

Department of Computer Science and Engineering
Bangladesh University of Business and Technology (BUBT)



CSE 478: Literature Review Records

Neural Network and Fuzzy System Project Title	Bangla Hate Speech Identification
Supervisor Name & Designation	Name: Khan Md. Hasib & Designation: Assistant Professor, Department of CSE, BUBT
Course Teacher's Name & Designation	Name: Khan Md. Hasib & Designation: Assistant Professor, Department of CSE, BUBT

Student's Information

Student's Name	Student's ID	Paper No.	Page No.
Arman Habib Shihab	19202103121	1, 2, 3, 4, 5	2-7
MD. Mehraz Hosen	19202103122	6, 7, 8, 9, 10	8-20
Sanzida Akter	19202103258	11, 12, 13, 14, 15	21-e
Afrina Akter Mim	19202103310	16, 17, 18, 19, 20	25-34
Sagor Kumar Saha	19202103423	21, 22, 23, 24, 25	35-44

Aspects	Paper # 1 (Title)
Title / Question (What is problem statement?)	Inculcating Context for Emoji Powered Bengali Hate Speech Detection using Extended Fuzzy SVM and Text Embedding Models.
Objectives / Goal (What is looking for?)	The goal of this paper is to develop an efficient automated tool for detecting hate speech in Bengali language on social media platforms.
Methodology / Theory (How to find the solution?)	Hate Similarity with Morphology Analysis (HS): Morphological analysis using the Word2Vec model to identify hate terms and similarities between hate lexicons and tweet terms. mBERT Vectorization: Utilizing the mBERT (Multilingual Bi-directional Encoder Representation from Transformer) model for contextual embedding of Bengali text. FSVMCIL (Fuzzy Support Vector Machine for class imbalanced learning): Extending fuzzy SVM to handle class imbalance, outliers, and noise in hate speech detection. Emoji-to-Text Conversion: Converting emojis in the text to text form for improved contextual analysis.
Software Tools (What program/software is used for design, coding and simulation?)	Python language
Test / Experiment How to test and characterize the design/prototype?	The proposed Bengali hate speech detection model was evaluated through extensive experiments using the Bengali hate speech dataset. The model's performance was assessed using metrics such as weighted F1 score, precision, recall, and accuracy. It achieves high accuracy and F1-score for detecting hate speech in Bengali text, addressing class imbalance and other challenges.
Simulation/Test Data (What parameters are determined?)	The paper uses a Bengali hate speech dataset called NNTI, consisting of 29,999 posts with labels of "Hate" and "Non-Hate." The dataset covers various categories, including sports, entertainment, religion, crime, politics, celebrity, and others. Parameters such as F1-score, accuracy, precision, and recall are determined to evaluate the model's performance.
Result / Conclusion (What was the final result?)	The final result of the research is that the proposed hate speech detection model, mBERT uncased + FSVMCIL + HS, consistently outperforms other approaches and baseline models. It achieved a 2.35% increase in F1-score and a 9.11% increase in accuracy compared to baseline models.
Obstacles/Challenges (List the methodological obstacles if authors mentioned in the article)	The paper mentions the challenges posed by the Bengali language's unique script and characteristics, including implicit hate speech, wrong spellings, and inaccurate sentence structures. Additionally, class imbalance and the presence of emojis in hate speech texts are addressed as challenges.
Terminology (List the common basic words frequently used in this research field)	Hate Speech Detection, Bengali Language, Morphological, Analysis, Fuzzy Classifier, FSVMCIL, mBERT, Hate Lexicon, Contextual Analysis, Emoji-to-Text Conversion, and Performance Metrics are some of the terms frequently used in this field of study.

Review Judgment (Briefly compare the objectives and results of all the articles you reviewed)	
Review Outcome (Make a decision how to use/refer the obtained knowledge to prepare a separate and new methodology for your own research project)	

Aspects	Paper # 2 (Title)
Title / Question (What is problem statement?)	Multimodal Hate Speech Detection from Bengali Memes and Texts.
Objectives / Goal (What is looking for?)	The goal of the research paper is to address the problem of hate speech detection in the under-resourced Bengali language by utilizing multimodal information from Bengali memes and texts.
Methodology / Theory (How to find the solution?)	The paper utilizes machine learning (ML) and deep learning (DL) approaches to detect hate speech from multimodal Bengali memes and texts.
Software Tools (What program/software is used for design, coding and simulation?)	Python language
Test / Experiment How to test and characterize the design/prototype?	Multiple machine learning and deep learning models are trained and evaluated using a labeled dataset. A key aspect of the experiment involves the application of 5-fold cross-validation to rigorously assess model performance. The research employs a suite of performance metrics, including F1-score and MCC, to ascertain the effectiveness of the models.
Simulation/Test Data (What parameters are determined?)	The process of experimentation is dependent on determining a number of parameters, including performance metrics such as F1-score and MCC. In order to determine the best successful method for Bengali hate speech identification, the research tests several model architectures and multimodal fusion combinations.
Result / Conclusion (What was the final result?)	The optimal model for hate speech detection in Bengali emerges as XLM-RoBERTa. Moreover, the multimodal fusion technique, particularly when applied to XLM-RoBERTa and DenseNet-161, yields an impressive MCC score of 0.67.
Obstacles/Challenges (List the methodological obstacles if authors mentioned in the article)	The authors highlight the lack of computational resources for natural language processing (NLP) in Bengali, despite its diversity and millions of native speakers. Moreover, some potential challenges could include limited labeled data and the need for explainable AI in hate speech detection.
Terminology (List the common basic words frequently used in this research field)	Some of the common basic words frequently used in this research field include hate speech, Bengali language, text, images, deep learning, machine learning, multimodal fusion, precision, recall, F1-score, and MCC.
Review Judgment (Briefly compare the objectives and results of all the articles you reviewed)	
Review Outcome (Make a decision how to use/refer the obtained knowledge to prepare a separate and new methodology for your own research project)	

Aspects	Paper # 3 (Title)
Title / Question (What is problem statement?)	G-BERT: An Efficient Method for Identifying Hate Speech in Bengali Texts on Social Media.
Objectives / Goal (What is looking for?)	The paper aims to propose an efficient method, G-BERT, for identifying hate speech in Bengali social media posts and mitigating online hate speech.
Methodology / Theory (How to find the solution?)	The methodology involves using a combination of deep learning models, namely BERT and GRU. BERT is utilized to extract contextual information from Bengali text, while GRU processes this information sequentially. The Softmax activation function is employed for classification.
Software Tools (What program/software is used for design, coding and simulation?)	Python for programming, Keras and TensorFlow for developing and training the deep learning model, and NumPy for basic mathematical operations.
Test / Experiment How to test and characterize the design/prototype?	The experiment involves training and evaluating the G-BERT model for hate speech detection in Bengali text. The dataset is divided into training, testing, and validation subsets. Various experiments are conducted by adjusting parameters like batch size, learning rate, and maximum sequence length. The performance is assessed using standard evaluation metrics.
Simulation/Test Data (What parameters are determined?)	The parameters determined in the experiments include batch size, learning rate, and maximum sequence length. These parameters impact the model's performance and are adjusted to optimize the results.
Result / Conclusion (What was the final result?)	The G-BERT model achieves a high accuracy of 95.56%, along with impressive precision (95.07%), recall (93.63%), and F1-score (92.15%). It outperforms other machine learning models and previous studies on Bengali hate speech detection.
Obstacles/Challenges (List the methodological obstacles if authors mentioned in the article)	Identifying hate speech in Bengali due to the language's unique script and grammar.
Terminology (List the common basic words frequently used in this research field)	Common basic words frequently used in this research field may include terms related to hate speech detection, natural language processing, deep learning, BERT, GRU, precision, recall, F1-score, and more.
Review Judgment (Briefly compare the objectives and results of all the articles you reviewed)	
Review Outcome (Make a decision how to use/refer the obtained knowledge to prepare a separate and new methodology for your own research project)	

Aspects	Paper # 4 (Title)
Title / Question (What is problem statement?)	A Robust Hybrid Machine Learning Model for Bengali Cyber Bullying Detection in Social Media.
Objectives / Goal (What is looking for?)	The aim is to address the prevalence of cyberbullying in the Bengali language on social media platforms, particularly Facebook, and improve the detection and classification of cyberbullying instances in Bengali.
Methodology / Theory (How to find the solution?)	The proposed methodology for addressing Bengali cyberbullying detection on social media involves data collection, text preprocessing, and feature extraction using TFIDF. It also includes dataset balancing through Instance Hardness Threshold resampling, followed by the application of machine learning models, specifically Logistic Regression for classifying data and Decision Trees for prediction based on characteristic subsets.
Software Tools (What program/software is used for design, coding and simulation?)	Python language
Test / Experiment How to test and characterize the design/prototype?	The proposed hybrid machine learning model for Bengali cyberbullying detection in social media was tested and characterized using a publicly available Bengali text dataset consisting of 44,001 comments.
Simulation/Test Data (What parameters are determined?)	The parameters for testing the hybrid machine learning model for Bengali cyberbullying detection include text preprocessing, TFIDF-based feature extraction, dataset resampling with IHT, and performance evaluation using metrics like accuracy, precision, recall, f1-score, ROC curve, confusion matrix, MSE, MAE, and RMSE.
Result / Conclusion (What was the final result?)	The model achieved high performance in binary and multilabel classification, with accuracy rates of 98.57% and 98.82% respectively, surpassing the performance of previous models.
Obstacles/Challenges (List the methodological obstacles if authors mentioned in the article)	The authors did not explicitly mention any methodological obstacles or challenges in the paper.
Terminology (List the common basic words frequently used in this research field)	Some of the common basic words frequently used in this research field Cyberbullying, Machine Learning, Text preprocessing, Feature extraction, TfidfVectorizer (TFID), Resampling, Deep Learning and Transformer-based models.
Review Judgment (Briefly compare the objectives and results of all the articles you reviewed)	
Review Outcome (Make a decision how to use/refer the obtained knowledge to prepare a separate and new methodology for your own research project)	

Aspects	Paper # 5 (Title)
Title / Question (What is problem statement?)	Reason Based Machine Learning Approach to Detect Bangla Abusive Social Media Comments.
Objectives / Goal (What is looking for?)	The goal is to identify abusive Bangla language using a novel approach that utilizes annotated translated Bengali corpora and adds a formal justification in each remark for classification.
Methodology / Theory (How to find the solution?)	The research methodology involves collecting comments from various Facebook pages, translating Bengali comments to English, annotating them as abusive or non-abusive, and then preprocessing the data. Feature extraction is done using unigram and bigram features with TF-IDF and Count Vectorizers, followed by employing multiple machine learning classifiers like Logistic Regression, Naïve Bayes, Decision Tree, Random Forest, SVM, AdaBoost, and Gradient Boost for evaluation using metrics such as precision, recall, accuracy, and F1-score.
Software Tools (What program/software is used for design, coding and simulation?)	Python language.
Test / Experiment How to test and characterize the design/prototype?	The testing phase involves training the machine learning classifiers with extracted features and evaluating their performance using metrics such as precision, recall, accuracy, and F1-score.
Simulation/Test Data (What parameters are determined?)	The primary parameters determined during experiments are precision, recall, accuracy, and F1-score for each classifier, both with and without the addition of semantic meaning to comments.
Result / Conclusion (What was the final result?)	The research concludes that incorporating semantic meaning into comments significantly enhances the performance of machine learning classifiers in detecting abusive language in Bengali. Logistic Regression achieved the highest accuracy of 97%, surpassing other classifiers.
Obstacles/Challenges (List the methodological obstacles if authors mentioned in the article)	It implies the challenges of working with a low-resource language like Bengali and the complexity of accurately detecting abusive language.
Terminology (List the common basic words frequently used in this research field)	Commonly used terms in this research field include abusive language, Bengali corpus, machine learning classifiers, TF-IDF Vectorizer, Count Vectorizer, precision, recall, accuracy, and F1-score.
Review Judgment (Briefly compare the objectives and results of all the articles you reviewed)	
Review Outcome (Make a decision how to use/refer the obtained knowledge to prepare a separate and new methodology for your own research project)	

Aspects	Paper # 6 (Title)
Title / Question (What is problem statement?)	G-BERT: An Efficient Method for Identifying Hate Speech in Bengali Texts on Social Media This study focuses on the identification of hate speech in Bengali texts posted on social media platforms.
Objectives / Goal (What is looking for?)	The primary aim of this research is to combat hate speech on social media platforms, particularly in the Bengali language. The researchers have developed an effective hate speech detection model called G-BERT.
Methodology / Theory (How to find the solution?)	<p>G-BERT combines the BERT architecture with a GRU model and Softmax activation function.</p> <ul style="list-style-type: none"> • The BERT architecture is a pre-trained language model that can be used to extract contextual information from text. • The GRU model is a recurrent neural network that can process text one word at a time, taking into account the context of the previous words. • The Softmax activation function is used to convert the output of the GRU model into a probability distribution over the different classes. <p>The G-BERT model is trained on a dataset of 16,800 posts and comments and evaluates the performance of the G-BERT model on a held-out test set of 2,000 posts and comments. The results show that the G-BERT model outperforms other baseline models, with an accuracy of 95.56</p>
Software Tools (What program/software is used for design, coding and simulation?)	<p>The authors used the following software tools for the design, coding, and simulation of the G-BERT model:</p> <ul style="list-style-type: none"> • Python • TensorFlow • PyTorch <p>TensorFlow with Keras was utilized for applying the deep learning model, and TensorFlow was employed for executing the neural network's GPU performance. NumPy, a Python library, was used for basic mathematical operations.</p>
Test / Experiment How to test and characterize the design/prototype?	<p>Test: The authors evaluated the performance of the G-BERT model on a held-out test set of 2,000 posts and comments.</p> <p>The G-BERT model was evaluated using various metrics, including accuracy (95.56%), precision (95.07%), recall (93.63%), and F1-score (92.15%), to assess its performance in identifying hate speech.</p> <p>The authors conducted a number of experiments to evaluate the performance of the G-BERT model by:</p> <ul style="list-style-type: none"> • Varying the size of the training dataset. • Varying the hyperparameters of the G-BERT model. • Evaluating the performance of the G-BERT model on different datasets.

<p>Simulation/Test Data (What parameters are determined?)</p>	<p>The key parameters determined in the experiments include the batch size, number of epochs, learning rate, and maximum sequence length of the input data.</p> <p>Parameters: The following parameters were determined for the simulation and test data:</p> <ul style="list-style-type: none"> • The size of the training dataset was 16,800 posts and comments. • The batch size was 32. • The learning rate was 0.0001. • The maximum sequence length was 128. <p>These parameters are adjusted to identify the optimal configuration for the G-BERT model.</p>
<p>Result / Conclusion (What was the final result?)</p>	<p>The G-BERT model achieved the following results:</p> <ul style="list-style-type: none"> • Accuracy: 95.56% • Precision: 95.07% • Recall: 93.63% • F1-score: 92.15% <p>These results suggest that the G-BERT model is a promising approach for identifying hate speech in Bengali texts.</p>
<p>Obstacles/Challenges (List the methodological obstacles if authors mentioned in the article)</p>	<ol style="list-style-type: none"> 1. The lack of a large dataset of labeled Bengali texts for training the G-BERT model. 2. The complexity of the Bengali language, which makes it difficult to identify hate speech.

<p>Terminology (List the common basic words frequently used in this research field)</p>	<p>Hate speech Speech that promotes or incites violence, discrimination, or hatred against individuals or groups based on attributes such as race, religion, ethnicity, etc.</p> <p>BERT Bidirectional Encoder Representations from Transformers.</p> <p>GRU Gated Recurrent Unit. A type of neural network architecture suited for sequential data processing.</p> <p>Softmax activation function A function used to convert the output of a neural network into a probability distribution.</p> <p>Precision A metric that measures the accuracy of positive predictions, indicating the proportion of correctly predicted positive samples.</p> <p>Recall A metric that measures the ability of a model to identify all relevant instances in the dataset, indicating the proportion of true positive samples.</p> <p>F1-Score A metric that combines precision and recall to provide a balance between them, often used when dealing with imbalanced datasets.</p>
<p>Review Outcome (Make a decision how to use/refer the obtained knowledge to prepare a separate and new methodology for your own research project)</p>	<p>Things that I learn for that paper:</p> <ol style="list-style-type: none"> 1. Collect a larger dataset of Bengali hate speech data. 2. Fine-tuning the G-BERT model on different hyperparameter settings. 3. Use a different pre-trained language model, such as XLNet or RoBERTa. 4. Train some base existing models (Example: G-BERT, G-BERT, Conv-LSTM, XLM-RoBERTa) and then train another model (meta-model) to learn how to best combine their predictions.***

Aspects	Paper # 7 (Title)
Title / Question (What is problem statement?)	Problem is to develop a novel stacked ensemble for hate speech recognition This study focuses on the identification of hate speech in Bengali texts posted on social media platforms.
Objectives / Goal (What is looking for?)	paper aims to develop a novel stacked ensemble for hate speech recognition
Methodology / Theory (How to find the solution?)	The methodology used in the study is a two-stage approach: Base-level classification: This stage involves training a set of base classifiers on the training set. The base classifiers used in the study were SVM, LR, XGBoost, KNN, NB, RF, and E-trees. Meta-level classification: This stage involves training a meta-classifier on the predictions of the base classifiers. The meta-classifier used in the study was also a logistic regression model.
Software Tools (What program/software is used for design, coding and simulation?)	. The software tools used in the study were: <ul style="list-style-type: none"> • Python • Scikit-learn • Keras • TensorFlow
Test / Experiment How to test and characterize the design/prototype?	The study used three publicly available datasets from Twitter to evaluate the proposed method: 1. Davidson dataset <ul style="list-style-type: none"> • 24,783 tweets • 18,825 hateful tweets • 5,958 non-hateful tweets • Training set: 16,495 tweets • Development set: 2,288 tweets • Test set: 5,958 tweets 2. HatEval dataset <ul style="list-style-type: none"> • 10,000 tweets • 4,000 hateful tweets • 6,000 non-hateful tweets • Training set: 7,000 tweets • Development set: 1,000 tweets • Test set: 2,000 tweets

<p>Test / Experiment How to test and characterize the design/prototype?</p>	<p>3. COVID-HATE dataset</p> <ul style="list-style-type: none"> • 10,000 tweets • 3,000 hateful tweets • 7,000 non-hateful tweets • Training set: 7,000 tweets • Development set: 1,000 tweets • Test set: 2,000 tweets <p>The proposed method was evaluated using the standard stacking approach, which is a machine learning ensemble meta-algorithm that combines the predictions of multiple base estimators in order to improve the performance on a given task.</p> <p>The training set was split into two parts: the development set and the test set. The development set was used to tune the hyperparameters of the models, and the test set was used to evaluate the final performance of the models.</p> <p>The study also evaluated the performance of the proposed method with different combinations of base classifiers and meta-classifiers. The best performance was achieved with the combination of SVM, LR, and E-tree as base classifiers and logistic regression as meta-classifiers.</p>
--	---

<p>Simulation/Test Data (What parameters are determined?)</p>	<p>The parameters that were determined in the study are:</p> <ul style="list-style-type: none"> • The number of base classifiers: The study used 8 base classifiers - SVM, LR, XGBoost, KNN, NB, RF, and E-trees. • The type of meta classifier: The study used a logistic regression model as the meta classifier. • The hyperparameters of the base classifiers and meta classifier: The hyperparameters of the base classifiers and meta classifier were tuned using the development set. • The split percentage of the training and test sets: 70/30
<p>Result / Conclusion (What was the final result?)</p>	<p>The proposed method was evaluated on three publicly available datasets from Twitter:</p> <ul style="list-style-type: none"> • The Davidson dataset • The HatEval dataset • The COVID-HATE dataset <p>The results showed that the proposed method outperformed the standard stacking approach for all datasets. The proposed method achieved an F1-score of 92.5% on the Davidson dataset, 88.0% on the HatEval dataset, and 85.5% on the COVID-HATE dataset. In comparison, the standard stacking approach achieved an F1-score of 90.5%, 85.0%, and 82.5% on the Davidson, HatEval, and COVID-HATE datasets, respectively.</p>
<p>Obstacles/Challenges (List the methodological obstacles if authors mentioned in the article)</p>	<ol style="list-style-type: none"> 1. The lack of large and well-labeled datasets for hate speech recognition 2. The difficulty of defining what constitutes hate speech 3. The dynamic nature of hate speech

<p>Terminology (List the common basic words frequently used in this research field)</p>	<p>The common basic words frequently used in this research field are:</p> <p>Hate speech: Any kind of communication that attacks a person or group on the basis of attributes such as race, religion, ethnic origin, national origin, sex, disability, sexual orientation, or gender identity.</p> <p>Stacking: A machine learning ensemble meta-algorithm that combines the predictions of multiple base estimators in order to improve the performance on a given task.</p> <p>Logistic regression: A statistical model that predicts the probability of a binary outcome.</p> <p>SVM: Support Vector Machine - A machine learning algorithm that finds the best hyperplane to separate two classes of data.</p> <p>XGBoost: A machine learning algorithm that combines decision trees in order to improve the performance on a given task.</p> <p>KNN: K-Nearest Neighbor - A machine learning algorithm that predicts the label of a new data point by finding the k most similar data points in the training set.</p> <p>NB: Naive Bayes - A machine learning algorithm that predicts the probability of a given class label by counting the number of times each feature appears in the training set.</p> <p>RF: A machine learning algorithm that builds multiple decision trees in order to improve the performance on a given task.</p> <p>E-trees:: A machine learning algorithm that builds decision trees that are ensembles of trees.</p>
--	---

Aspects	Paper # 8 (Title)
Title / Question (What is problem statement?)	Bangla hate speech detection on social media using attention-based recurrent neural network The problem statement of the paper is to develop a method for detecting Bangla hate speech on social media. It is difficult to detect hate speech, especially in a language like Bangla, where there is limited research on the topic.
Objectives / Goal (What is looking for?)	The objectives of the paper are to: <ul style="list-style-type: none"> • Develop a method for detecting Bangla hate speech on social media that is accurate and efficient. • Evaluate the proposed method on a dataset of Bangla comments from Facebook pages. • Compare the proposed method to other state-of-the-art methods.
Methodology / Theory (How to find the solution?)	The authors propose a novel approach to detecting Bangla hate speech on social media using an attention-based recurrent neural network. The authors use a Bangla Emot Module to detect the emotions lying behind emojis and emoticons, and they use an attention-based decoder to focus on the most important words in the text.
Software Tools (What program/software is used for design, coding and simulation?)	. The software tools used in the paper are: <ul style="list-style-type: none"> • Python • TensorFlow • Keras • Scikit-learn
Test / Experiment How to test and characterize the design/prototype?	The authors evaluated their approach on a dataset of Bangla comments from Facebook pages. The dataset was divided into training, development, and test sets. The authors trained their model on the training set, and they evaluated its performance on the development set. The final performance of the model was evaluated on the test set.
Simulation/Test Data (What parameters are determined?)	A dataset of 7,425 Bengali comments from Facebook pages, consisting of seven distinct categories of hate speeches, was used to train and evaluate the model. The dataset contains a mix of hateful and non-hateful comments.
Result / Conclusion (What was the final result?)	The attention-based decoder algorithm achieved the highest accuracy of 77 The model gives the best performance with high precision (0.78), high recall (0.75), and high F1-score (0.78), while the LSTM-based model achieved precision of 0.72, recall of 0.71, and F1-score of 0.72, and the GRU-based model achieved precision, recall, and F1-score of 0.70 and 0.69 respectively. The final outcome for both LSTM decoder and GRU decoder models was 74% accuracy, while the attention-based decoder outperformed the previous two with 77% accuracy.

<p>Obstacles/Challenges (List the methodological obstacles if authors mentioned in the article)</p>	<p>Attention-based recurrent neural network A type of neural network that uses attention to focus on the most important parts of a sequence.</p> <p>Bangla Emote Module A module that detects the emotions lying behind emojis and emoticons.</p> <p>Development set A set of data used to tune the hyperparameters of a model.</p> <p>Test set A set of data used to evaluate the performance of a model on unseen data.</p>
<p>Terminology (List the common basic words frequently used in this research field)</p>	<p>Hate speech Speech that promotes or incites violence, discrimination, or hatred against individuals or groups based on attributes such as race, religion, ethnicity, etc.</p> <p>BERT Bidirectional Encoder Representations from Transformers.</p> <p>GRU Gated Recurrent Unit. A type of neural network architecture suited for sequential data processing.</p> <p>Softmax activation function A function used to convert the output of a neural network into a probability distribution.</p> <p>Precision A metric that measures the accuracy of positive predictions, indicating the proportion of correctly predicted positive samples.</p> <p>Recall A metric that measures the ability of a model to identify all relevant instances in the dataset, indicating the proportion of true positive samples.</p> <p>F1-Score A metric that combines precision and recall to provide a balance between them, often used when dealing with imbalanced datasets.</p>
<p>My Thinking/ Improvement idea</p>	<ol style="list-style-type: none"> 1. Use of multimodal data, such as images and videos, in addition to text data. 2. Model on other languages. Can the given model perform on other languages, such as Hindi, Urdu, or English? 3. GNN (Graph Neural Network): Used to model the relationships between the comments. GNNs can be used to learn from graph-structured data. 4. PLM (Pre-trained Language Model): A type of model that is trained on a large corpus of text and can be fine-tuned for specific natural language processing tasks.

Aspects	Paper # 9 (Title)
Title / Question (What is problem statement?)	Imbalanced Toxic Comments Classification using Data Augmentation and Deep Learning Classification of toxic comments in online user-generated content. The problem further entails dealing with a highly imbalanced dataset where a large majority of comments are non-toxic, making it challenging to build an effective classification model.
Objectives / Goal (What is looking for?)	Objective of the research is to develop a multi-label classification scheme capable of detecting different types of toxicity in online comments. Specifically, the authors aim to create a model that can accurately classify comments as toxic or non-toxic and, for toxic comments, identify the specific types of toxicity they contain.
Methodology / Theory (How to find the solution?)	<p>The authors use deep learning, specifically an ensemble model, to classify online comments into toxic and non-toxic categories, and further categorize toxic comments into different types of toxicity.</p> <p>Ensemble Model: The ensemble model combines three neural network types: CNN (Convolutional Neural Network), LSTM (Long Short-Term Memory), and GRU (Gated Recurrent Unit).</p> <p>Classification Steps:</p> <ol style="list-style-type: none"> 1. Toxicity Determination: The model first decides if a comment is toxic or non-toxic. 2. Type of Toxicity Identification: If a comment is toxic, the model identifies the specific type of toxicity. <p>Addressing Imbalance: The dataset is imbalanced, with more non-toxic comments. To address this, data augmentation techniques are used:</p> <ul style="list-style-type: none"> • Unique Words Augmentation • Random Masking • Synonyms Replacement
Software Tools (What program/software is used for design, coding and simulation?)	<p>The software tools used in the paper are:</p> <ul style="list-style-type: none"> • Python • TensorFlow • Keras
Test / Experiment How to test and characterize the design/prototype?	<p>The models were trained and evaluated on the Wikipedia talk edits dataset. The dataset was split into training, validation, and test sets with a ratio of 80:10:10.</p> <p>The models were evaluated using the following metrics:</p> <ul style="list-style-type: none"> • F1-score • ROC AUC

<p>Simulation/Test Data (What parameters are determined?)</p>	<p>The following parameters were determined:</p> <ul style="list-style-type: none"> • Number of epochs • Learning rate • Optimizer • Loss function <p>It discusses the choice of vocabulary size ($V = 50,000$ words) and the fixed comment length ($N = 150$ words) for preprocessing. Additionally, it references the use of Fast-Text word embedding ($D = 300$) for representing words.</p>
<p>Result / Conclusion (What was the final result?)</p>	<p>The proposed ensemble model outperformed all other methods on both the toxic/nontoxic classification and toxicity types prediction tasks. It achieved an F1-score of 0.828 for toxic/nontoxic classification and 0.872 for toxicity types prediction.</p>
<p>Terminology (List the common basic words frequently used in this research field)</p>	<p>Toxic comments: Comments that contain harmful or offensive content.</p> <p>Data augmentation: Techniques used to increase the diversity of training data by generating new data samples based on existing ones.</p> <p>Deep learning: A subset of machine learning that uses artificial neural networks to model and solve complex problems.</p> <p>Convolutional Neural Network (CNN): A type of neural network that is particularly effective for image and sequence data.</p> <p>Bidirectional Long Short-Term Memory (LSTM): A type of recurrent neural network that processes sequences in both forward and backward directions.</p> <p>Bidirectional Gated Recurrent Unit (GRU): A type of recurrent neural network similar to LSTM, designed for sequential data processing.</p> <p>F1-score: A metric that combines precision and recall to provide a balance between them.</p> <p>ROC AUC: Receiver Operating Characteristic Area Under the Curve, a metric for measuring the quality of a binary classification model.</p>
<p>My Thinking/ Improvement idea</p>	<p>Attention Mechanisms: Mechanisms used in neural networks to focus on specific parts of input data when making predictions.</p> <p>BERT-Based Models: Models built upon the BERT (Bidirectional Encoder Representations from Transformers) architecture, which is a pre-trained language model.</p> <p>Advanced Data Augmentation: More sophisticated techniques for generating additional training data to improve model performance.</p>

Aspects	Paper # 10 (Title)
Title / Question (What is problem statement?)	Bangla Toxic Comment Classification (Machine Learning and Deep Learning Approach) The paper is to address the issue of toxic comments in on-line conversations in the Bangla language and develop an effective classification system for detecting various forms of toxicity within these comments.
Objectives / Goal (What is looking for?)	The primary objective of the research is to develop a classification system for Bangla toxic comments. The goal is to classify comments into different categories of toxicity, including toxic, severe toxic, obscene, threat, insult, and identity hate.
Methodology / Theory (How to find the solution?)	<p>Classification Methods:</p> <ul style="list-style-type: none"> • Binary Relevance: This method involves treating each label separately as a binary classification problem. • Support Vector Machine (SVM): It's a type of algorithm that finds the best way to separate different types of comments. • Gaussian Naive Bayes: This method uses probabilities to classify comments. • Classifier Chain: It transforms the problem into multiple single-label classification tasks. • Label Powerset: This method also transforms the problem into a multi-class problem. • MLkNN (Multi-Label k-Nearest Neighbor): It's a variation of the k-Nearest Neighbor method adapted for multi-label classification. • BP-MLL (Backpropagation for Multi-Label Learning): This is a neural network-based approach for multi-label classification. <p>Data Preprocessing: Before using these methods, the researchers prepared the data by:</p> <ul style="list-style-type: none"> • Removing Punctuation: They got rid of punctuation marks like periods and commas. • Tokenizing Text: This means splitting comments into individual words or tokens. • Eliminating Stop Words: They removed common words in Bangla that don't carry much meaning, like "and" or "the." <p>Transforming Multi-Label to Single-Label: To make it easier for these methods to work, they transformed the problem of classifying multiple labels into simpler single-label problems.</p>

Software Tools (What program/software is used for design, coding and simulation?)	. They used Python and Scikit-learn machine learning library to develop their models. They also use the Tensor-Flow library to train and deploy their NN models.
Test / Experiment How to test and characterize the design/prototype?	<ul style="list-style-type: none"> • The research involves intensive experimentation with different classification methods. • The authors evaluate the performance of each method using accuracy metrics and hamming loss. • They also visualize the results to compare the performance of different classifiers, including accuracy, precision, recall, and F1-score, to measure their performance.
Simulation/Test Data (What parameters are determined?)	The researchers use a variety of parameters to determine the performance of their models. These parameters include the accuracy, precision, recall, and F1 score. The researchers also use confusion matrices to visualize the performance of their models on different types of toxic comments.
Result / Conclusion (What was the final result?)	The researchers find that the best-performing model is a BP-MLL Neural Network. This model achieves an accuracy of 60.00% on the test set. Additionally, the BP-MLL Neural Network has the lowest hamming loss and log-loss of all the models they evaluated.

Aspects	Paper # 11 (Title)
Title / Question (What is problem statement?)	Bengali Hate Speech Detection with BERT and Deep Learning Models
Objectives / Goal (What is looking for?)	<ul style="list-style-type: none"> • The research paper aims to address the issue of hate speech detection in the Bengali language on social media platforms by compiling a new dataset and evaluating the performance of different deep learning and transformer models. • The goal is to contribute to the study of identifying hate speech in Bengali and create a safer online environment by accurately identifying and removing hateful content from online platforms.
Methodology / Theory (How to find the solution?)	<ul style="list-style-type: none"> • The authors compiled a new dataset of 8,600 user comments from Facebook and YouTube, categorized into five groups: sports, religion, politics, entertainment, and others. • Five distinct models were used to study abusive language in Bengali, and the BERT model achieved the highest accuracy of 80%. • The same models were also tested on an existing dataset of 30,000 records, achieving an accuracy of 97%. • The models underwent training, validation, and testing using distinct sets of data. Statistical measures such as accuracy, recall, precision, F1-score, and confusion matrix were used to evaluate the models' performance. • The performance of the models was assessed using a confusion matrix, which provides a graphical summary of the model's results. The confusion matrix includes metrics like accuracy, recall, precision, and F1-score.
Software Tools (What program/software is used for design, coding and simulation?)	Implementation work was carried out at Intel(R) Core (TM) i7 CPU M60 @ 2.80 GHz in Python.
Test / Experiment How to test and characterize the design/prototype?	This research paper tests and compares models for detecting hate speech in Bengali. Using 8,600 user comments from Facebook and YouTube, the models were trained, validated, and tested. The BERT model had the highest accuracy of 80%. They were also evaluated using an existing dataset of 30,000 records and achieved 97% accuracy. The architecture and performance of the BERT model were analyzed.

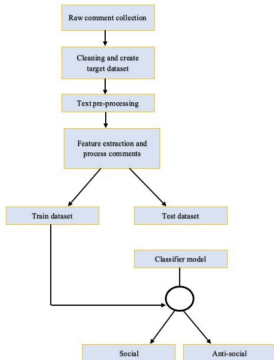
Simulation/Test Data (What parameters are determined?)	<p>Dataset Example:</p> <p>Before preprocessing</p> <p>“শুয়োরের বাচ্চারা তোদের খাইয়া কাজ নাই। এ দেশের যুবসমাজকে কোনদিকে নিয়ে যাচ্চিস। তোদের কোন দায়িত্ব নাই? পরিমনির হল কার্তিক মাসের কুন্তি আর এর পিছনে তোরা সবকটা কুন্ডা।🙄🙄🙄”</p> <p>After preprocessing</p> <p>“শুয়োরের বাচ্চারা তোদের খাইয়া কাজ নাই। এ দেশের যুবসমাজকে কোনদিকে নিয়ে যাচ্চিস। তোদের কোন দায়িত্ব নাই পরিমনির হল কার্তিক মাসের কুন্তি আর এর পিছনে তোরা সবকটা কুন্ডা”</p> <p>Fig. 5: Data before and after preprocessing</p> <div><p>before tokenization</p><div>“কুন্ডায় ও খাইবেনা ওর পঁচা দেহ”</div><p>after tokenization</p><p>[‘কুন্ডায়’, ‘ও’, ‘খাইবেনা’, ‘ওর’, ‘পঁচা’, ‘দেহ’]</p></div> <p>Fig. 6: Sentence before and after tokenization.</p>																																																								
Result / Conclusion (What was the final result?)	<p>They report the precision (Pabus), recall (Rabus), and F1 scores (F1abus) and etc. of various classifiers rates given</p> <p>Table 3: Summary of several model’s performances on our dataset</p> <table><tr><th>Model</th><th>Label</th><th>Precision</th><th>Recall</th><th>F1-score</th><th>Accuracy</th></tr><tr><td rowspan="2">CNN</td><td>NHS (class 0)</td><td>78</td><td>79</td><td>79</td><td rowspan="2">77</td></tr><tr><td>HS (class 1)</td><td>75</td><td>73</td><td>74</td></tr><tr><td rowspan="2">LSTM</td><td>NHS (class 0)</td><td>79</td><td>79</td><td>79</td><td rowspan="2">77</td></tr><tr><td>HS (class 1)</td><td>75</td><td>75</td><td>75</td></tr><tr><td rowspan="2">Bi-LSTM</td><td>NHS (class 0)</td><td>79</td><td>78</td><td>78</td><td rowspan="2">77</td></tr><tr><td>HS (class 1)</td><td>75</td><td>75</td><td>75</td></tr><tr><td rowspan="2">GRU</td><td>NHS (class 0)</td><td>78</td><td>76</td><td>77</td><td rowspan="2">78</td></tr><tr><td>HS (class 1)</td><td>73</td><td>75</td><td>74</td></tr><tr><td rowspan="2">BERT</td><td>NHS (class 0)</td><td>82</td><td>82</td><td>82</td><td rowspan="2">80</td></tr><tr><td>HS (class 1)</td><td>79</td><td>79</td><td>79</td></tr></table>	Model	Label	Precision	Recall	F1-score	Accuracy	CNN	NHS (class 0)	78	79	79	77	HS (class 1)	75	73	74	LSTM	NHS (class 0)	79	79	79	77	HS (class 1)	75	75	75	Bi-LSTM	NHS (class 0)	79	78	78	77	HS (class 1)	75	75	75	GRU	NHS (class 0)	78	76	77	78	HS (class 1)	73	75	74	BERT	NHS (class 0)	82	82	82	80	HS (class 1)	79	79	79
Model	Label	Precision	Recall	F1-score	Accuracy																																																				
CNN	NHS (class 0)	78	79	79	77																																																				
	HS (class 1)	75	73	74																																																					
LSTM	NHS (class 0)	79	79	79	77																																																				
	HS (class 1)	75	75	75																																																					
Bi-LSTM	NHS (class 0)	79	78	78	77																																																				
	HS (class 1)	75	75	75																																																					
GRU	NHS (class 0)	78	76	77	78																																																				
	HS (class 1)	73	75	74																																																					
BERT	NHS (class 0)	82	82	82	80																																																				
	HS (class 1)	79	79	79																																																					
Obstacles/Challenges (List the methodological obstacles if authors mentioned in the article)	<ul style="list-style-type: none">• The lack of available public Bengali datasets for hate speech detection on social media sites like Facebook and YouTube hindered the research process.• Existing online datasets for Bengali hate speech detection were sparse, poorly sequenced, and lacked necessary data types .• Limited studies on identifying hate speech in languages other than English, including Bengali, posed a challenge in understanding and addressing the issue.																																																								
Terminology (List the common basic words frequently used in this research field)	Hate speech, Abusive language, Bengali hate speech detection, Deep learning models, BERT model.																																																								

Aspects	Paper # 12 (Title)																					
Title / Question (What is problem statement?)	Abusive content detection in transliterated Bengali-English social media corpus																					
Objectives / Goal (What is looking for?)	This research paper introduces an annotated corpus of 3,000 transliterated Bengali comments to detect abusive content on social media. The authors use supervised machine learning and deep learning-based classifiers for evaluations and compare the performance of different classifiers such as LR, SVM, RF, and BiLSTM. The goal is to provide a publicly available corpus to aid in detecting abusive content in Bengali social media.																					
Methodology / Theory (How to find the solution?)	<ul style="list-style-type: none">• The research paper employs several supervised machine learning (ML) and deep learning-based classifiers, including logistic regression (LR), support vector machine (SVM), random forest (RF), and bidirectional long short-term memory (BiLSTM) architecture, for identifying abusive content in transliterated Bengali comments in social media.• The Authors use TF-IDF to calculate the importance of words in a document. They use ML classifiers with default settings from the scikit-learn library.																					
Software Tools (What program/software is used for design, coding and simulation?)	The authors utilize the scikit-learn library for implementing the ML classifiers with default parameter settings using Python.																					
Test / Experiment How to test and characterize the design/prototype?	Researchers randomly selected 100 abusive and nonabusive comments and manually examined them for English and transliterated Bengali words. They found that 76% of offensive words and almost 80% of non-abusive words were transliterated Bengali. The final corpus had no Bengali words due to filtering.																					
Simulation/Test Data (What parameters are determined?)	<div>Dataset Example:</div> <table><thead><tr><th>Transliterated Bengali Comment</th><th>English Translation</th><th>Class</th></tr></thead><tbody><tr><td>1. amar mathay dhorena somoy tv kivabe a khankire office aneche</td><td>I don't understand why shomoy TV brought this slut</td><td>Abusive</td></tr><tr><td>2. Really onk valolaglo vaia Apnr question gulo khubbi mojar silo</td><td>Really liked it a lot bro, your questions were very funny</td><td>Non-abusive</td></tr><tr><td>3. Magi tore to amio chudmo na</td><td>Whore not even I will fuck you</td><td>Abusive</td></tr><tr><td>4. Joy, tumar show r dekbona</td><td>Joy, I won't watch your show again</td><td>Non-abusive</td></tr><tr><td>5. Sobay to tor Moto khanki magi na tor family o khanki.</td><td>Not everyone is slut like you. Your family is slut too.</td><td>Abusive</td></tr><tr><td>6. Bro please tader k interview te ane highlights na kora tai valo</td><td>Bro Please don't highlight them in your interview</td><td>Non-abusive</td></tr></tbody></table> <div>Figure 1: Examples of annotated abusive and non-abusive reviews</div>	Transliterated Bengali Comment	English Translation	Class	1. amar mathay dhorena somoy tv kivabe a khankire office aneche	I don't understand why shomoy TV brought this slut	Abusive	2. Really onk valolaglo vaia Apnr question gulo khubbi mojar silo	Really liked it a lot bro, your questions were very funny	Non-abusive	3. Magi tore to amio chudmo na	Whore not even I will fuck you	Abusive	4. Joy, tumar show r dekbona	Joy, I won't watch your show again	Non-abusive	5. Sobay to tor Moto khanki magi na tor family o khanki.	Not everyone is slut like you. Your family is slut too.	Abusive	6. Bro please tader k interview te ane highlights na kora tai valo	Bro Please don't highlight them in your interview	Non-abusive
Transliterated Bengali Comment	English Translation	Class																				
1. amar mathay dhorena somoy tv kivabe a khankire office aneche	I don't understand why shomoy TV brought this slut	Abusive																				
2. Really onk valolaglo vaia Apnr question gulo khubbi mojar silo	Really liked it a lot bro, your questions were very funny	Non-abusive																				
3. Magi tore to amio chudmo na	Whore not even I will fuck you	Abusive																				
4. Joy, tumar show r dekbona	Joy, I won't watch your show again	Non-abusive																				
5. Sobay to tor Moto khanki magi na tor family o khanki.	Not everyone is slut like you. Your family is slut too.	Abusive																				
6. Bro please tader k interview te ane highlights na kora tai valo	Bro Please don't highlight them in your interview	Non-abusive																				
Result / Conclusion (What was the final result?)	<div>They report the precision (P_{abus}), recall (R_{abus}), and F1 scores ($F1_{abus}$) of various classifiers for identifying abusive comments, rates given below-</div> <table><thead><tr><th>Classifier</th><th>R_{abus}</th><th>P_{abus}</th><th>$F1_{abus}$</th></tr></thead><tbody><tr><td>SVM</td><td>0.790 ± 0.008</td><td>0.865 ± 0.015</td><td>0.827 ± 0.010</td></tr><tr><td>LR</td><td>0.779 ± 0.006</td><td>0.876 ± 0.004</td><td>0.823 ± 0.006</td></tr><tr><td>BiLSTM</td><td>0.781 ± 0.031</td><td>0.800 ± 0.036</td><td>0.790 ± 0.031</td></tr><tr><td>RF</td><td>0.781 ± 0.013</td><td>0.762 ± 0.028</td><td>0.770 ± 0.020</td></tr></tbody></table>	Classifier	R_{abus}	P_{abus}	$F1_{abus}$	SVM	0.790 ± 0.008	0.865 ± 0.015	0.827 ± 0.010	LR	0.779 ± 0.006	0.876 ± 0.004	0.823 ± 0.006	BiLSTM	0.781 ± 0.031	0.800 ± 0.036	0.790 ± 0.031	RF	0.781 ± 0.013	0.762 ± 0.028	0.770 ± 0.020	
Classifier	R_{abus}	P_{abus}	$F1_{abus}$																			
SVM	0.790 ± 0.008	0.865 ± 0.015	0.827 ± 0.010																			
LR	0.779 ± 0.006	0.876 ± 0.004	0.823 ± 0.006																			
BiLSTM	0.781 ± 0.031	0.800 ± 0.036	0.790 ± 0.031																			
RF	0.781 ± 0.013	0.762 ± 0.028	0.770 ± 0.020																			

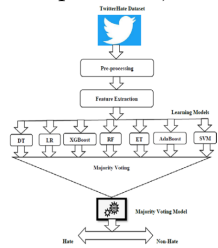
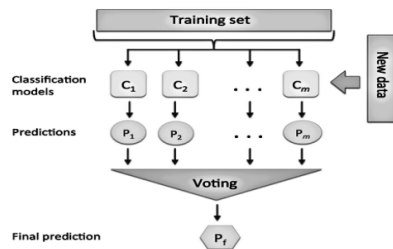
<p>Obstacles/Challenges (List the methodological obstacles if authors mentioned in the article)</p>	<ul style="list-style-type: none"> • Inadequacy of resources and tools for abusive text detection in low-resource languages like Bengali. • Difficulty in capturing transliterated Bengali comments using monolingual approaches. • Lack of publicly available transliterated Bengali corpus for abusive content analysis. • Missing annotated training data for transliterated Bengali text, which is required for supervised machine learning (ML) classifiers. • The need for manual annotation of a large corpus of transliterated Bengali comments for abusive content detection. • Inter-annotator agreement between native Bengali speakers for assigning the class of transliterated comments into abusive and non-abusive categories.
<p>Terminology (List the common basic words frequently used in this research field)</p>	<p>Transliterated Bengali, Social media, Corpus, Machine learning (ML), Deep learning, Classifiers, Support vector machine (SVM), Supervised learning, Annotated corpus, Low-resource languages, Code-mixing, Sentiment scores, Offensive words, Threatening words, Natural language processing (NLP), classifier, KNN-K-Nearest Neighbor. RF- Random Forest, SVM- Support Vector Machine</p>

Aspects	Paper # 13 (Title)
Title / Question (What is problem statement?)	Comparative analysis of deep learning based Afaan Oromo hate speech detection
Objectives / Goal (What is looking for?)	<p>This research paper examines the use of deep learning models for identifying hate speech in Afaan Oromo. The study collects and annotates a large dataset and compares the performance of different models, finding that a CNN-BiLSTM model performs the best with an F1-score of 87%. The paper also discusses the impact of pre-trained word embeddings and data augmentation on classification performance. The research aims to contribute to the development of hate speech detection in Afaan Oromo using machine learning.</p>
Methodology / Theory (How to find the solution?)	<ul style="list-style-type: none"> • The research paper focuses on the comparative analysis of deep learning models for Afaan Oromo hate speech detection. • The paper collects and annotates a large dataset of hate speech in Afaan Oromo, which is then used to evaluate the performance of different deep-learning models, including CNN, BiLSTM, LSTM, GRU, and CNN-LSTM. • The results show that the model based on CNN and BiLSTM achieves the best performance, with an average F1-score of 87%. • The paper also discusses the impact of pre-trained word embeddings and data augmentation on classification performance. Training the model with embedded representation slightly increases the classification performance by 1.5 on average. • The experiments conducted in the study involve different circumstances, such as pre-trained word embeddings, training word embeddings with the model, and data augmentation. The performance of each model is evaluated using precision, recall, and F1-score metrics. • The research aims to contribute to the development of machine learning models for detecting hate speech in Afaan Oromo.
Software Tools (What program/software is used for design, coding and simulation?)	<p>Deep learning frameworks such as TensorFlow, PyTorch, or Keras are commonly used for implementing and training deep learning models. Natural language processing libraries like NLTK (Natural Language Toolkit) or spaCy and Python programming language.</p>

<p>Test / Experiment How to test and characterize the design/prototype?</p>	<p>Each deep learning model investigated in the study was trained independently using a set for parameter optimization and a development set for validation purposes. The performance of the models was evaluated using metrics such as precision, recall, and F1-score. The results of the experiments were presented in separate tables, comparing the performance of different models under different circumstances, such as pre-trained word embeddings and data augmentation</p>
<p>Simulation/Test Data (What parameters are determined?)</p>	<p>The paper mentions that the researchers collected and annotated the biggest dataset of hate speech in Afaan Oromo, indicating that they had their own dataset for training and testing the models.</p>
<p>Result / Conclusion (What was the final result?)</p>	<p>The results of the experiments were presented in separate tables, comparing the performance of different models under different circumstances, such as pre-trained word embeddings and data augmentation. The best-performing model based on CNN and BiLSTM achieved an average F1-score of 87%</p>
<p>Obstacles/Challenges (List the methodological obstacles if authors mentioned in the article)</p>	<ul style="list-style-type: none"> • Lack of available resources and research in this area. • The paper highlights the challenge of collecting and annotating a sufficient dataset of hate speech in Afaan Oromo, which required the involvement of language experts. • The researchers faced the challenge of evaluating and comparing different deep learning models, including CNN, LSTMs, BiLSTMs, LSTM, GRU, and CNN-LSTM, to identify the most effective approach for Afaan Oromo hate speech recognition. • The lack of previous research and established methodologies for hate speech detection in Afaan Oromo could have posed challenges in terms of benchmarking and comparing the results of this study with existing literature.
<p>Terminology (List the common basic words frequently used in this research field)</p>	<p>Social media platforms, Machine learning models, Hate speech, Deep learning models, Afaan Oromo</p>

Aspects	Paper # 16
Title / Question (What is problem statement?)	A Machine Learning Approach to Classify Anti-social Bengali Comments on Social Media
Objectives / Goal (What is looking for?)	The research aims to prevent anti-social activities in the Bangla community by studying and classifying Bengali comments on social media.
Methodology / Theory (How to find the solution?)	<p>The work was divided into some phases.</p> <ul style="list-style-type: none"> • The researchers collected 2000 Bengali comments from Facebook and YouTube, did data cleaning and preprocessing, • They utilized various machine learning models, including Gated Recurrent Unit (GRU), Logistic Regression (LR), Random Forest (RF), Multinomial Naive Bayes (MNB), and Support Vector Machine (SVM), to classify the comments as anti-social or socially acceptable • The researchers used TF-IDF (Term Frequency-Inverse Document Frequency) to evaluate the significance of words in the comments and model N-grams, and • Supervised classifiers were applied, and the performance of different models, such as LR, RF, MNB, Linear SVM, and GRU, was evaluated based on accuracy, recall, and precision.
Software Tools (What program/software is used for design, coding and simulation?)	Implementation work was carried out at Intel(R) Core (TM) i7 CPU M60 @ 2.80 GHz in Python.
Test / Experiment How to test and characterize the design/prototype?	The dataset contains 2000 comments equally balanced between both classes, indicating that the dataset is evenly split between anti-social and socially acceptable comments . The train set and test set were respectively 90% and 10% of the dataset . Label encoding was applied to the train and test sets to convert them into binary values.
Simulation/Test Data (What parameters are determined?)	<p>The research paper determines several parameters related to the classification of anti-social Bengali comments on social media. The datasets was collected from social media platforms from the authors.</p>  <pre> graph TD A[Raw comment collection] --> B[Cleaning and create target dataset] B --> C[Text pre-processing] C --> D[Feature extraction and process comments] D --> E[Train dataset] D --> F[Test dataset] E --> G[Classifier model] F --> G G --> H[Social] G --> I[Anti-social] </pre> <p>Fig. 1. Flowchart of the anti-social comment detection model</p>

Result / Conclusion (What was the final result?)	The results showed that Multinomial Naive Bayes (MNB) had the highest accuracy of 80.51%, followed by GRU with an accuracy of 78.89% . LR, RF, and Linear SVM had accuracies of 74.36%, 71.28%, and 70.26% respectively . The performance metrics of MNB showed the precision, recall, and F1 score all were 81.55%,
Obstacles/Challenges (List the methodological obstacles if authors mentioned in the article)	The authors highlight the scarcity of requisite datasets for studying Bengali comments on social media, indicating a potential challenge in obtaining sufficient data for their research.
Terminology (List the common basic words frequently used in this research field)	Artificial intelligence. Machine learning Gated Recurrent Unit (GRU) Logistic Regression (LR) Random Forest (RF) Multinomial Naive Bayes (MNB) Support Vector Machine (SVM)

Aspects	Paper # 17
Title / Question (What is problem statement?)	Multi-Model Learning to Detect Twitter Hate Speech
Objectives / Goal (What is looking for?)	The paper proposes a multi-model learning approach for identifying hate speech and non-hate speech on Twitter and aims to achieve high detection results.
Methodology / Theory (How to find the solution?)	The work was divided into three phases. <ul style="list-style-type: none"> • The first phase was the collection of data, • The second phase was pre-processing and computation of data, and • Third phase was to acquire detection results, precision, recall, and f1-score.
Software Tools (What program/software is used for design, coding and simulation?)	Implementation work was carried out at Intel(R) Core (TM) i7 CPU M60 @ 2.80 GHz in Python.
Test / Experiment How to test and characterize the design/prototype?	They used machine learning classifiers such as Decision tree, Logistic regression, XGBoost, Random forest, Extra tree, AdaBoost, and Support vector machine for analyzing the dataset. The dataset was divided into an 80:20 ratio for training and testing the classifiers . Detected precision, recall, and f1-score using TF-IDF features.  <p>Figure 1. Framework of Proposed Twitter Hate Speech Detection System.</p>
Simulation/Test Data (What parameters are determined?)	Datasets was collected from - Kaggle repository, which contained 31,962 tweets categorized as binary hate or non-hate, to evaluate their multi-model learning strategy for detecting hate speech on Twitter. The dataset was significantly skewed, with 93 tweets (or 2,965 non-hate labeled Twitter data) and 7 tweets (or 2,240 hate-labeled Twitter data). The dataset was divided into an 80:20 ratio for training and testing the classifiers. 

<p>Result / Conclusion (What was the final result?)</p>	<p>The multi-model learning strategy for detecting hate speech on Twitter achieved detection</p> <ul style="list-style-type: none"> • Accuracy of 96.29%, • Precision of 96%, • Recall of 96%, and • F1-score of 96
<p>Obstacles/Challenges (List the methodological obstacles if authors mentioned in the article)</p>	<p>The article mentions that various approaches have been proposed by researchers for hate speech detection on different social networking platforms, indicating that the diversity and complexity of these platforms can be methodological obstacles in developing universal detection models.</p>
<p>Terminology (List the common basic words frequently used in this research field)</p>	<p>Multi-model learning. Machine learning (ML) · TF-IDF (Term Frequency-Inverse Document Frequency) Deep learning (DL) · Classifier performance XGBoost-Extreme Gradient Boosting RF-Random Forest SVM- Support Vector Machine</p>

Aspects	Paper # 18																															
Title / Question (What is problem statement?)	Hate Speech and Offensive Language Detection in Bengali																															
Objectives / Goal (What is looking for?)	The paper addresses the gap in research on hate speech detection in low-resource languages like Bengali and develops an annotated dataset of Bengali posts for classification of hateful content. The authors explore different models and techniques, including interlingual transfer mechanisms, and find that XLM-Roberta performs the best for training actual and Romanized datasets.																															
Methodology / Theory (How to find the solution?)	<p>The work was divided into four phases.</p> <ul style="list-style-type: none">• The first phase was the dataset collection and sampling,• The second phase was pre-processing and computation of data,• Third phase was experimenting the dataset with a wide range of models, and• The fourth and last phase was to evaluate their models in terms of accuracy, F1-score and AUROC score.																															
Software Tools (What program/software is used for design, coding and simulation?)	Implementation work was carried out at Intel(R) Core (TM) i7 CPU M60 @ 2.80 GHz in Python.																															
Test / Experiment How to test and characterize the design/prototype?	The models are trained, validated, and tested using a stratified split of the dataset, ensuring representation from each class across the splits. The models in the paper are trained, validated, and tested using a 70:10:20 train-validation-test split, stratified by class across the splits.																															
Simulation/Test Data (What parameters are determined?)	<p>The researchers collected the dataset by sampling Bengali (actual and Romanized) tweets from Twitter.</p> <table><thead><tr><th>Type</th><th>Tweet</th><th>Translation</th><th>Label</th><th>Target</th></tr></thead><tbody><tr><td rowspan="3">Actual</td><td>এই ব্যক্তি যে বর্তমানের বর্ণাশ্রম বৈজ্ঞানিক চিন্তা করেছেন তাহলে... কিন্তু তার মনে আছে যে ৩০০ বছর আগে 'The word 'colored'' should not be associated with the Bengali those who have been enslaved all their lives.</td><td>Do you understand today that this race is barbaric, stupid, mean, fanatical? But the British understood more than 300 years ago! The word 'colored' should not be associated with the Bengali those who have been enslaved all their lives.</td><td>Hate</td><td>Bengali</td></tr><tr><td>@user আমার ছোট্ট মেয়ে, আমার মেয়ে, আমার মেয়ে... @user I fuck you in dream, daughter of a bitch, this is why I get nightmare</td><td></td><td>Offensive</td><td>Individual, Woman</td></tr><tr><td>@user মাদারিষ্টা জেনে নিলু তার কুলায় পণ্ডিত কাদেরই https://bit.ly/3uqjzj1</td><td>@user Dalits are questioning the citizenship law https://bit.ly/3uqjzj1</td><td>Normal</td><td>Others</td></tr><tr><td rowspan="3">Romanized</td><td>@user @user 42 e 42 e ki holo re ganduchoda choti chota nicher koto ??? Tor baper gnare dhakke dho 42 is ?? Shanker poka... Kangle mal... Super jute... 🤔🤔🤔</td><td>@user @user What happened to him ass fucker, shoe ickes, circumcise man? Out of 42, 42 is your father's son... Son of a bitch... Kangle (derogatory term for Bangladesh)... Pig breed... 🤔🤔🤔</td><td>Hate</td><td>Bangladeshi</td></tr><tr><td>Shankar chete dwijen barik, kal tui usesh. kal tui isophane. kal ami tor bou ke chudbo. kochi maal. LENOVO THE LAORA.</td><td>Son of a bitch dwijen barik. Tomorrow you are finish. Tomorrow you will be in the crematorium. I will fuck your wife tomorrow. Young wife. LENOVO THE LAORA.</td><td>Offensive</td><td>Individual</td></tr><tr><td>@user He got best debutante wid #58617? 🤔 Then what abt his film #Paanchdhyay? Sala amra audience ra ki bokachoda? koto lobby cholebi!</td><td>@user He got best debutante wid #58617? 🤔 Then what about his film #Paanchdhyay? Damn, are we fucking dumb audiences? How much longer will the lobby last?!</td><td>Normal</td><td>Others</td></tr></tbody></table> <p>Table 2: Samples of Actual and Roman Bengali tweets for each label from the dataset</p>	Type	Tweet	Translation	Label	Target	Actual	এই ব্যক্তি যে বর্তমানের বর্ণাশ্রম বৈজ্ঞানিক চিন্তা করেছেন তাহলে... কিন্তু তার মনে আছে যে ৩০০ বছর আগে 'The word 'colored'' should not be associated with the Bengali those who have been enslaved all their lives.	Do you understand today that this race is barbaric, stupid, mean, fanatical? But the British understood more than 300 years ago! The word 'colored' should not be associated with the Bengali those who have been enslaved all their lives.	Hate	Bengali	@user আমার ছোট্ট মেয়ে, আমার মেয়ে, আমার মেয়ে... @user I fuck you in dream, daughter of a bitch, this is why I get nightmare		Offensive	Individual, Woman	@user মাদারিষ্টা জেনে নিলু তার কুলায় পণ্ডিত কাদেরই https://bit.ly/3uqjzj1	@user Dalits are questioning the citizenship law https://bit.ly/3uqjzj1	Normal	Others	Romanized	@user @user 42 e 42 e ki holo re ganduchoda choti chota nicher koto ??? Tor baper gnare dhakke dho 42 is ?? Shanker poka... Kangle mal... Super jute... 🤔🤔🤔	@user @user What happened to him ass fucker, shoe ickes, circumcise man? Out of 42, 42 is your father's son... Son of a bitch... Kangle (derogatory term for Bangladesh)... Pig breed... 🤔🤔🤔	Hate	Bangladeshi	Shankar chete dwijen barik, kal tui usesh. kal tui isophane. kal ami tor bou ke chudbo. kochi maal. LENOVO THE LAORA.	Son of a bitch dwijen barik. Tomorrow you are finish. Tomorrow you will be in the crematorium. I will fuck your wife tomorrow. Young wife. LENOVO THE LAORA.	Offensive	Individual	@user He got best debutante wid #58617? 🤔 Then what abt his film #Paanchdhyay? Sala amra audience ra ki bokachoda? koto lobby cholebi!	@user He got best debutante wid #58617? 🤔 Then what about his film #Paanchdhyay? Damn, are we fucking dumb audiences? How much longer will the lobby last?!	Normal	Others
Type	Tweet	Translation	Label	Target																												
Actual	এই ব্যক্তি যে বর্তমানের বর্ণাশ্রম বৈজ্ঞানিক চিন্তা করেছেন তাহলে... কিন্তু তার মনে আছে যে ৩০০ বছর আগে 'The word 'colored'' should not be associated with the Bengali those who have been enslaved all their lives.	Do you understand today that this race is barbaric, stupid, mean, fanatical? But the British understood more than 300 years ago! The word 'colored' should not be associated with the Bengali those who have been enslaved all their lives.	Hate	Bengali																												
	@user আমার ছোট্ট মেয়ে, আমার মেয়ে, আমার মেয়ে... @user I fuck you in dream, daughter of a bitch, this is why I get nightmare		Offensive	Individual, Woman																												
	@user মাদারিষ্টা জেনে নিলু তার কুলায় পণ্ডিত কাদেরই https://bit.ly/3uqjzj1	@user Dalits are questioning the citizenship law https://bit.ly/3uqjzj1	Normal	Others																												
Romanized	@user @user 42 e 42 e ki holo re ganduchoda choti chota nicher koto ??? Tor baper gnare dhakke dho 42 is ?? Shanker poka... Kangle mal... Super jute... 🤔🤔🤔	@user @user What happened to him ass fucker, shoe ickes, circumcise man? Out of 42, 42 is your father's son... Son of a bitch... Kangle (derogatory term for Bangladesh)... Pig breed... 🤔🤔🤔	Hate	Bangladeshi																												
	Shankar chete dwijen barik, kal tui usesh. kal tui isophane. kal ami tor bou ke chudbo. kochi maal. LENOVO THE LAORA.	Son of a bitch dwijen barik. Tomorrow you are finish. Tomorrow you will be in the crematorium. I will fuck your wife tomorrow. Young wife. LENOVO THE LAORA.	Offensive	Individual																												
	@user He got best debutante wid #58617? 🤔 Then what abt his film #Paanchdhyay? Sala amra audience ra ki bokachoda? koto lobby cholebi!	@user He got best debutante wid #58617? 🤔 Then what about his film #Paanchdhyay? Damn, are we fucking dumb audiences? How much longer will the lobby last?!	Normal	Others																												
Result / Conclusion (What was the final result?)	<p>The MuRIL model performed the best in the study on hate speech and offensive language detection in Bengali.</p> <ul style="list-style-type: none">• When fine-tuning the model with 32 instances, m-BERT initially performed the best, but as the number of instances increased, MuRIL consistently outperformed all other models,• MuRIL achieved a macro F1-Score of 0.751 with 128 instances per label, indicating its strong performance in detecting hateful content.,																															

Obstacles/Challenges (List the methodological obstacles if authors mentioned in the article)	Communication through emojis and the presence of sarcastic or ambiguous content can lead to misclassification of posts, highlighting the challenge of accurately interpreting the intent behind certain expressions.
Terminology (List the common basic words frequently used in this research field)	Machine learning · Natural Language Processing (NLP) · Hate speech · Social media Performance Matrix MuRIL-Multilingual Representations for Indian Languages.

Aspects	Paper # 19																																			
Title / Question (What is problem statement?)	BD-SHS: A Benchmark Dataset for Learning to Detect On-line Bangla Hate Speech in Different Social Contexts																																			
Objectives / Goal (What is looking for?)	The goal of the paper is to address the problem of hate speech (HS) detection in the context of Bangla language on online social networking sites and streaming services.																																			
Methodology / Theory (How to find the solution?)	Phases followed in this paper for the whole process: <ul style="list-style-type: none">• Identification of shortcomings in existing Bangla hate speech datasets,• Creation of a large manually labeled dataset called BD-SHS,• Splitting the dataset,• Experimentation with various models and linguistic features,• Training classifier models with different combinations of features• Evaluation of the performance of the models																																			
Software Tools (What program/software is used for design, coding and simulation?)	Implementation work was carried out at Intel(R) Core (TM) i7 CPU M60 @ 2.80 GHz in Python.																																			
Test / Experiment How to test and characterize the design/prototype?	Splitting the dataset into train (70%), validation (15%), and test (15%) sets after performing a random shuffle. . Accuracy, specificity. Experimentation with various models and linguistic features to develop benchmark results for three classification tasks: HS Identification (Binary classification), Identify the Target of HS (Multi-label classification), and Categorization of HS Types (Multi-label classification).																																			
Simulation/Test Data (What parameters are determined?)	<p>The paper used a hierarchical annotation process to label the BD-SHS dataset, which is the first of its kind in Bangla hate speech detection .</p> <table><tr><th>#</th><th>Comment</th><th>HS</th><th>Target</th><th>Type</th></tr><tr><td>1</td><td>কি বালের মুভি What an ass movie.</td><td>NH</td><td>-</td><td>-</td></tr><tr><td>2</td><td>বোরকা পরে না, ধর্ষিত তো হবেই। It is obvious that girls who do not wear burqas would be raped.</td><td>HS</td><td>female</td><td>gender</td></tr><tr><td>3</td><td>তর কাজ দেখে বমি করে দিতে ইচ্ছে করছে। Seeing your work makes me want to puke.”</td><td>HS</td><td>IND</td><td>slander</td></tr><tr><td>4</td><td>দাড়ি ওয়ালা তো এক নাম্বার বাইনচোত আর বাকি গুলা রে জুতা মারা দরকার That bearded guy is number one asshole and the rest should be beaten with shoes.</td><td>HS</td><td>male, group</td><td>slander, CV</td></tr><tr><td>5</td><td>তুই কুত্তার বচ্চা। তোর মত মালান্ডিনের দেশে থাকার কোন অধিকার নাই। তোরে আমি পিটায়ে মেরে ফেললাম। You son of a bitch. A malaun (swear for Hindu) like you has no right to live in our country. I will beat you to death.</td><td>HS</td><td>IND, group</td><td>slander, religion, CV</td></tr><tr><td>6</td><td>অডমিন সালা তুমি কই সালা পাগল কই থেকে আসে Admin, where are you bastard? Where does these kinds of lunatics come from?</td><td>HS</td><td>IND</td><td>slander</td></tr></table> <p>Table 2: Representative snapshot of our dataset</p>	#	Comment	HS	Target	Type	1	কি বালের মুভি What an ass movie.	NH	-	-	2	বোরকা পরে না, ধর্ষিত তো হবেই। It is obvious that girls who do not wear burqas would be raped.	HS	female	gender	3	তর কাজ দেখে বমি করে দিতে ইচ্ছে করছে। Seeing your work makes me want to puke.”	HS	IND	slander	4	দাড়ি ওয়ালা তো এক নাম্বার বাইনচোত আর বাকি গুলা রে জুতা মারা দরকার That bearded guy is number one asshole and the rest should be beaten with shoes.	HS	male, group	slander, CV	5	তুই কুত্তার বচ্চা। তোর মত মালান্ডিনের দেশে থাকার কোন অধিকার নাই। তোরে আমি পিটায়ে মেরে ফেললাম। You son of a bitch. A malaun (swear for Hindu) like you has no right to live in our country. I will beat you to death.	HS	IND, group	slander, religion, CV	6	অডমিন সালা তুমি কই সালা পাগল কই থেকে আসে Admin, where are you bastard? Where does these kinds of lunatics come from?	HS	IND	slander
#	Comment	HS	Target	Type																																
1	কি বালের মুভি What an ass movie.	NH	-	-																																
2	বোরকা পরে না, ধর্ষিত তো হবেই। It is obvious that girls who do not wear burqas would be raped.	HS	female	gender																																
3	তর কাজ দেখে বমি করে দিতে ইচ্ছে করছে। Seeing your work makes me want to puke.”	HS	IND	slander																																
4	দাড়ি ওয়ালা তো এক নাম্বার বাইনচোত আর বাকি গুলা রে জুতা মারা দরকার That bearded guy is number one asshole and the rest should be beaten with shoes.	HS	male, group	slander, CV																																
5	তুই কুত্তার বচ্চা। তোর মত মালান্ডিনের দেশে থাকার কোন অধিকার নাই। তোরে আমি পিটায়ে মেরে ফেললাম। You son of a bitch. A malaun (swear for Hindu) like you has no right to live in our country. I will beat you to death.	HS	IND, group	slander, religion, CV																																
6	অডমিন সালা তুমি কই সালা পাগল কই থেকে আসে Admin, where are you bastard? Where does these kinds of lunatics come from?	HS	IND	slander																																

<p>Result / Conclusion (What was the final result?)</p>	<p>The paper experimented with various models and linguistic features, including linear Support Vector Machine (SVM) and Bidirectional Long Short Term Memory (Bi-LSTM) architectures. BiLSTM trained with informal embeddings from IFT achieved performed the best with</p> <ul style="list-style-type: none"> • Accuracy of 85.08%, • Precision of 91.0%, • Recall of 91.0%, and • F1 score of 91.0%
<p>Obstacles/Challenges (List the methodological obstacles if authors mentioned in the article)</p>	<p>The lack of formal language syntax, spelling mistakes, and the use of various swear words and non-standard acronyms in the comment sections of social media and online streaming sites made the task of hate speech detection harder.</p>
<p>Terminology (List the common basic words frequently used in this research field)</p>	<p>HS-Hate Speech. Offensive comments Linguistic diversity NLP-Natural Language Processing models Bidirectional Long Short Term Memory (Bi-LSTM) IFT-InFormal Text MFT-Multilingual FastText</p>

Aspects	Paper # 20																																								
Title / Question (What is problem statement?)	Social Media Hate Speech Detection Using Explainable Artificial Intelligence (XAI)																																								
Objectives / Goal (What is looking for?)	The goal of the research study was to apply explainable artificial intelligence (XAI) characteristics to detect hate speech in social media using deep learning models.																																								
Methodology / Theory (How to find the solution?)	<p>The work was divided into multiple phases.</p> <ul style="list-style-type: none">• Two datasets were used in the study,• Data pre-processing was performed on the datasets,• Exploratory data analysis to uncover patterns and insights,• Experimentation with various models, and• Evaluating the performance of the models.																																								
Software Tools (What program/software is used for design, coding and simulation?)	Implementation work was carried out at Intel(R) Core (TM) i7 CPU M60 @ 2.80 GHz in Python.																																								
Test / Experiment How to test and characterize the design/prototype?	For the experimental work, the datasets were divided into the ratio of 80used to train classification algorithms, and the remaining 20% used for testing purposes. Accuracy, specificity, sensitivity, and area under the curve were evaluated for the seven classifiers.																																								
Simulation/Test Data (What parameters are determined?)	<p>Two datasets were used in the study: the Jigsaw dataset and the HateXplain dataset from Kaggle.</p> <p>Table 2. Google Jigsaw dataset details.</p> <table><tr><th>Classification</th><th>Frequency</th></tr><tr><td>Clean</td><td>201,081</td></tr><tr><td>Toxic</td><td>21,384</td></tr><tr><td>Obscene</td><td>12,140</td></tr><tr><td>Insult</td><td>11,304</td></tr><tr><td>Identity hate</td><td>2117</td></tr><tr><td>Severe toxic</td><td>1962</td></tr><tr><td>Threat</td><td>689</td></tr></table> <p>Table 3. HateXplain dataset details.</p> <table><tr><th></th><th>Twitter</th><th>Gab</th><th>Total</th></tr><tr><td>Hateful</td><td>708</td><td>5227</td><td>5935</td></tr><tr><td>Offensive</td><td>2328</td><td>3152</td><td>5480</td></tr><tr><td>Normal</td><td>5770</td><td>2044</td><td>7814</td></tr><tr><td>Undecided</td><td>249</td><td>670</td><td>919</td></tr><tr><td>Total</td><td>9055</td><td>11,093</td><td>20,148</td></tr></table>	Classification	Frequency	Clean	201,081	Toxic	21,384	Obscene	12,140	Insult	11,304	Identity hate	2117	Severe toxic	1962	Threat	689		Twitter	Gab	Total	Hateful	708	5227	5935	Offensive	2328	3152	5480	Normal	5770	2044	7814	Undecided	249	670	919	Total	9055	11,093	20,148
Classification	Frequency																																								
Clean	201,081																																								
Toxic	21,384																																								
Obscene	12,140																																								
Insult	11,304																																								
Identity hate	2117																																								
Severe toxic	1962																																								
Threat	689																																								
	Twitter	Gab	Total																																						
Hateful	708	5227	5935																																						
Offensive	2328	3152	5480																																						
Normal	5770	2044	7814																																						
Undecided	249	670	919																																						
Total	9055	11,093	20,148																																						
Result / Conclusion (What was the final result?)	<p>The BERT variants (BERT + MLP and BERT + ANN) performed significantly better than other linear explainable models, achieving accuracies of 93.67% and 93.55% respectively. The BERT ANN model demonstrated the highest comprehensiveness score of</p> <ul style="list-style-type: none">• Accuracy of 93.55%• Precision of 95.2%,• Specificity of 83%and• F1-score of 94.14%																																								

Obstacles/Challenges (List the methodological obstacles if authors mentioned in the article)	Complex AI models with a large number of parameters, iterations, and optimization make it challenging to validate their results. Conventional AI models have limitations in terms of explainability, and their combination with XAI is not discussed in the research study.
Terminology (List the common basic words frequently used in this research field)	AU-Area under the ROC Curve XAI-Explainable Artificial Intelligence KNN-K-Nearest Neighbor RF-Random Forest SVM- Support Vector Machine LSTM-Long Short-Term Memory ANN-Artificial Neural Network
Review Judgment (Briefly compare the objectives and results of all the articles you reviewed)	<ul style="list-style-type: none"> • Manash Sarker et al. proposed a system to classify anti-social Bengali comments on social media and Multinomial Naive Bayes (MNB) had an accuracy of 80.51 • Dharmaraj R. Patil et al. proposed a system to detect Twitter hate speech and multi-model learning strategy had an accuracy of 96.29% • Mithun Das et al. proposed a system to detect hate speech and offensive language in Bengali and MuRIL model had an accuracy of 83.3% • Nauros Romim et al. proposed a system to detect Twitter hate speech and BiLSTM had an accuracy of 85.08% • Harshkumar Mehta et al. proposed a system of social media Hate Speech Detection Using Explainable Artificial Intelligence (XAI) and BERT+ANN had an accuracy of 93.55%
Review Outcome (Make a decision how to use/refer the obtained knowledge to prepare a separate and new methodology for your own research project)	By seeing all the machine learning classification algorithms used above, I would use a Hybrid BERT model due to its ability to capture rich contextual language understanding, Fine-Tuning, Open Source availability and increase the accuracy in Bengali Language.

Aspects	Paper # 21 (Title)
Title / Question (What is problem statement?)	Advances in Machine Learning Algorithms for Hate Speech Detection in Social Media: A Review
Objectives / Goal (What is looking for?)	<p>This paper aims to review machine learning (ML) algorithms and techniques for hate speech detection in social media (SM) and provide a comprehensive and updated state-of-the-art in this field.</p> <p>The paper aims to equip readers with information on the critical steps involved in hate speech detection using ML algorithms, evaluate the weaknesses and strengths of each method, and identify research gaps and open challenges in hate speech detection.</p> <p>The goal is to bridge the gap and keep professionals, old and new researchers informed about the current developments in hate speech detection using ML approaches.</p>
Methodology / Theory (How to find the solution?)	<p>The researchers used several databases, including IEEE Explore, ACM, ScienceDirect, Scopus, and Universiti Sains Malaysia databases, to gather relevant articles for the review work. These databases were chosen for their reputation and because they are subscribed by Universiti Sains Malaysia Library.</p> <p>The search was limited to a span of ten years (2010-2020) and specific key terms and phrases related to hate speech detection, such as offensive comments, aggressive comments, cyberbullying, profanity, and toxic comments on social media, were used for retrieval.</p> <p>Filter tools available in each database were utilized to narrow down the articles, focusing on computer science, engineering, and mathematics subjects. Only the most relevant articles that passed the inclusion test were downloaded.</p> <p>The inclusion criteria required that the papers addressed issues related to offensive comments (hate speech, cyberbullying, aggressive comments, toxic comments, etc.) on social media. The title and abstract of each paper were used to determine the inclusion.</p>
Software Tools (What program/software is used for design, coding and simulation?)	Implementation work was carried out at Intel(R) Core (TM) i7 CPU M60 @ 2.80 GHz in Python.
Test / Experiment How to test and characterize the design/prototype?	<p>The paper does not explicitly mention any specific test or experiment conducted by the authors. Instead, it focuses on reviewing and analyzing existing machine-learning algorithms and techniques for hate speech detection in social media.</p> <p>The researchers collected and explored data, extracted features, performed dimensionality reduction, selected and trained classifiers, and evaluated models as part of the baseline components of hate speech classification using ML algorithms.</p> <p>The paper also discusses the different variants of ML techniques, including classical ML, ensemble approaches, and deep learning methods, that have been employed for hate speech detection.</p>

Simulation/Test Data (What parameters are determined?)	While the paper does not provide specific simulation or test data, it highlights the improvements in ML algorithms and the proposal of new datasets and performance metrics in the literature.
Result / Conclusion (What was the final result?)	From this paper, we can say that between ML and DL algorithms, the Deep learning algorithms performed better results than Machine learning algorithms.
Obstacles/Challenges (List the methodological obstacles if authors mentioned in the article)	<p>One of the challenges mentioned is the need to handle hate speech messages that are contextual in nature, indicating that more research is required in this area.</p> <p>The paper also highlights the limitation of ignoring numeric symbols and special characters that may convey hate speech messages, suggesting the need for further investigation in this aspect.</p> <p>Another challenge could be the lack of a comprehensive coding guide benchmark to guide annotators in hate speech detection, which may require additional research and development.</p> <p>Overall, the paper acknowledges the need for ongoing research to address these challenges and improve the automatic detection of hate speech in social media using machine learning algorithms.</p>
Terminology (List the common basic words frequently used in this research field)	Artificial intelligence. Machine learning · Internet of things (IoT) · Healthcare · Fog computing · Learning classifier. KNN-K-Nearest Neighbor RF-Random Forest, SVM- Support Vector Machine
Review Judgment (Briefly compare the objectives and results of all the articles you reviewed)	<ul style="list-style-type: none"> • Amin et al. proposed a system to classify people with heart disease and healthy people and had an accuracy of 89% • Samuel et al. proposed a system to predict heart failure risks, decision support systems based on artificial neural networks and had an accuracy of 91.10%
Review Outcome (Make a decision how to use/refer the obtained knowledge to prepare a separate and new methodology for your own research project)	Along with the seven machine learning classification algorithms, I would use (analytic hierarchy process) AHP technique due to its simplicity, scalability, mathematical background, and ability to assess qualitative and quantitative factors to evaluate the effectiveness and efficacy of monitoring patients.

Aspects	Paper # 22 (Title)
Title / Question (What is problem statement?)	Bangla hate speech detection on social media using attention-based recurrent neural network
Objectives / Goal (What is looking for?)	The objective of the paper is to detect hate speech in the Bengali language on social media using an attention-based recurrent neural network.
Methodology / Theory (How to find the solution?)	<p>Data was extracted using Facebook API and saved into a CSV file. The dataset was split into a training set (80%) and a testing set (20%).</p> <p>Label Encoder was used to convert the training and testing sets into corresponding binary values for input into machine learning approaches.</p> <p>Preprocessing methods were applied to extract information from the data, including extracting information from emoticons and emojis to detect the type of speech.</p> <p>Bangla natural language tokenization was used to split sentences into words. Features were extracted using TF-IDF vectorization and word embedding.</p> <p>Classification approaches such as CNN, Bidirectional LSTM, and GRU were applied and compared for performance.</p> <p>Recurrent neural network (RNN) with Attention Mechanism was used for text classification. The performances of all the classification approaches were analyzed and compared.</p> <p>The paper also mentions the creation of data collection from popular public pages in Bangladesh, including news portals, celebrities, politicians, and cricketers</p>
Software Tools (What program/software is used for design, coding and simulation?)	Implementation work was carried out at Intel(R) Core (TM) i7 CPU M60 @ 2.80 GHz in Python.
Test / Experiment How to test and characterize the design/prototype?	<p>The paper conducted experiments to evaluate the performance of different machine learning algorithms in detecting hate speech in the Bengali language on social media.</p> <p>The dataset used for training and evaluation consisted of 7,425 Bengali comments from Facebook pages, categorized into seven distinct categories of hate speech.</p> <p>Three encoder-decoder algorithms, including attention-based decoder, LSTM, and GRU-based decoders, were used for predicting hate speech categories. Among these, the attention-based decoder achieved the best accuracy of 77%.</p> <p>The paper also compared the performance of different ML algorithms using binary-class and multiclass labels and found that the attention mechanism performed well and better than other algorithms in both cases.</p> <p>The test accuracy and f1-score for different hate speech categories were also analyzed and presented in the paper</p>

Simulation/Test Data (What parameters are determined?)	The dataset consisted of seven distinct categories of hate speeches. The dataset was split into a training set (80%) and a testing set (20%).
Result / Conclusion (What was the final result?)	The model showed high precision (0.78), high recall (0.75), and high f1-score (0.78) in classifying hate speech categories.
Obstacles/Challenges (List the methodological obstacles if authors mentioned in the article)	<p>Limited existing works: The paper highlights that very few works have been carried out in the context of hate speech detection in the Bengali language, despite millions of people communicating on social media in Bengali. This indicates a lack of prior research and established methodologies in this specific domain.</p> <p>Need for accuracy and interpretability improvements: The existing works in Bengali hate speech detection require improvements in both accuracy and interpretability. This suggests that the current methods may not be sufficiently effective in accurately identifying and classifying hate speech in Bengali language text.</p> <p>Dataset limitations: The paper mentions the use of a dataset of 7,425 Bengali comments for training and evaluation. However, it does not provide details about the representativeness or diversity of the dataset, which could potentially impact the generalizability of the model's performance.</p> <p>Lack of information on computational resources: The paper does not provide information on the computational resources used for training and evaluating the models. This could be a potential challenge as complex machine learning models like recurrent neural networks often require significant computational power and memory.</p>
Terminology (List the common basic words frequently used in this research field)	Artificial intelligence. Machine learning · Internet of things (IoT) · Healthcare · Fog computing · Learning classifier. KNN-K-Nearest Neighbor RF-Random Forest, SVM- Support Vector Machine
Review Judgment (Briefly compare the objectives and results of all the articles you reviewed)	<ul style="list-style-type: none"> • Amin et al. proposed a system to classify people with heart disease and healthy people and had an accuracy of 89% • Samuel et al. proposed a system to predict heart failure risks, decision support systems based on artificial neural networks and had an accuracy of 91.10%
Review Outcome (Make a decision how to use/refer the obtained knowledge to prepare a separate and new methodology for your own research project)	Along with the seven machine learning classification algorithms, I would use (analytic hierarchy process) AHP technique due to its simplicity, scalability, mathematical background, and ability to assess qualitative and quantitative factors to evaluate the effectiveness and efficacy of monitoring patients.

Aspects	Paper # 23 (Title)
Title / Question (What is problem statement?)	Detection of Hate Speech using BERT and Hate Speech Word Embedding with Deep Model
Objectives / Goal (What is looking for?)	<p>The objective of the paper is to detect hate speech in on-line social media platforms using advanced natural language processing techniques. The authors aim to address the challenges posed by hate communities, such as their use of abbreviations, intentional spelling mistakes, and coded words to evade detection.</p> <p>The paper investigates the feasibility of leveraging domain-specific word embedding as features and a bidirectional LSTM-based deep model as a classifier to automatically detect hate speech. The authors also explore the use of the transfer learning language model BERT for hate speech detection.</p>
Methodology / Theory (How to find the solution?)	<p>The paper proposes two approaches for detecting hate speech: Approach 1 involves using domain-specific word embedding features with a bidirectional LSTM-based deep model classifier, while Approach 2 utilizes the transfer learning language model BERT for binary classification of hate speech.</p> <p>Approach 1 includes steps such as data collection from hate speech datasets and Twitter, pre-processing of the collected data to remove nonmeaningful words and symbols, feature extraction using domain-specific word embedding (HSW2V), and classification using a deep sequential model with a bidirectional LSTM layer.</p> <p>Approach 2 focuses on using BERT, a pre-trained language model, for hate speech detection. BERT is trained on a large data corpus and provides high-performance results for various NLP tasks. The paper compares the performance of both approaches on a combined balanced dataset from available hate speech datasets. Approach 1 achieves a 93% f1-score, while BERT achieves a 96% f1-score.</p> <p>The authors highlight the influence of the size of the trained data on the performance of pre-trained models like BERT. Despite the variation in corpus size, Approach 1 achieves a close result to BERT because it is trained on data related to the same domain.</p>
Software Tools (What program/software is used for design, coding and simulation?)	Implementation work was carried out at Intel(R) Core (TM) i7 CPU M60 @ 2.80 GHz in Python.
Test / Experiment How to test and characterize the design/prototype?	<p>The paper conducts experiments to evaluate the performance of two approaches for hate speech detection: Approach 1 using domain-specific word embedding with a bidirectional LSTM-based deep model, and Approach 2 using the BERT language model.</p> <p>The experiments are performed on a combined balanced dataset from available hate speech datasets, including the Davidson-ICWSM, Waseem-EMNLP, and Waseem-NAACL datasets.</p> <p>Approach 1 achieves a 93% f1-score, while Approach 2 using BERT achieves a higher 96% f1-score.</p>

Simulation/Test Data (What parameters are determined?)	<p>The paper does not explicitly mention the details of the test data used for the experiments. However, it states that the experiments were conducted on a combined balanced dataset from available hate speech datasets, including the Davidson-ICWSM, Waseem-EMNLP, and Waseem-NAACL datasets.</p> <p>The authors also mention that the combined dataset was created by randomly selecting a similar number of examples for each class and balancing the dataset according to the lowest class number.</p>
Result / Conclusion (What was the final result?)	<p>The paper investigates the feasibility of using domain-specific word embedding and a bidirectional LSTM-based deep model, as well as the transfer learning language model BERT, for hate speech detection.</p> <p>The experiments show that the domain-specific word embedding with the bidirectional LSTM-based deep model achieved a 93% f1-score, while BERT achieved a 96% f1-score on a combined balanced dataset from available hate speech datasets.</p>
Obstacles/Challenges (List the methodological obstacles if authors mentioned in the article)	<p>The paper highlights the challenge of detecting coded words and emphasizes the importance of leveraging domain-specific word embedding to assign negative meanings to such words.</p> <p>The performance of pre-trained models like BERT is influenced by the size of the trained data, and there is a huge variation in the corpus size of hate speech datasets.</p>
Terminology (List the common basic words frequently used in this research field)	Artificial intelligence. Machine learning · Internet of things (IoT) · Healthcare · Fog computing · Learning classifier. KNN-K-Nearest Neighbor RF-Random Forest, SVM- Support Vector Machine
Review Judgment (Briefly compare the objectives and results of all the articles you reviewed)	<ul style="list-style-type: none"> • Amin et al. proposed a system to classify people with heart disease and healthy people and had an accuracy of 89% • Samuel et al. proposed a system to predict heart failure risks, decision support systems based on artificial neural networks and had an accuracy of 91.10%
Review Outcome (Make a decision how to use/refer the obtained knowledge to prepare a separate and new methodology for your own research project)	Along with the seven machine learning classification algorithms, I would use (analytic hierarchy process) AHP technique due to its simplicity, scalability, mathematical background, and ability to assess qualitative and quantitative factors to evaluate the effectiveness and efficacy of monitoring patients.

Aspects	Paper # 24 (Title)
Title / Question (What is problem statement?)	G-BERT: An Efficient Method for Identifying Hate Speech in Bengali Texts on Social Media
Objectives / Goal (What is looking for?)	The objective of the paper is to propose a technique, G-BERT, for identifying hate speech in Bengali social media posts and mitigating its negative impact on individuals.
Methodology / Theory (How to find the solution?)	The paper utilized a data crawling process to automatically collect data from various sources such as Bengali online news portals and social media sites. The BeautifulSoup Python library was used to extract posts and comments related to hashtags and emojis from social media platforms. Specific criteria and keywords were used to select data, focusing on content published within a specific time period and targeting platforms with a significant user base and active engagement. The collected data, consisting of posts, comments, and memes, were manually labeled by graduate students from the Advanced Machine Learning (AML) lab. The dataset was divided into training, testing, and validation sets using a stratified sampling approach. The proposed G-BERT model combined the Bidirectional Encoder Representations from Transformers (BERT) architecture and the Gated Recurrent Units (GRU) model to identify hate speech in Bengali social media posts.
Software Tools (What program/software is used for design, coding and simulation?)	Implementation work was carried out at Intel(R) Core (TM) i7 CPU M60 @ 2.80 GHz in Python.
Test / Experiment How to test and characterize the design/prototype?	The paper conducted an experiment to evaluate the performance of the proposed G-BERT model for identifying hate speech in Bengali social media posts. The experiment involved comparing the performance of the G-BERT model with several other classification algorithms. The evaluation metrics used to measure the performance of the model included accuracy, precision, recall, and F1-score. The results of the experiment showed that the G-BERT model outperformed all other classification algorithms tested, achieving an accuracy of 95.56%, precision of 95.07%, recall of 93.63%, and F1-score of 92.15%.
Simulation/Test Data (What parameters are determined?)	The specific details about the test data used in the paper are not mentioned in the research paper.
Result / Conclusion (What was the final result?)	The paper proposes a new model called G-BERT for identifying hate speech in Bengali texts on social media. The model combines the Bidirectional Encoder Representations from Transformers (BERT) architecture for extracting Bengali text properties and a Gated Recurrent Units (GRU) model with a Softmax activation function for categorizing hate speech. The performance of the G-BERT model was compared with several other algorithms, and it achieved an accuracy, precision, recall, and F1-score of 95.56%, 95.07%, 93.63%, and 92.15% respectively. The proposed model outperformed all other classification algorithms tested, indicating its effectiveness in locating hate speech in Bengali texts on social media platforms.

<p>Obstacles/Challenges (List the methodological obstacles if authors mentioned in the article)</p>	<p>Lack of research on recognizing hate speech in Bengali texts: The paper highlights the research gap in the field of hate speech detection in Bengali texts, indicating the lack of previous studies addressing this issue.</p> <p>The rapid increase in online hate speech: The paper acknowledges the rapid increase in Internet users and the corresponding rise in concerns such as hate speech, abusive texts, and harassment. This highlights the urgency and importance of addressing hate speech in Bengali texts on social media platforms.</p> <p>Complexity of hate speech detection: Identifying hate speech in Bengali texts poses challenges due to the nuances of language, cultural context, and the evolving nature of hate speech. Developing an efficient and accurate model for hate speech detection requires addressing these complexities.</p> <p>Performance comparison with other algorithms: The paper compares the performance of the proposed G-BERT model with several other classification algorithms. The challenge lies in achieving superior performance in terms of accuracy, precision, recall, and F1-score, which the G-BERT model successfully accomplishes</p>
<p>Terminology (List the common basic words frequently used in this research field)</p>	<p>Artificial intelligence. Machine learning · Internet of things (IoT) · Healthcare · Fog computing · Learning classifier. KNN-K-Nearest Neighbor RF-Random Forest, SVM- Support Vector Machine</p>
<p>Review Judgment (Briefly compare the objectives and results of all the articles you reviewed)</p>	<ul style="list-style-type: none"> • Amin et al. proposed a system to classify people with heart disease and healthy people and had an accuracy of 89% • Samuel et al. proposed a system to predict heart failure risks, decision support systems based on artificial neural networks and had an accuracy of 91.10%
<p>Review Outcome (Make a decision how to use/refer the obtained knowledge to prepare a separate and new methodology for your own research project)</p>	<p>Along with the seven machine learning classification algorithms, I would use (the analytic hierarchy process) AHP technique due to its simplicity, scalability, mathematical background, and ability to assess qualitative and quantitative factors to evaluate the effectiveness and efficacy of monitoring patients.</p>

Aspects	Paper # 25 (Title)
Title / Question (What is problem statement?)	A hybrid approach based on personality traits for hate speech detection in Arabic social media
Objectives / Goal (What is looking for?)	The objective of the paper is to propose a novel approach for detecting hate speech in Arabic social media by using personality trait features. The paper aims to address the increasing frequency of cyber hate speech in the Arab region and the need for automated detection of Arabic hate speech.
Methodology / Theory (How to find the solution?)	<p>The paper presents a hybrid approach for hate speech detection in Arabic social media based on personality traits. The proposed model consists of two phases, as depicted in Figure 1 of the paper. The functions of each phase are discussed in detail in the following subsections.</p> <p>The first phase involves extracting personality trait features from the social media data. The second phase focuses on using these extracted features to detect hate speech in Arabic social media.</p> <p>Detailed explanations of the framework model and the specific techniques used in each phase are provided in the paper.</p>
Software Tools (What program/software is used for design, coding and simulation?)	Implementation work was carried out at Intel(R) Core (TM) i7 CPU M60 @ 2.80 GHz in Python.
Test / Experiment How to test and characterize the design/prototype?	<p>The paper describes the experimental setup and findings in Section 4.</p> <p>The first phase of the experiments involves inferring personality-based features from the text using five models for each binary class.</p> <p>The second phase investigates the correlation between offensive language and personality characteristics using models proposed in previous work.</p> <p>In the third phase, the findings are compared after adding personality features to the original dataset. The experiments were conducted using Google Colab Pro and various libraries such as NumPy, Pandas, Re, Alphabet Detector, Sklearn, and Keras.</p> <p>The baseline models used in the experiments include supervised machine learning algorithms such as MNB, SVM, DT, and KNN, trained using the TF-IDF technique.</p> <p>A sample of 340 tweets from the Arabic hate speech dataset's development set was randomly selected for the experiment, with 170 offensive and 170 non-offensive tweets.</p>
Simulation/Test Data (What parameters are determined?)	The paper does not provide specific details about conducting simulations. The experimental setup focuses on inferring personality-based features, investigating the correlation between offensive language and personality traits, and comparing the findings after adding personality features to the dataset. The experiments involve using various machine learning algorithms, such as MNB, SVM, DT, and KNN, trained using the TF-IDF technique. The dataset used in the experiments consists of 92 Egyptian Twitter users' profiles, their Twitter feeds, and their personality ratings based on the five major personality traits.

<p>Result / Conclusion (What was the final result?)</p>	<p>The proposed hybrid approach based on personality traits achieved a superior macro-F1 score of 82.3% compared to previous work reported in the literature.</p> <p>The experimental results demonstrate the ability of the model to recognize offensive speech with a high precision rate of 58%.</p> <p>The models (A, C) outperformed the rest in terms of recall, achieving 95% and 81% respectively, indicating their capability to detect non-offensive speech .</p>
<p>Obstacles/Challenges (List the methodological obstacles if authors mentioned in the article)</p>	<p>Scouting for hate speech on social media is a significant difficulty due to the abundance of online content, making manual inspections almost impossible [1] . The intersection of personality learning and hate speech detection is a relatively less studied niche, indicating a lack of research in this area.</p> <p>The increase in the frequency of cyber hate speech in Arabic social media poses a major concern for stakeholders, highlighting the need for automated detection methods.</p> <p>Preparing the dataset for feature extraction involves several text preparation steps, such as removing punctuation, strange letters, numbers, and diacritics, as well as normalizing Arabic characters using a pre-trained word embedding model.</p> <p>The dataset used in the experiments consists of profiles of 92 Egyptian Twitter users, their Twitter feeds, and their personality ratings based on the five major personality traits, which may limit the generalizability of the findings to other contexts</p>
<p>Terminology (List the common basic words frequently used in this research field)</p>	<p>Artificial intelligence. Machine learning · Internet of things (IoT) · Healthcare · Fog computing · Learning classifier. KNN-K-Nearest Neighbor RF-Random Forest, SVM- Support Vector Machine</p>
<p>Review Judgment (Briefly compare the objectives and results of all the articles you reviewed)</p>	<ul style="list-style-type: none"> • Amin et al. proposed a system to classify people with heart disease and healthy people and had an accuracy of 89% • Samuel et al. proposed a system to predict heart failure risks, decision support systems based on artificial neural networks and had an accuracy of 91.10%
<p>Review Outcome (Make a decision how to use/refer the obtained knowledge to prepare a separate and new methodology for your own research project)</p>	<p>Along with the seven machine learning classification algorithms, I would use (analytic hierarchy process) AHP technique due to its simplicity, scalability, mathematical background, and ability to assess qualitative and quantitative factors to evaluate the effectiveness and efficacy of monitoring patients.</p>