain of misogyny detection in code-mixed texts is relatively underexplored and there is lack of studies that

Journal of Reliability and Statistical Studies, Vol. 19, Issue 1 (2026), 173–198.
doi: 10.13052/jrss0974-8024.1918

tackle misogyny detection in under-resourced languages. One of the major causes is unavailability of appropriate Hindi-English mixed-coded language dataset. Therefore, in attempt to bridge this research gap our study focuses on developing a dataset and leveraging deep learning techniques on this high-quality curated dataset containing Hindi-English code-mixed comments from multiple social media platforms. This dataset contains 17,234 comments from different social media platforms, annotated manually into misogynistic and non-misogynistic based on the content. Our study also demonstrates a detailed comparison between baseline machine learning, deep learning, and transformer-based approaches utilising our own curated Hinglish dataset. The results indicated that fine-tuned BERT outperformed the deep learning algorithms with highest 0.92 accuracy.

Keywords: Misogyny detection, Hindi English code-mixed text, deep learning algorithm, BERT, offensive language, social media platform.

1 Introduction

Due to the exponential growth of user-generated content on the websites like Twitter, Facebook, Instagram, and YouTube, there is a significant growth in negative rhetoric, such as misogynistic speech. Despite the fact that the social media platform was initially designed to enable people to communicate freely, express themselves, and establish a community, this space has gradually turned out to be the place where harmful, abusive, and hate-racking materials spread [1]. Misogyny, both overt and covert, is one of the most widespread manifestations of hate speech in the Internet and has extended implications at both personal and social levels.

The consequences of misogyny in the Internet are far reaching. Various researchers show that misogynistic material has the effect of causing psychological damage, including anxiety, low self-esteem, emotional distress and retreating participation in the online world. At the social level, unceasing streams of humiliating material towards women recreates unhealthy gender stereotypes, venerates discrimination, disheartens civic contributions of women, and puts up obstacles towards their involvement in the discourse [4, 5]. There is also online misogyny translated onto the offline where it influences attitudes that support a person in terms of violence, discrimination at the work place and exclusion. These effects are additionally compounded in multilingual societies by the fact that negative stories are spread very fast amongst different people with various languages.

Although contents moderation policies can be utilized to reduce such abuse, it has been noted that it is a big challenge to detect misogyny in culturally diverse settings [2]. In India and other comparable multilingual nations, code-mixed language (and especially Hinglish, a mixture of Hindi and English [3]) is used by its users. Code-mixed content contains informal grammar, non-standardized spelling, and cultural overtones making the automated detection particularly tricky. Misogyny, which is a type of gender hate speech, is a language that departs, stereotypes, or encourages hatred against women. The United Nations Sustainable Development Goals (SDGs 3, 5, and 11) label women as a vulnerable segment that has to face continuous inequality and safety issues in the digital realm and in-person [4, 5]. The creation of systems that will deal with online misogyny on a large scale and determine it reliably is necessary, given the psychological, social, and cultural consequences thereof.

Social media is also a place where women discuss personal experience, create awareness of harassment, and solidarity movements [6–8]. But when women speak up or break the rules of patriarchy, they are often faced with further criticism, trolling, victim-blaming and even orchestrated hate campaigns. These antagonistic groups harm their participation in digital spaces, psychological health, and security [9, 10]. Though a number of academic studies deal with the detection of offensive language and hate speech, little exists regarding studies on misogyny, and particularly in code-mixed and under-resourced languages. The lack of high-quality and annotated datasets in Hindi-English mixed content is one of the primary bottlenecks since Hindi is spoken by over 300 million people spread across 24 countries. The absence of this is a major deterrent to the formation of proper misogynist detection devices [11].

Our research paper provides a resource to fill this gap, a high-quality, manually annotated Hinglish dataset and a problem-solving tool, with an objective of identifying and detecting misogyny in social media posts. The data is informative and language-rich and is task-oriented in nature, thus making it easier to conduct future studies on the development of effective systems that could detect and mitigate harmful online communication and promote safer online communication practices.

Objectives of this Study:

- To analyse the linguistic characteristics of Hinglish social media texts.
- To design a pre-processing pipeline suitable for code-mixed misogyny detection.

- To evaluate and compare the effectiveness of traditional machine learning classifiers with advanced deep learning approaches.
- To highlight the challenges and propose potential solutions for detecting hate speech in multilingual and code-mixed contexts.

Major Contributions of this study:

- A curated and annotated Hinglish dataset for detecting misogyny, enriched with detailed linguistic and statistical analysis.
- Implementation of various deep learning approaches for capturing sequential dependencies in code-mixed text.
- A Comparative performance evaluation of deep learning methods against baseline machine learning classifiers such as Naive Bayes and SVM.
- Insights into error patterns and recommendations for future research.

Outline of the paper

This paper is structured as follows: Section Related Work examines the literature on hate speech that is currently available. The suggested work methodology is explained in Section Methodology. The obtained results are presented and interpreted in the Results and Discussion section. Comparing our findings to the most advanced techniques currently in use is done in the Section Comparison with the State of the Art. The study's limitations are described in Section Work Limitations, and final thoughts and recommendations for further research are given in Section Future Work and Conclusion.

2 Related Work

To close this gap, this study presents a carefully selected, high-quality dataset of Hindi-English mixed comments gathered from websites such as YouTube, Facebook, Twitter, and Reddit. The dataset is specifically designed to automatically identify misogynistic remarks in Hindi-English mix language. To measure this dataset's efficacy for autonomous misogyny detection, a few algorithmic models have been applied. The development of algorithms that can automatically detect and highlight sexist sentiments in social media conversation is made possible by this work, which greatly advances the field's research.

A study [12] introduced a large-scale English hate speech dataset curated from sources such as Kaggle and GitHub. The dataset contains 451,709 sentences (371,452 hateful and 80,250 non-hateful) and was further expanded

into a balanced augmented dataset of 726,120 samples. It includes 145,046 unique words, 6,403 contractions, and 377 offensive terms, with each sentence limited to 180 words. This dataset supports NLP pre-processing tasks, reduces out-of-vocabulary issues, and provides a benchmark for hate speech classification. A misogyny [13] dataset was developed from 8,000 Austrian German online forum comments, some containing dialectal or English elements. Comments were annotated on a five-level scale (0–4), ranging from non-sexist to highly sexist, with guidelines and labelling conducted by professional forum moderators. Initial experiments using transformer-based models were reported for both binary and classification across several class tasks. Another study [14] compiled 0.45 million comments from 18 digital sources and proposed a hybrid deep learning approach combining CNN and BiLSTM with attention. The fused model outperformed existing methods, demonstrating strong generalization with 89% accuracy, 0.88 precision, and 0.91 recall. A small dataset (2,258 comments), a new Bengali misogyny dataset [15], Ben-Misog, was curated from diverse online sources. It includes five misogynistic categories and one non-misogynistic class. Various ML and DL models were evaluated, with Bi-LSTM using BERT (multilingual-cased) embedding's achieving the best performance of 91.24% accuracy. A [16] Reddit-based misogyny dataset was created using comments from 12 misogynistic subreddits and 71 randomly selected subreddits for broader coverage. While baseline models achieved high accuracy ($\sim$0.93), they showed low F1-scores ($\sim$0.43) for misogynistic content, highlighting the difficulty of nuanced detection. A novel [17] dataset of 2,229 code-mixed Hinglish YouTube comments on Indian social issues was analysed through exploratory data analysis (EDA), examining distribution, sentiment polarity, keyword importance, and clustering. PCA was further applied to uncover underlying patterns and groupings within the comments.

The study [18] discusses the SemEval-2022 Task 5: Multimedia Automatic Misogyny Identification (MAMI), which aims to detect misogynous content in online memes by analysing both textual and visual information. The dataset consists of approximately 15,000 memes; 10,000 for training and 1,000 for testing, balanced between misogynous and non-misogynous classes. Achieved F1-scores around 0.83 for binary detection and approximately 0.73 for multi-label classification [10] proposed a curated high-quality dataset of 12,698 YouTube comments and replies in Hindi-English code-mixed language for misogynistic attitude detection is proposed. The mBERT model gives best performance on both subtasks, with macro average F1 scores of 0.59 and 0.52, and weighted average F1 scores of 0.66 and

0.65, respectively. A Bengali misogyny dataset [19] was developed from ∼4,000 manually collected social media comments, later expanded to 15,000 via BERT-based augmentation. Expert validation was conducted with sociologists, and classification experiments showed that LSTM with Bangla BERT embedding's performed best, achieving 82.59% accuracy (binary) and 67.27% (multi-class). A study on workplace sexism [20] detection, grounded in Ambivalent Sexism Theory, distinguished between hostile and benevolent forms. Using a curated dataset, various neural models were trained, with a Bi-LSTM plus attention mechanism achieving the best performance (F1 = 0.88), outperforming simpler architectures. A curated dataset of 12,698 YouTube comments was introduced, annotated at two levels for sentiment polarity and content-specific categorization, enabling nuanced analysis [22]. The model achieves accuracy of 63.71% with subtask 1 and 64.78% with subtask 2 respectively. One study Combines Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN) algorithms with GloVe embedding's for sentiment analysis and achieved an accuracy of 90.95% for user comments in Bangla and 97.42% for comments in English [23, 24]. Another study [25] demonstrates the effectiveness of BiLSTM models in capturing language-specific contextual patterns, achieving superior performance across both languages in real-world Reddit data, yielding an F1-score Accuracy of 80% using logistic Regression.

While several studies have introduced datasets and models for detecting hate speech and misogyny across platforms and languages [28], but still there are significant research gaps. Existing datasets are often small, domain-specific, and limited in language variety [15, 26]. Most datasets focus on English, while low-resource and code-mixed languages such as Bengali, Indian English, and regional dialects remain underexplored [27, 28]. Furthermore, most corpora suffer from class imbalance and inconsistent annotation schemes. Few corpora incorporate expert verification or address subtle forms of misogyny such as benevolent sexism. Methodologically, while some studies have adopted advanced Transformer-based and hybrid deep learning architectures, traditional machine learning and shallow models still dominate, and multimodal and cross-lingual approaches remain underexplored. Evaluation practices are also flawed: many studies report only accuracy, which is insufficient in imbalanced contexts [29, 30], leading to unreliable performance assessments [31, 32]. Finally, few studies have explored the practical applicability, scalability, or ethical implications of misogyny detection systems on social media platforms. All these limitations indicate the need for large-scale, diverse, multilingual, and multimodal datasets, as well

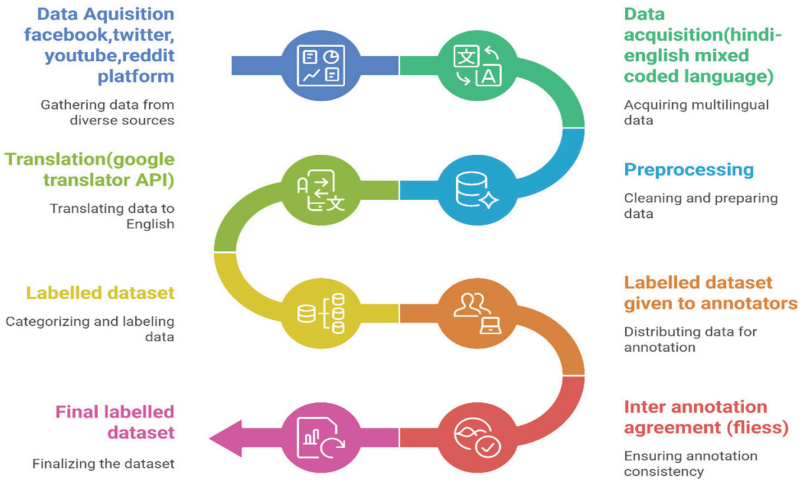

Figure 1 Creating a multilingual dataset: A step-by-step journey.

as robust and interpretable models that can handle nuanced expressions of online misogyny in different contexts.

3 Dataset Creation

The dataset creation process for misogynistic comments involved multiple systematic steps. Initially, data was acquired from diverse social media platforms such as Facebook, Twitter, YouTube, and Reddit, including Hindi-English code-mixed text. The collected data was then translated into English using the Google Translator API, followed by pre-processing for cleaning and preparation. Next, the dataset was categorized and labelled for misogyny detection. Annotators were provided with labelled data for detailed annotation, and inter-annotator agreement (Fleiss' Kappa) was measured to ensure consistency. Finally, after validation and refinement, the final labelled dataset was prepared, forming a reliable resource for automatic misogyny detection. Figure 1 depicts various steps involved in dataset preparation.

Data Acquisition

The Hindi-English code-mixed text comments are gathered from prevalent social networking sites, such as Facebook, Twitter, Reddit, and YouTube. The comments are collected by employing web scraping methods supplemented with manual downloading from these platforms. Figure 1 illustrates the

various steps involved in the data acquisition. To ensure a realistic variety of miscellaneous misogynistic comments, we followed steps like:

***Exploration of Discussions about Prominent Women*:** This aspect focuses on gathering information and discussions related to women who have achieved public recognition in fields like entertainment (actors, models) and arts.

***Specifying keywords*:** The use of broad and targeted terms like "women" alongside more specific or culturally nuanced terms like "papa ki pari" (a Hindi phrase often meaning "father's angel" or "daddy's girl," which can be used endearingly or sometimes in a possessive context) and "moti" (Hindi for "fat," which could be used in a derogatory or descriptive way depending on context) suggests an interest in a wide range of discussions, potentially including those related to appearance, relationships, and public perception.

***Searching threads*:** Searching of Reddit threads dedicated to or written by individuals who identify as anti-women or antifeminist and searching words like-'maalhai', 'item lag rahihai', 'aurat ki aukat', etc.

***Identification of Anti-Women/Antifeminist Content*:** This component specifically targets content generated by individuals or groups holding misogynistic or antifeminist views. YouTube video IDs were searched on the topic of women's safety, political videos that contain comments about women.

To fetch these comments, we used a customized scraper employing platform API's. Initially downloaded comments contained information like text as comments, unique id, username, and timestamp. We downloaded a total of 35,540 comments. These comments were further pre-processed to eliminate extraneous content. The pre-processing included the removal of copies and small comments having less than 12 words. Conversion of entire text to lowercase to maintain consistency. Elimination of unnecessary symbols (e.g., "!!!", "@", "#") that do not contribute to meaning. Figure 2 represents the steps of data pre-processing.

After performing pre-processing of the data, a total of 17,234 comments were extracted and further used for the study.

Annotation Agreement

Before using the curated dataset for further study, to assess the consistency of the annotations, we performed an inter-annotator agreement analysis, which quantifies the extent to which annotators coincide in their labelling decisions.

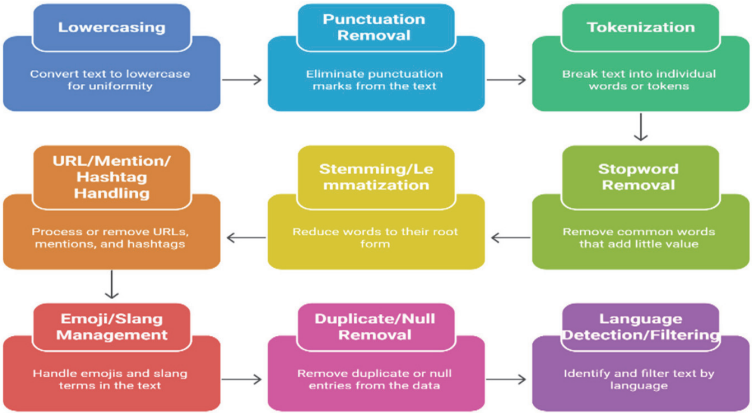

Figure 2 Block diagram of Pre-processing steps.

This evaluation confirms the reliability of the annotation scheme by determining whether multiple annotators can consistently allocate identical labels to the same comment. For calculation purpose we created a matrix where:

Rows = each comment
Columns = each annotator (Annotator1 to Annotator10)
Cells = the label given by each annotator (e.g., 0 = non-misogynistic, 1 = misogynistic)

We calculated the Fleiss Kappa measure, which is applied when three or more annotators are involved in assessment on a categorical scale. The value obtained is 0.84. The resulting score signifies substantial agreement among annotators, indicating well refined and reliable annotation process.

Data Annotation Procedure

The collected corpus of text needs to be annotated into misogynistic and non-misogynistic comments written in Hindi-English coded-mix language. The language of each comment was determined and then translated to English using Google translate API. Then the dataset is labelled into misogynistic and non-misogynistic comments using, a group of 10 annotators. These annotators were selected with the condition that they have a good understanding of both Hindi and English languages. Each annotator is provided 1700 comments to carry out annotation. Clear instructions were given to the annotators that the process of annotation involves classifying comments into

Table 1 Hindi comment translations and labelling

Comment	Translation	Label
Larkiyo Khud Apni Protect KrneChaheyHamesa Larki Dusro Help Ummid Q rakhti h brave apni help khud krohamesakoe	Girls should always protect themselves. A girl expects help from others, but she should always help herself first.	0
अरुंधतिकाकोईविचारहीनहीहैतोविचारों कीलड़ाईक्याखाकलड़ेगी।उसकाकेवल एकहीविचारहैभारतवीरोध।इससेअस्पष्ट मेरेपासकोईशब्दनही।	If there is no idea of Arundhati, then what will fight the fight of ideas. The idea is Bharat Virodh. I have no words unclear from this.	1
jab koi larkigaltikare news channel uskachehra blur diya jata	When a girl makes a mistake, the news channel Blur her face	1
Bhautburalgaaapkastorijankr love u much diidaapbhautkhubsurat	I felt very bad knowing your story, love u that you are very beautiful	0

two classes misogynistic and non-misogynistic and provided the criteria of finding misogynistic comments as –

1. Misogyny refers to hatred, dislike, or prejudice against women.
2. Comment includes disrespectful, abusive, demeaning, or stereotypical remarks toward women.

To ensure the annotation process was well-understood, annotators were briefed prior to commencing their work. Table 1 illustrates comment translation and annotation. The comments are categorized into two classes i.e. Misogynistic (1) and non-misogynistic (0).

Dataset Statistics

The annotated dataset comprises 17234 comments gathered from different social networking platforms like YouTube, Facebook, Twitter and Reddit. The dataset is classified into two categories: misogynistic (1) and non-misogynistic (0) which is task 1.

Figure 3 is the representation of word cloud of Misogynistic comments and figure 4 indicates word length distribution of the curated dataset.

4 Methodology

The methodology for misogynistic comments dataset analysis involves a structured pipeline to ensure effective classification and reliable results. Data was collected from multiple social media platforms, including Facebook,

Figure 3 Misogynistic comments word cloud.

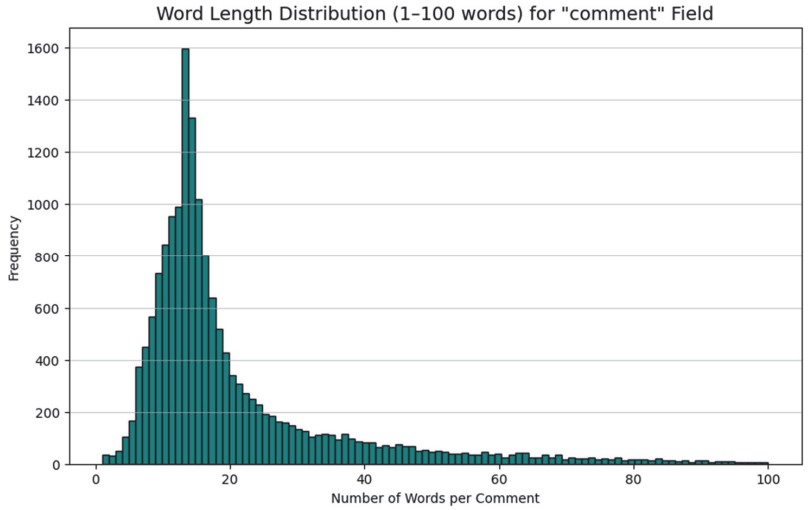

Figure 4 Dataset Word length distribution chart.

Twitter, YouTube, and Reddit, as these platforms host diverse user-generated content. The raw data underwent pre-processing and word embedding techniques to prepare it for further analysis. Pre-processing steps included HTML tag removal, special character removal, data tokenization, and stop word

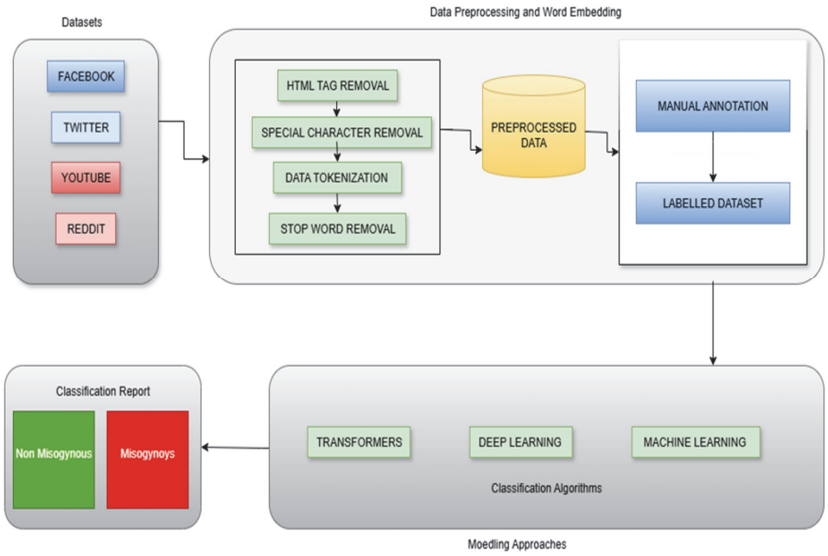

Figure 5 Proposed methodology.

removal, which helped in cleaning the text and making it suitable for computational models. The cleaned and pre-processed data was then subjected to manual annotation, where human annotators categorized comments into misogynistic and non-misogynistic classes, resulting in a labelled dataset.

Once the labelled dataset was finalized, it was used for training and testing classification models. Three categories of modelling approaches were employed: transformers, deep learning, and traditional machine learning algorithms. These models were applied to identify patterns and improve classification accuracy. The final step involved generating a classification report, where results were categorized into two classes: misogynistic and non-misogynistic. This systematic methodology ensures that the dataset is both clean and accurately labelled, while advanced modelling approaches provide robust performance in detecting misogynistic content across multilingual and code-mixed data sources.

Deep Learning Approaches

BiLSTM

An extension of the LSTM (Long Short-Term Memory) network that can process sequences both forward and backward is called a BiLSTM. This is especially useful in NLP, where understanding both the left and right context

of a word improves accuracy.

$$\overrightarrow{h_t} = LSTM_{fwd}\left(x_{t,}\overrightarrow{h_{t-1}}\right) \quad \text{and} \quad \overleftarrow{h_t} = LSTM_{bwd}\left(x_{t,}\overleftarrow{h_{t+1}}\right) \tag{1}$$

$$\textbf{Final} \quad h_t = (\overrightarrow{h_t} : \overleftarrow{h_{t,}}) \tag{2}$$

Support Vector Machine (SVM)

Support Vector Machines (SVM) represent a potent supervised machine learning algorithm utilized for the purpose of both classification and regression tasks. The mechanism of SVMs revolves around the identification of the most suitable hyperplane that effectively segregates distinct classes within the feature space.
Linear hyper plane equation:

$$\omega^T x + b = 0 \tag{3}$$

The distance between a data point xiand the decision boundary can be calculated as

$$d_i = \frac{\omega^T x + b}{\|\omega\|} \tag{4}$$

$$\text{Our aim to } \textit{Min } \frac{1}{2}\|\omega\|^2 \text{ Subject to conditions } y_i(\omega^T x + b) \geq 1 \tag{5}$$

Convolutional Neural Network (CNN)

The convolutional neural network (CNN) is the most emblematic deep learning model. It comprises the input, convolution, pooling, and full connection layers.
The convolutional:

$$h_i = f(W \cdot x_{i:i+k-1} + b) \tag{6}$$

The pooling layer:

$$h_{pool} = \max_i \; h_i \tag{7}$$

Sigmoid:

$$f(s) = \frac{1}{1 + \exp(-s)} \tag{8}$$

Tanh:

$$f(s) = \frac{\sinh(s)}{\cosh(s)} = \frac{e^s - e^{-s}}{e^s + e^{-s}} \tag{9}$$

ReLU:

$$f(s) = Max(0, s) \tag{10}$$

Logistic Regression

Logistic Regression models binary outcomes using a sigmoid function. It estimates weights using maximum likelihood optimization.

$$P(y = 1 \mid x) = \frac{1}{1 + \exp(-(wX + b))} \tag{11}$$

Multilingual-BERT

Multilingual BERT uses transformer encoders with self-attention to learn contextual embedding's across many languages.

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{12}$$

These models were applied to identify patterns and improve classification accuracy. The final step involved generating a classification report, where results were categorized into two classes: misogynistic and non-misogynistic. This systematic methodology ensures that the dataset is both clean and accurately labelled, while advanced modelling approaches provide robust performance in detecting misogynistic content across multilingual and code-mixed data sources.

5 Results & Analysis

To evaluate misogyny detection performance, a range of models were trained using the same dataset. Table 2 represents a comparative analysis of various machine learning and deep learning models on classification task, evaluated using common metrics: Precision, Recall, F1-score, and Accuracy.

The comparative analysis shows that the Fine-tuned BERT model achieved the best overall results with an accuracy of 92.05%, surpassing both traditional and deep learning baselines. The BERT model's contextual

Table 2 Analysis of curated dataset and matrices

Model	Class	Precision	Recall	F1-Score	Accuracy
Logistic Regression	Non-Misogynistic	0.8517	0.9319	0.89	0.8871
	Misogynistic	0.9281	0.8442	0.8842	
	macro avg	0.8899	0.8881	0.8871	
	weighted avg	0.8907	0.8871	0.887	
Linear SVM	Non-Misogynistic	0.8546	0.9259	0.8888	0.8866
	Misogynistic	0.9227	0.8488	0.8842	
	macro avg	0.8887	0.8874	0.8865	
	weighted avg	0.8894	0.8866	0.8865	
Random Forest	Non-Misogynistic	0.8926	0.9064	0.8995	0.9008
	Misogynistic	0.9088	0.8954	0.9021	
	macro avg	0.9007	0.9009	0.9008	
	weighted avg	0.9009	0.9008	0.9008	
BiLSTM	Non-Misogynistic	0.85	0.9	0.87	0.87
	Misogynistic	0.9	0.85	0.88	
	macro avg	0.88	0.88	0.87	
	weighted avg	0.88	0.87	0.87	
CNN	Non-Misogynistic	0.89	0.87	0.88	0.89
	Misogynistic	0.88	0.9	0.89	
	macro avg	0.89	0.89	0.89	
	weighted avg	0.89	0.89	0.89	
Fine-tuned BERT	Non-Misogynistic	0.9218	0.9153	0.9185	**0.9205**
	Misogynistic	0.9193	0.9255	0.9224	
	macro avg	0.9205	0.9204	0.9205	
	weighted avg	0.9205	0.9205	0.9205	

embedding's enabled superior precision and recall balance, leading to the highest F1-score (0.9205). Among traditional machine learning classifiers, the Random Forest model attained the highest accuracy (90.08%), followed by Logistic Regression (88.71%) and Linear SVM (88.66%). Although these models performed reliably, they were limited by their inability to capture deeper linguistic semantics and contextual relationships present in textual data.

The BiLSTM and CNN architectures showed moderate yet strong performance with accuracies of 87% and 89%, respectively. The BiLSTM model's sequential processing capabilities helped it understand contextual flow within sentences, while CNN captured n-gram features effectively. However, both models were unable to match transformer-based models in representing nuanced context and subtle misogynistic cues. The Fine-tuned BERT model

Table 3 Reasoning of each model performance on curated dataset

Model	Reasoning/Interpretation
Logistic Regression	Performs decently as the dataset has somewhat linearly separable patterns. However, struggles with complex contextual or implicit misogyny, limiting recall for the Misogynistic class.
Linear SVM	Similar to Logistic Regression but better in handling high-dimensional text features. Its margin optimization aids generalization, though subtle misogynistic tones reduce recall.
Random Forest	Captures non-linear relationships and interactions. Ensemble learning improves robustness, but context understanding remains limited compared to deep learning models.
BiLSTM	Captures sequential dependencies effectively, useful for word-order-sensitive expressions. Slightly lower performance due to dataset size or lack of pertained embedding's.
CNN	Good at capturing local phrase-level features. Achieves balanced precision and recall but may miss long-range contextual cues needed for subtle misogyny detection.
Fine-tuned BERT	Best-performing model. Its transformer-based architecture captures deep contextual semantics and nuanced tone variations, ensuring superior generalization and balance.

outperformed all others due to its bidirectional attention mechanism, which enables contextual understanding at both word and sentence levels. The model maintained nearly identical precision and recall for both classes, demonstrating strong generalization across non-misogynistic and misogynistic samples. Its superior macro and weighted averages indicate robustness and consistency across the dataset. The fine-tuned BERT model clearly surpasses traditional baseline models by approximately 3 to 4 % accuracy. This means that the dataset has both lexical and mild contextual cues. Table 3. Depicts the reasoning behind the application of various machine learning and deep learning models.

Across most models, recall for Non-Misogynistic samples was slightly higher, suggesting a stronger ability to identify neutral or non-offensive text. Conversely, the Misogynistic class achieved higher precision, indicating that when models labelled content as misogynistic, it was generally accurate. The Fine-tuned BERT model displayed balanced behaviour across both categories, minimizing false positives and false negatives. In summary, the results clearly establish the Fine-tuned BERT model as the most effective architecture for misogyny detection, outperforming traditional and neural network models.

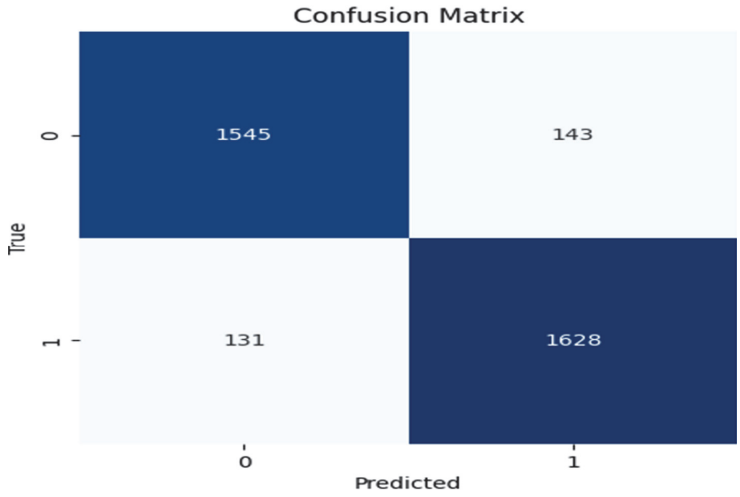

Figure 6 Confusion matrix for fine-tuned BERT.

The confusion matrix in figure 6, shows how well the refined BERT model performs in differentiating between comments that are misogynistic and those that are not. Of all the predictions, only 143 and 131 instances were misclassified, respectively, while 1545 non-misogynistic and 1628 misogynistic comments were correctly identified. High precision and recall are reflected in the near-diagonal dominance, which shows balanced sensitivity in both classes. With a low false negative rate, BERT outperforms conventional machine learning models in identifying subtle linguistic patterns and demonstrates its efficacy in detecting implicit and context-dependent misogyny.

To further validate model robustness the Fine-tuned BERT model was subjected to 5-fold cross-validation (Table 5) in order to further confirm the model's robustness. Accuracy: 0.9218, Precision: 0.9213, Recall: 0.9238, and F1: 0.9227 are the average results across folds, indicating consistent performance with little variation. This stability implies that the model is not over fitting to particular data subsets and shows a strong capacity for generalization. The accuracy variation across folds (from 0.908 to 0.929) demonstrates that BERT continues to exhibit dependable predictive behaviour even in the face of data variability, reflecting the inherent diversity of textual expressions found in the dataset. Overall, the results show a distinct trend: models' capacity to manage contextual, subtle, and non-explicit manifestations of misogyny greatly improves as they move from conventional statistical classifiers to

Table 4 Results for stratified k-fold cross-validation

Fold	Accuracy	Precision	Recall	F1-Score
1	0.925152	0.929594	0.923252	0.926412
2	0.928924	0.927684	0.933485	0.930575
3	0.924282	0.923164	0.928937	0.926041
4	0.918456	0.925158	0.914107	0.919599
5	0.908299	0.902848	0.919272	0.910986
Average	**0.921823**	**0.92129**	**0.923811**	**0.922723**

transformer-based architectures. Table 4. represents the results of stratified k-fold cross-validation.

6 Error Analysis

Although the Fine-tuned BERT model achieved the highest performance among all classifiers, a deeper analysis of misclassified samples reveals several linguistic and contextual challenges inherent in Hindi–English code-mixed misogyny detection. Understanding these model failures provides valuable insights into dataset complexity and opportunities for future improvement.

1. Implicit and Subtle Misogyny

Many comments express misogynistic attitudes indirectly, without explicit abusive terms. Models often failed to detect:

- **Stereotypical gender expectations:**
  "Ladkiyon ko ghar ki zimmedari nibhaani chahiye, bahar kaam karne ki kya zarurat?"
  (Women should handle household duties; why do they need to work outside?)
- **Patronizing benevolence/disguised sexism:**
  "Girls are too delicate, they cannot handle serious responsibilities."

Since such comments lack overt negativity, models – especially traditional ones – classified them as non-misogynistic.

2. Sarcasm and Irony

Sarcasm is difficult for models because literal meaning differs from intended meaning.

- *"Wah, kya 'sanskari' ladki hai, bilkul Instagram pe naachti rehti ho."*
  (Oh, what a 'cultured' girl, always dancing on Instagram.)

Sarcastic remarks often fooled models due to their reliance on surface-level lexical cues.

3. Contextual Ambiguity

Some comments require external or conversational context to infer misogyny:

- Comments referring to a previous post or incident
- Replies that only contain short phrases like *"Typical ladki behaviour"*
- Comments with pronouns like *"she", "her", "wo"* without context

Because the dataset consists of standalone comments, context-dependent cues were missed.

4. Code-Mixed Spelling Variations

Hinglish text often contains inconsistent spellings:

- *"ladki / larki / ldkzi"*
- *"aurat / awrat / auret"*

Deep learning models handled this better than machine learning approaches, but severe misspellings still resulted in misclassification.

5. Cultural/Metaphorical Expressions

Indian socio-cultural expressions carry misogynistic meaning that is implicit:

- *"Aurat ki aukat yaad rakh"* (Remember a woman's place)
- *"Izzat wali ladki aise nahi karti"* (A 'respectable' girl doesn't behave like this)

These expressions were often misclassified, as they carry cultural bias rather than explicit abuse.

6. Mixed Sentiment Comments

Some comments included both positive and negative elements, confusing the models:

- *"She speaks well but like all girls, drama toh hota hi hai."*
  (She speaks well but like all girls, there is always drama.)

Models tended to classify such mixed-content comments as non-misogynistic due to positive sentiment words. In Table 5. Some of the misclassified examples are shown.

Summary of Error Trends

- **Implicit misogyny and sarcasm** were the most challenging categories.
- **Cultural expressions** were frequently misinterpreted.

Table 5 Representative misclassified examples

Comment (Hinglish)	Translation	Ground Truth	Model Prediction	Error Reason
"Aise kapde pehenogi toh log toh bolenge hi."	If you dress like this, people will obviously comment.	1	0	Victim-blaming, implicit misogyny
"Badi shareef banti ho Insta reels banate hue."	Pretending to be innocent while making Instagram reels.	1	0	Sarcasm detected as neutral
"She is too emotional, typical girl."	–	1	0	Stereotyping misinterpreted as descriptive
"Moti ho gayi ho tum, diet karo."	You've become fat, go on a diet.	1	0	Body-shaming but no explicit abuse words

- **Ambiguity and lack of conversational context** led to reduced performance.
- **Fine-tuned BERT** performed better than classical and deep learning models but still struggled with complex socio-linguistic cues.

The error analysis highlights the need for:

- richer contextual datasets,
- sarcasm-aware architectures,
- socio-linguistic feature integration, and
- Multi-label models capturing different misogyny types.

7 Discussion and Future Scope

This study distinguishes itself by presenting a Hindi-English code-mixed language dataset curated from prominent social media platforms, aimed at the automatic detection of misogynistic comments. The availability of appropriate datasets and strong modelling approach that can meaningfully contribute to advancing research in this area and support the development of practical tools for detecting and mitigating misogynistic content in real-world social media environments. This work introduces a dataset consisting of 17234 Hindi-English code-mixed comments, sourced from widely used social media platforms like YouTube, twitter, Facebook and reddit. To ensure consistent labelling and high data quality, ten independent annotators participated in a methodical annotation process.

According to calculations, the inter-annotator agreement for the classification task is high. Next, various algorithmic models are used to identify misogynistic and non-misogynistic comments in the generated dataset. The models incorporate techniques from transformer-based models, deep learning, and machine learning. This study can be extended to

- Multi-class framework that distinguishes between different forms and intensities of misogyny – such as hostile, benevolent, and subtle sexism – to enable deeper and more nuanced analysis.
- Expand beyond Hinglish to include other regional code-mixed languages such as Tamil-English and Bengali-English.
- Integrate contextual and temporal modelling by utilizing conversation-level datasets and context-aware architectures such as hierarchical attention networks and dialogue transformers.

Academic research into misogynistic comments helps predict gender-based violence, understand online harassment, and analyze social identity and group behavior related to violence and discrimination.

This research is crucial for comprehending how misogynistic ideologies form and contribute to harm against women in both digital and physical realms.

Practically, studying misogynistic comments aids in developing automated systems for identifying such content on social media platforms. These practical applications support more effective reporting of hate speech and influence the creation of anti-harassment laws and policies for online environments.

8 Conclusion

This study addresses that gap by introducing a curated dataset in Hinglish code-mixed language, specifically designed for misogyny detection. The dataset comprises 17234 Hindi-English mix coded language comments acquired from various poplar social media platforms through web scrapping; each comment is manually annotated into misogynistic and non-misogynistic. The calculated inter-annotator agreement was high, confirming the reliability and quality of the annotation.

The results indicate strong potential, although further study could explore more advanced or hybrid models to enhance performance. The findings suggest that this dataset is a valuable resource for future research and experimentation in automatic misogyny detection within social media content.

Developing such detection systems has significant real-world implications, including improving online safety for women and mitigating exposure to harmful or offensive content. This dataset marks a significant step forward in extending misogyny detection efforts to non-English and code-mixed linguistic contexts, helping advance inclusive and responsible AI in multilingual digital environments.

Data Availability

The dataset used in this study can be accessed upon reasonable request to the authors for research purposes.

References

[1] E. Aïmeur, S. Amri, and G. Brassard, "Fake news, disinformation and misinformation in social media: a review," Soc Netw Anal Min, vol. 13, no. 1, 2023, doi: 10.1007/s13278-023-01028-5.

[2] M. Mozafari, R. Farahbakhsh, and N. Crespi, "Cross-Lingual Few-Shot Hate Speech and Offensive Language Detection Using Meta Learning," IEEE Access, vol. 10, pp. 14880–14896, 2022, doi: 10.1109/ACCESS .2022.3147588.

[3] S. Bhaskara, S. P, S. Seth, S. Mohanty, and P. Kanwal, "Detection and Comparison of Abusive and Hate Speech in English and Hinglish with Emojis using Deep Learning and Non-Deep Learning Techniques," in 2023 4th International Conference for Emerging Technology (INCET), IEEE, May 2023, pp. 1–7. doi: 10.1109/INCET57972.2023.10170633.

[4] S. Kumbale, S. Singh, G. Poornalatha, and S. Singh, "BREE-HD: A Transformer-Based Model to Identify Threats on Twitter," IEEE Access, vol. 11, no. June, pp. 1–1, 2023, doi: 10.1109/access.2023.3291072.

[5] S. Frenda, B. Ghanem, M. Montes-Y-Gómez, and P. Rosso, "Online hate speech against women: Automatic identification of misogyny and sexism on twitter," Journal of Intelligent and Fuzzy Systems, vol. 36, no. 5, pp. 4743–4752, 2019, doi: 10.3233/JIFS-179023.

[6] I. Kayes and A. Iamnitchi, "Privacy and security in online social networks: A survey," Online Soc Netw Media, vol. 3–4, pp. 1–21, 2017, doi: 10.1016/j.osnem.2017.09.001.

[7] S. Ali, N. Islam, A. Rauf, I. U. Din, M. Guizani, and J. J. P. C. Rodrigues, "Privacy and security issues in online social networks," Future Internet, no. 12, pp. 1–12, 2018, doi: 10.3390/fi10120114.

[8] A. Sharma and R. Kaushal, "Detecting Hate Speech in Hindi in Online Social Media," in 2023 3rd International Conference on Intelligent Communication and Computational Techniques, ICCT 2023, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ICCT 56969.2023.10075749.

[9] P. Parikh, H. Abburi, N. Chhaya, M. Gupta, and V. Varma, "Categorizing Sexism and Misogyny through Neural Approaches," ACM Transactions on the Web, vol. 15, no. 4, Jul. 2021, doi: 10.1145/3457189.

[10] A. Singh, D. Sharma, and V. K. Singh, "Misogynistic attitude detection in YouTube comments and replies: A high-quality dataset and algorithmic models," Comput Speech Lang, vol. 89, Jan. 2025, doi: 10.1016/j.csl.2024.101682.

[11] M. S. Jahan and M. Oussalah, "A systematic review of hate speech automatic detection using natural language processing," Neurocomputing, vol. 546, p. 126232, 2023, doi: 10.1016/j.neucom.2023.126232.

[12] Devansh Mody, YiDong Huang, Thiago Eustaquio Alves de Oliveira, A curated dataset for hate speech detection on social media text, Data in Brief, Volume 46, 2023, https://doi.org/10.1016/j.dib.2022.108832.

[13] B. Krenn, J. Petrak, M. Kubina, and C. Burger, "GERMS-AT: A Sexism/Misogyny Dataset of Forum Comments from an Austrian Online Newspaper," 2024. [Online]. Available: https://www.britannica.

[14] W. Sharif, S. Abdullah, S. Iftikhar, D. Al-madani, and S. Mumtaz, "Enhancing Hate Speech Detection in the Digital Age: A Novel Model Fusion Approach Leveraging a Comprehensive Dataset," IEEE Access, vol. 12, no. December 2023, pp. 27225–27236, 2024, doi: 10.1109/AC CESS.2024.3367281.

[15] R. Kumar, B. Lahiri, and A. K. Ojha, "Aggressive and Offensive Language Identification in Hindi, Bangla, and English: A Comparative Study," SN Comput Sci, vol. 2, no. 1, pp. 1–20, 2021, doi: 10.1007/ s42979-020-00414-6.

[16] E. Guest, B. Vidgen, N. Sastry, G. Tyson, and H. Margetts, "An Expert Annotated Dataset for the Detection of Online Misogyny," 1350. [Online]. Available: https://github.com/ellamguest/.

[17] S. Yadav, A. Kaushik, and K. McDaid, "Exploratory Data Analysis on Code-mixed Misogynistic Comments," Mar. 2024, [Online]. Available: http://arxiv.org/abs/2403.09709.

[18] E. Fersini et al., "SemEval-2022 Task 5: Multimedia Automatic Misogyny Identification." [Online]. Available: https://cloud.google.com/visio n/docs/.

[19] S. Sultan Saruar Jahan et al., "Deep learning based misogynistic Bangla text identification from social media," Computing and Informatics, vol. 42, pp. 993–1012, 2023, doi: 10.31577/cai.

[20] D. Grosz and P. Conde-Cespedes, "Automatic Detection of Sexist Statements Commonly Used at the Workplace," Jul. 2020, [Online]. Available: http://arxiv.org/abs/2007.04181.

[21] Karishma, S., and Akila, V. Multiclass Classification of Hindi-English Code Mixed Misogyny Comments Using Recurrent Neural Networks. 2025 International Conference on Emerging Technologies in Engineering Applications (ICETEA), 1–6.

[22] S. R. R. Rahman, J. U. Tanvin and M. N. Islam, "A Hybrid Deep Learning Model for Sentiment Analysis of Multilingual Comments on Trending YouTube Videos" *2025 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, Chittagong, Bangladesh, 2025, pp. 1–6, doi: 10.1109/ECCE64574.2025.11013083.

[23] A. Phadte and M. L. Dhore, "Sentiment Analysis of English-Marathi-Konkani Code-Mixed Social Media Text: A Multilingual Approach," 2025 International Conference on Computing Technologies (ICOCT), Bengaluru, India, 2025, pp. 1–10, doi: 10.1109/ICOCT64433.2025.11118344.

[24] A. Phadte and M. L. Dhore, "Advancements in Sentiment Analysis of Code-Mixed Text: A Survey of Multilingual Models and Emerging Innovations," *2025 9th International Conference on Computing, Communication, Control and Automation (ICCCBEA)*, Pune, India, 2025, pp. 01–10, doi: 10.1109/ICCUBEA65967.2025.11283754.

[25] D. S. AbdElminaam et al., "Harnessing Machine Learning and Deep Learning for Multilingual Sentiment Analysis: A Comparative Study on Arabic and English Social Media Data," *2025 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC)*, Cairo, Egypt, 2025, pp. 198–205, doi: 10.1109/MIUCC66482.2025.11196877.

[26] A. Tontodimamma, E. Nissi, A. Sarra, and L. Fontanella, "Thirty years of research into hate speech: topics of interest and their evolution," Scientometrics, vol. 126, no. 1, pp. 157–179, Jan. 2021, doi: 10.1007/s11192-020-03737-6.

[27] M. Mozafari, R. Farahbakhsh, and N. Crespi, "A BERT-Based Transfer Learning Approach for Hate Speech Detection in Online Social Media," Oct. 2019, [Online]. Available: http://arxiv.org/abs/1910.12574.

[28] S. Khan et al., "BiCHAT: BiLSTM with deep CNN and hierarchical attention for hate speech detection," Journal of King Saud University –

Computer and Information Sciences, vol. 34, no. 7, pp. 4335–4344, 2022, doi: 10.1016/j.jksuci.2022.05.006.

[29] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep learning for hate speech detection in tweets," 26th International World Wide Web Conference 2017, WWW 2017 Companion, no. 2, pp. 759–760, 2017, doi: 10.1145/3041021.3054223.

[30] S. Kamble and A. Joshi, "Hate Speech Detection from Code-mixed Hindi-English Tweets Using Deep Learning Models," 2018, [Online]. Available: http://arxiv.org/abs/1811.05145.

[31] A. Khan, A. Ahmed, S. Jan, M. Bilal, and M. F. Zuhairi, "Abusive Language Detection in Urdu Text: Leveraging Deep Learning and Attention Mechanism," IEEE Access, vol. PP, p. 1, 2024, doi: 10.1109/ACCESS .2024.3370232.

[32] V. B. M. L. V. P. Komal Florio, 2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW). 2019.

Biographies

Deepti Negi received her master's degree in computer application from Hemavati Nandan Bahuguna University in 2008. She is currently working as an Assistant Professor at the School of computing, Graphic Era Hill University. Her research areas include natural language processing, deep learning, and social network analysis.

Himani Maheshwari is currently working as Assistant Professor in School of Computing, Graphic Era Hill University, Dehradun. She completed her Graduation from MJP Rohilkhand University, Bareilly and Post-Graduation from Uttarakhand Technical University, Dehradun. She received her Ph.D. from IIT, Roorkee. She has qualified UGC NET and GATE. She has published 45 research papers in different reputed national and international journals, 10 book chapters and attained 8 copy rights. Her area of specialization is Artificial Intelligence, Big Data Analysis and Machine Learning.

Chandrakala Arya is currently working as Assistant Professor in School of Computing, Graphic Era Hill University, Dehradun. She completed her

Graduation from Kumaon University, Nainital and Post Graduation from Uttarakhand Technical University, Dehradun. She Received her Ph.D. from Babasaheb Bhimrao Ambedkar University (Central University) Lucknow. She has qualified UGC NET and GATE. She has published 30 research papers in different reputed national and international conferences and journals. Her area of specialization is Artificial Intelligence, and Machine Learning.

Umesh Chandra is currently working as Assistant Professor in Department of Statistics & Computer Science, College of Agriculture, Banda University of Agriculture & Technology, Banda. He completed his Graduation from Kumaun University, Nainital and Post Graduation from Uttarakhand Technical University, Dehradun. He Received his Ph.D. from IIT, Roorkee. He has published 42 research papers in different reputed national and international journals, 8 book chapters and attained 7 copy rights. He is author of one edited book. His area of specialization is GIS, Artificial Intelligence, Big Data Analysis and Machine Learning.

Gaurav Shukla is currently working as Assistant Professor in Department of Statistics & Computer Science, College of Agriculture, Banda University of Agriculture & Technology, Banda. He completed his Graduation and Post Graduation from MJP Rohilkhand University, Bareilly. He received his Ph.D. also from MJP Rohilkhand University, Bareilly. He has published 48 research papers in different reputed national and international journals, 7 book chapters and attained 1 copy right. He is author of one book. His area of specialization is Life Testing Models, Applied Statistics and Machine Learning.